ChainSpace: A Chained-Reasoning Paradigm for Spatial Intelligence

Xiaohan Zhang[1,2,*], Feng Gu[2,*,†], Xudong Rao[2], Xuhao Pan[2], Tao Wei[2], Zhou Pan[2,†], Kun Zhan[2]

[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]Li Auto Inc.

[*]Equal contribution. [†]Corresponding authors.
zhangxh2023@zju.edu.cn, gufeng@lixiang.com

Abstract

Spatial intelligence requires foundation models to maintain coherent spatial state across interactions with the physical world. However, existing data-centric approaches typically treat spatial reasoning as independent question-answer instances, enabling shortcut-based answering and providing limited supervision for persistent spatial understanding. To address this, we introduce `ChainSpace`, a chained-reasoning paradigm that structures spatial reasoning as a state-preserving multi-round process. In this paradigm, spatial questions are organized into logically constrained and jointly consistent chains, where later questions depend on spatial constraints established in earlier rounds. Following this principle, we instantiate `ChainSpace-Bench`, a manually annotated real-world multi-round benchmark with a *Chain-Aware Metric*, and `ChainSpace-Pipeline`, a simulator-based chain-structured supervision generation framework for spatial intelligence training. Experiments show that `ChainSpace-Bench` exposes chain-level failures that are not captured by isolated question accuracy. Additionally, with a relatively small amount of simulator-generated chained data, models trained by `ChainSpace-Pipeline` achieve the best performance among open-source models on `ChainSpace-Bench` and transfer competitively to multiple external spatial intelligence benchmarks. These results establish `ChainSpace` as an effective paradigm for more faithful evaluation and more data-efficient learning of spatial intelligence.

1 Introduction

Spatial intelligence is foundational for multimodal large language models [3, 49, 22] to interact with the physical world, with broad applications in vision-and-language navigation [1, 30], autonomous driving [45, 47], and robotics [9, 44]. However, despite rapid progress in multimodal foundation models [3, 49, 22], spatial intelligence remains a major challenge [8].

Existing works on spatial intelligence can be roughly grouped into architectural solutions [11, 12] and data-centric approaches [41, 8, 31]. Among them, data-centric approaches are particularly attractive for general-purpose foundation models, as they evaluate and improve spatial intelligence without altering the underlying model architecture. These approaches span evaluation benchmarks [41, 16, 43, 20], which diagnose spatial capabilities through curated test instances, and training-data construction [8, 32, 42], which improves spatial capability by scaling spatial supervision.

However, existing data-centric approaches, especially on the training-data side, often rely on collecting and curating large-scale real-world data [8, 42, 31], which often incurs substantial collection, annotation, and curation costs. More importantly, current spatial datasets [8, 42, 31, 16, 20, 41], whether for training or evaluation, typically treat spatial reasoning as a set of independent question-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

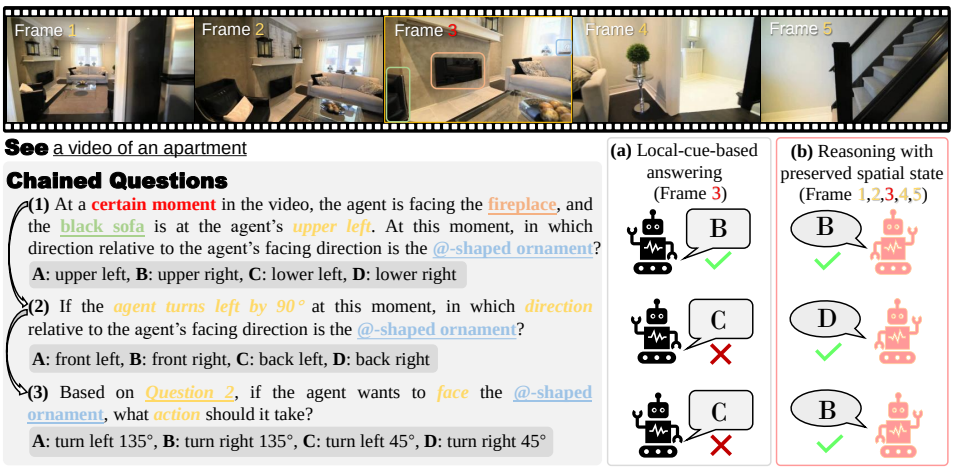


Figure 1: Illustration of the ChainSpace paradigm. Given a video trajectory, spatial questions are organized into a logically constrained chain, where later questions depend on spatial constraints established in earlier rounds. (a) A model following an independent-QA behavior may answer an early question correctly using local cues from a single frame, but fail on later questions that require preserving and updating spatial state. This illustrates why isolated question accuracy can overestimate spatial intelligence. (b) By reasoning with the preserved spatial state across frames and question turns, the model can produce answers that remain jointly consistent throughout the chain.

answer (QA) instances. Since models can answer them in isolation without maintaining a coherent understanding of the space, this leaves room for shortcut-based answering (Fig. 1(a)) through commonsense priors, local cues, or dataset-specific regularities [34], thus not only making evaluation less faithful to true spatial intelligence, but also providing only limited supervision during training.

At the core of these limitations is the lack of a structured formulation of persistent spatial state. In real-world embodied settings, spatial reasoning is rarely a collection of isolated questions. Instead, it requires a model to accumulate observations over time, maintain a coherent understanding of the environment, and answer multiple interdependent queries grounded in the same underlying spatial context [50], as shown in Fig. 1(b).

Based on this insight, we introduce ChainSpace, a chained-reasoning paradigm that moves beyond independent QA and organizes spatial reasoning as a multi-round process grounded in persistent spatial state. The core principle lies in organizing multiple questions as a logically constrained process, where later questions depend on earlier spatial constraints and the full answer sequence must remain jointly consistent. Under this paradigm, we instantiate both evaluation and learning. For evaluation, we construct ChainSpace-Bench, a manually annotated real-world multi-round benchmark with a *Chain-Aware Metric* for measuring chain-level consistency beyond isolated question accuracy. For learning, we propose ChainSpace-Pipeline, which generates simulator-based chained supervision from embodied trajectories. Experiments show that ChainSpace-Bench exposes chain-level failures missed by isolated question accuracy. With only 50K simulator-generated chained samples, models trained by ChainSpace-Pipeline achieve the best performance among open-source models on ChainSpace-Bench, outperform spatial models trained with hundreds of thousands to millions of real-world samples, and transfer competitively to external spatial intelligence benchmarks. Our contributions are summarized as

- We introduce ChainSpace, a chained-reasoning paradigm for spatial intelligence, which organizes spatial questions into logically constrained and jointly consistent chains grounded in persistent spatial state, making shortcut-based answering substantially harder.
- We instantiate ChainSpace for evaluation through ChainSpace-Bench, a manually annotated real-world multi-round benchmark, together with a *Chain-Aware Metric* for measuring chain-level spatial reasoning beyond isolated per-question accuracy.
- We instantiate ChainSpace for learning through ChainSpace-Pipeline, a simulator-based data construction pipeline that generates chained supervision from embodied trajectories, providing more informative supervision for learning spatial intelligence.

- Experiments show that `ChainSpace-Bench` exposes chain-level failures missed by isolated question accuracy. With only 50K simulator-generated chained samples, `ChainSpace-Pipeline` achieves the best performance among open-source models on `ChainSpace-Bench`, outperforming spatial models trained with hundreds of thousands to millions of real-world samples, and transfers competitively across multiple external spatial intelligence benchmarks.

2 Related Work

Existing efforts to evaluate and improve spatial intelligence in foundation models are largely built upon isolated supervision signals. For evaluation, benchmarks such as VSI-Bench [41], ViewSpatial-Bench [20], and MMSI-Bench [43] evaluate spatial reasoning limitations in multimodal models from different aspects. However, because these benchmarks typically score questions independently, they provide limited evidence of whether a model can preserve, update, and reuse spatial state across related queries. As a result, isolated accuracy may overestimate spatial intelligence, since models can answer individual questions using local cues, commonsense priors, or dataset-specific regularities [34]. A similar issue appears in training data. Recent data-centric efforts, such as SenseNova-SI [8], Cambrian-S [42], and RoboBrain [31], show that scaling spatial supervision improves model capability, but their samples are still largely organized as independent QA pairs. Such supervision provides limited pressure for models to preserve and update spatial constraints across steps, and may favor local pattern matching over reusable spatial reasoning. In contrast, `ChainSpace` organizes spatial data as logically dependent and jointly consistent chains. `ChainSpace-Bench` enables chain-level evaluation beyond isolated accuracy, while `ChainSpace-Pipeline` generates chained supervision that requires models to reuse accumulated spatial state. Together, they instantiate a data-centric paradigm for more faithful evaluation and more informative learning of spatial intelligence.

3 Chained Spatial Intelligence

Spatial intelligence in embodied settings is better viewed as *chained reasoning under a shared latent spatial state*. A spatial context x (*e.g.*, a video trajectory) induces an underlying spatial state, including object locations, viewpoint changes, room topology, and spatial relations accumulated over time [41]. Multiple spatial questions are therefore not isolated labels, but different queries over the same latent state [35, 50]. Given a multi-round question set $q_{1:T} = (q_1, \dots, q_T)$ and the corresponding answers $a_{1:T} = (a_1, \dots, a_T)$, the conventional independent-QA formulation implicitly treats prediction as

$$p(a_{1:T} \mid x, q_{1:T}) \approx \prod_{t=1}^{T} p(a_t \mid x, q_t), \tag{1}$$

which assumes that each round can be answered in isolation once the question and observation are given. This factorization ignores the structural coupling among answers that should be induced by the same spatial state. As a result, a model may achieve high per-question accuracy by exploiting local cues, commonsense priors, or dataset-specific regularities, without maintaining a coherent spatial understanding of the full scene [34].

Chained Spatial Reasoning. In contrast, we formulate spatial reasoning as a chained process. Let s_t denote the latent spatial state maintained after round t, which conceptually summarizes the spatial constraints established so far. Here, s_t is not assumed to be an explicit architectural module, but a representation of the persistent spatial information that a model should preserve, update, and reuse across question turns. We model chained spatial reasoning as

$$p(a_{1:T} \mid x, q_{1:T}) = \prod_{t=1}^{T} p(a_t \mid x, q_{\le t}, s_{t-1}), \tag{2}$$

with state transition $s_t = \Phi(s_{t-1}, x, q_t, a_t)$, where later predictions are conditioned on the spatial constraints accumulated from earlier rounds. From this perspective, Eq. 1 can be viewed as a degenerate case in which the intermediate spatial state is ignored or repeatedly reset, while Eq. 2 treats the answer sequence as a structured prediction problem under a shared spatial context. This formulation induces two key properties for valid chained spatial reasoning. The first is **dependency**,

i.e., at least one later-round answer should require constraints established in previous rounds, rather than being reducible to the current question alone. The second is **consistency**, *i.e.*, the full answer tuple $a_{1:T}$ should be jointly valid under the same spatial context. Let $\mathcal{C}(x, q_{1:T})$ denote the set of answer chains that satisfy the spatial constraints induced by x and $q_{1:T}$. A valid prediction should satisfy

$$a_{1:T} \in \mathcal{C}(x, q_{1:T}), \tag{3}$$

rather than merely maximizing isolated correctness for each a_t. This formulation motivates our evaluation and training design. For evaluation, isolated per-question accuracy only measures marginal correctness and may miss whether answers satisfy pairwise and full-chain spatial constraints. For training, Eq. 2 suggests that supervision should preserve cross-round dependencies, requiring later rounds to reuse and update spatial constraints established earlier. Thus, `ChainSpace` provides a paradigm for organizing spatial evaluation and supervision around shared-state, logically constrained answer chains. `ChainSpace-Bench` and `ChainSpace-Pipeline` are two concrete instantiations of this paradigm, for evaluation and learning respectively.

Why chained spatial reasoning reduces shortcut answering. Since each answer is optimized and scored marginally, a shortcut predictor can produce a plausible local answer from (x, q_t) without ensuring that the full answer tuple is spatially valid. In other words, independent QA does not explicitly penalize arbitrary answer combinations in $\mathcal{A}^T$, so correct early answers do not necessarily constrain later predictions. Chained reasoning instead turns spatial reasoning into constrained answer-chain prediction. For a T-round chain with answer space $\mathcal{A}$, valid predictions are restricted to $\mathcal{C}(x, q_{1:T}) \subseteq \mathcal{A}^T$, where cross-round constraints eliminate locally plausible but globally inconsistent answer combinations. Thus, state-agnostic shortcuts become less viable, since later answers must remain compatible with spatial constraints established earlier.

4 ChainSpace-Bench

4.1 Overview

Following the formulation in Sec. 3, we introduce `ChainSpace-Bench` to provide a faithful benchmark for evaluating spatial intelligence. The core distinction of `ChainSpace-Bench` from prior spatial intelligence benchmarks [41, 16, 20, 43, 35, 23, 36] lies in its chained multi-round reasoning. We organize each question set as a logically constrained reasoning process, where later questions depend on the spatial constraints established by earlier ones and the full chain must remain jointly consistent. `ChainSpace-Bench` contains 326 real-world videos and 1,304 3-rounds QA pairs. To ensure benchmark reliability, all annotations are manually verified under a strict quality-control protocol. We carefully remove erroneous labels, refine ambiguous cases, and only retain samples that reach agreement within the annotation team. More benchmark details can be found in the appendix.

4.2 Benchmark Construction

Data Collection. We build `ChainSpace-Bench` from RoomTour3D [18], which provides continuous first-person room-tour videos collected from real indoor environments. These videos contain natural temporal continuity, diverse room layouts, rich object arrangements, and multi-room trajectories, making them suitable for evaluating spatial understanding across time. Before annotation, we manually filter out videos with overly complex spatial structures, poor camera viewpoints, or insufficient visual coverage that prevents reliable annotation of spatial relations.

Question-Answer Annotation. Spatial intelligence involves a broad range of abilities, including counting, spatial relations, navigation, and size or distance estimation [8, 41, 20]. For each selected video, we annotate four spatial reasoning tasks, *i.e.*, object count, appearance order, relative direction, and route plan. We choose these 4 tasks because they are reliably annotatable on real-world videos and naturally support chained logical dependencies across questions. In contrast, size- and distance-centric tasks are more sensitive to viewpoint changes, making precise human annotation harder on real-world trajectories [46, 20]. To ensure consistency and reduce annotation ambiguity, we design a dedicated question template for each task type. Each template forms a 3-round QA chain, where the questions are progressively organized to reflect the intended logical dependency within the task. We adopt a 3-round design because it provides the minimal structure needed to move beyond isolated

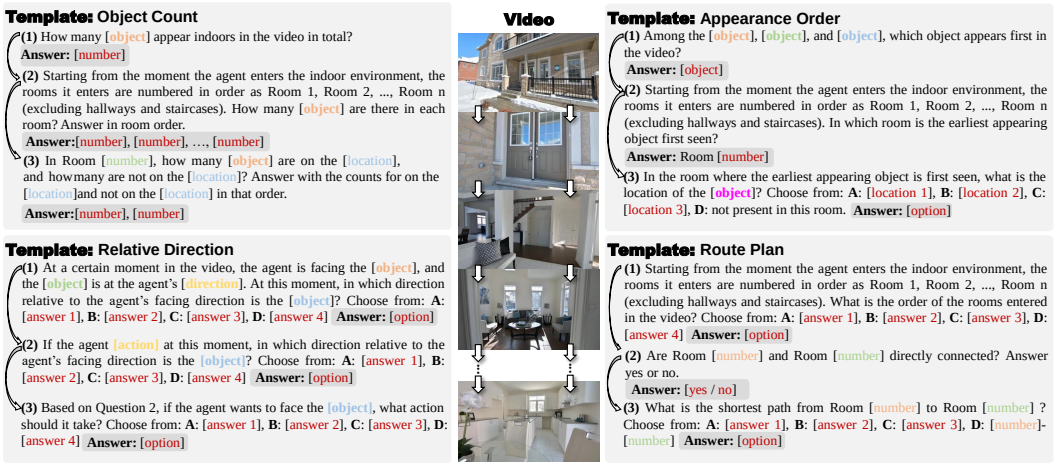


Figure 2: Illustration of our task templates. Each trajectory is instantiated into four 3-round question templates, *i.e.*, *object count*, *appearance order*, *relative direction*, and *route plan*. All templates follow the same chained design principle, where later questions depend on the spatial constraints established by earlier ones. For *object count*, the chain goes from global counting to room-wise decomposition and then to finer-grained counting within a specific room. For *appearance order*, it goes from earliest object identification to first-appearance room localization and then to object-location reasoning within that room. For *relative direction*, it goes from direction perception to hypothetical viewpoint change and then to action prediction. For *route plan*, it goes from room visitation sequence to room adjacency and then to shortest-path reasoning.

QA and simple pairwise dependency, while still remaining practical for reliable human annotation and review. The question templates for the four task types are illustrated in Fig. 2.

Annotation Quality. We adopt a two-stage quality-control process that combines automatic rule-based filtering with human-in-the-loop review. We first apply rule-based checks to remove QA chains with invalid answer formats, inconsistent object or room references, ambiguous options, or violations of the intended cross-round dependencies. The reviewer team then continuously monitors the annotation progress, examines the correctness of labeled QA chains, and provides timely feedback to the annotators whenever ambiguities or inconsistencies arise. Annotations are only retained when agreement is reached within the team.

4.3 Chain-Aware Metric

To align evaluation with the principle of `ChainSpace`, we propose a *Chain-Aware Metric*. Instead of scoring questions independently [41, 43, 20], it rewards both local correctness and pairwise/full-chain consistency. Because `ChainSpace-Bench` contains manually verified cross-round dependencies, correctness patterns indicate whether predictions follow the intended spatial chain. The metric is defined as

$$R_{\text{chain}}(\mathbf{s}) = \sum_{t=1}^{3} \alpha_t s_t + \sum_{i<j} \beta_{ij} s_i s_j + \gamma s_1 s_2 s_3, \tag{4}$$

where $s_t \in \{0, 1\}$ denotes whether the answer at round t is correct and $\mathbf{s} = (s_1, s_2, s_3)$. The α terms measure local correctness, the β terms reward pairwise cross-round consistency, and γ captures full-chain consistency. The score is normalized by the maximum attainable value. In our experiments, we set $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$, $\beta_{12} = 2.0$, $\beta_{23} = 1.5$, $\beta_{13} = 1.25$, and $\gamma = 3.0$. We assign smaller weights to isolated correctness and larger weights to chained terms, so that the metric emphasizes whether the full spatial reasoning chain is established rather than being dominated by per-question accuracy. Among pairwise terms, β_{12} is largest because (s_1, s_2) captures the most direct local dependency, while β_{23} and β_{13} are lower to reflect incomplete or less reliable chain evidence without the full intermediate reasoning path. Finally, γ receives the largest weight to reward fully correct and jointly consistent chains. We further ablate these hyper-parameters in Table 7 to show the robustness of weight choices.

5 ChainSpace-Pipeline

5.1 Overview

Following the principle in Sec. 3, we propose `ChainSpace-Pipeline`, a simulator-based chain-structured supervision generation framework for spatial intelligence training. We use ProcTHOR [14] and AI2-THOR [19] to generate diverse indoor scenes and collect continuous exploration trajectories through automatic embodied navigation. For each trajectory, we construct 4 task instances, corresponding to object count, appearance order, relative direction, and route plan. Each instance is instantiated as a 3-round chained template from trajectory-specific spatial states and metadata, so that later rounds depend on spatial constraints established earlier. Unlike prior data-construction pipelines that largely organize supervision as independent QA [8, 42, 5], `ChainSpace-Pipeline` preserves the dependency structure of spatial reasoning. This turns simulator metadata into chained supervision that requires models to reuse and update accumulated spatial state across rounds, providing a more structured training signal for learning spatial intelligence.

5.2 Pipeline Construction

Trajectory and Scene Generation To construct training data at scale, we adopt a simulator-based pipeline. This enables low-cost generation of a large amount of data, which is difficult to obtain through manual annotation on real-world videos. Moreover, since the model will be evaluated on real-world benchmarks, this setup helps verify that performance gains come from the proposed chained design principle itself, rather than from additional training on benchmark-specific data. To support such a pipeline, the simulator needs to generate diverse multi-room indoor scenes with rich object layouts and provide precise spatial states together with object-level metadata for automatic question construction. We therefore use ProcTHOR [14] to generate indoor houses, and AI2-THOR [19] to collect continuous exploration trajectories through automatic embodied navigation. During exploration, the agent traverses reachable positions while we record RGB observations, agent poses, room transitions, visible objects, and house-level structural metadata. As a result, the generated trajectories exhibit diverse spatial layouts, realistic navigation patterns, and continuous spatial context over time, making them well suited for constructing logically constrained multi-round supervision.

Chained Question-Answer Construction. For each scene and exploration trajectory, we construct four task instances, one for each task type (*e.g.*, object count, appearance order, relative direction, and route plan). Each instance is formulated as a 3-round chained question template (Fig. 2) instantiated from trajectory-specific spatial states and metadata, where later questions depend on earlier spatial constraints. Fig. 2 shows the canonical template form. In the actual training data, we vary the phrasing so that supervision is tied to the chained reasoning structure rather than to fixed linguistic patterns.

Why choose these 4 tasks? We focus on these 4 tasks because they cover representative spatial abilities, while naturally supporting explicit cross-round dependencies. Other spatial abilities, such as size- and distance-centric reasoning, are also important but less suitable for our current pipeline, as they require precise numerical estimation and are more sensitive to simulator-to-real discrepancies [8]. Heavy training on such synthetic numerical signals may further bias the model's depth-related priors and hurt generalization [2, 24].

Why choose 3-round QA chain? We adopt a 3-round design because it is sufficient to create meaningful chained supervision. The first round establishes a spatial constraint, the second refines it, and the third requires further reasoning under the accumulated context [48]. This yields non-trivial cross-round dependencies while remaining stable for large-scale automatic construction, whereas longer chains are more likely to introduce ambiguity, noise, and error propagation [48, 27].

Quality Control and Data Filtering. We ensure training data quality through rule-based filtering followed by human inspection. For each generated QA chain, we first apply automatic checks to remove invalid answer formats, inconsistent object or room references, ambiguous options, violated cross-round dependencies, and trajectories with excessively long videos (more than 100 frames). We then manually review a validation subset of 100 trajectories (400 chained QA samples) using the same reviewer team as in Sec. 4. Full-scale data generation is finalized only after this subset passes human inspection, and the remaining filtered samples constitute the final training set.

6 Experiment

6.1 Experimental Setup

Datasets. For training, we use the simulation data generated by `ChainSpace-Pipeline`. For evaluation, we use two groups of benchmarks. Our annotated `ChainSpace-Bench` is used to directly test the proposed chained design principle, while five existing spatial intelligence benchmarks, including VSI-Bench [41], MMSI-Bench [43], MindCube [35], ViewSpatial [20], and EmbSpatial [16], are used to assess the generalization of the learned spatial reasoning ability.

Evaluation Metrics On `ChainSpace-Bench`, we report both per-question accuracy and our proposed *Chain-Aware Metric*. Per-question accuracy measures local correctness at each round, while the Chain-Aware Metric further evaluates logical consistency across the full reasoning chain. Besides, we report Full-chain Exact Match as a strict unweighted chain-level metric, ensuring that the observed gain does not depend on a particular weighting scheme. For existing spatial intelligence benchmarks, we follow their original evaluation protocols. Specifically, MMSI-Bench [43], MindCube [35], ViewSpatial [20], and EmbSpatial [16] are evaluated using accuracy. VSI-Bench [41] reports exact-match accuracy for multiple-choice questions and Mean Relative Accuracy (MRA) for numerical questions. For the end-to-end embodied navigation evaluation on NavBench [29], we report *Success Rate* (SR) and *Success weighted by Path Length* (SPL) under the easy, medium, and hard difficulty levels. We additionally report *Exec. Avg.*, defined as the average of SR and SPL across the three difficulty levels, to summarize overall embodied-navigation performance.

Implementation Details We use Qwen3-VL-8B-Instruct [3] as the main base model. InternVL3-8B [49] is also used to evaluate our generalization ability. The models are trained on 8 NVIDIA H20 GPUs with full fine-tuning, where the ViT encoder is frozen. We adopt bfloat16 training with FlashAttention, gradient checkpointing, and DeepSpeed ZeRO-2 for memory-efficient optimization. For training on `ChainSpace-Pipeline`, we use 1 epoch, a per-device batch size of 1, gradient accumulation steps of 4, and a learning rate of 1×10^{-7} with a warmup ratio of 0.05. The maximum sequence length is set to 32768. We also enable DFT loss during training.

6.2 Main Results

Faithful reflection of spatial intelligence. As shown in Table 1, `ChainSpace-Bench` *more faithfully reflects spatial intelligence.* RoboBrain2.5 [31] improves the Avg. Question Accuracy over the Qwen3-VL-8B [3] baseline after large-scale spatial training, yet its Chain-Aware Score decreases. This mismatch indicates that isolated question accuracy alone may overestimate a model's spatial ability [34]. The proprietary-model results also suggest that `ChainSpace-Bench` is non-trivial even for strong frontier systems, while the larger differences under the Chain-Aware score indicate that our metric provides a discriminative power beyond isolated question accuracy.

Reliable chained spatial reasoning. Tables 1, 2, and 3 provide a progressive evidence chain for the effectiveness of `ChainSpace` **(1)** `ChainSpace-Pipeline` *improves chain-level spatial reasoning with much less data.* As shown in Table 1, our model outperforms open-source spatial models trained on substantially larger real-world datasets, despite using only $\sim$50K simulator-generated samples. This suggests that organizing supervision under `ChainSpace` can be more effective than simply scaling independent spatial QA data. **(2)** *The improvement is reliable and concentrated on chain-level metrics.* Table 2 compares different tuned models against the same untuned baseline. Our method significantly improves both Full-chain Exact Match and Chain-Aware Score, while SenseNova-SI [8] does not show significant gains on either chain-level metric. This indicates that the benefit of our pipeline is not merely better isolated answering, but more complete chained reasoning. **(3)** *The gain comes from better use of accumulated spatial state.* Table 3 performs an oracle-history intervention. Notably, our model obtains the largest gains in Q3 Acc. under correct history, revealing that correct spatial history unlocks substantially stronger downstream reasoning. This suggests that the main bottleneck of previous trained spatial models is the ability to preserve and propagate correct spatial state across rounds. `ChainSpace-Pipeline` encourages models to update and reuse accumulated spatial constraints, enabling them to benefit more effectively from correct reasoning history.

Table 1: Results on `ChainSpace-Bench`. Open-source models and our method are all built on Qwen-family (Qwen2.5-VL-7B-Instruct [4] and Qwen3-VL-8B-Instruct [3]) for fair comparison. Existing baselines are trained with different data scales, including CamBrain-S (∼590K), RoboBrain2.0 (∼5.3M), RoboBrain2.5 (∼12.4M), and SenseNova-SI (∼8M), while our `ChainSpace-Pipeline` uses only ∼50K simulator-generated samples. The text in red and blue denotes the best and second-best results among open-source models.

Model	Avg. Question Accuracy	Avg. Chain-Aware Score
Proprietary Models		
Seed-2.0-Pro [7]	0.4289	0.2722
GPT-5 [26]	0.4825	0.3262
Gemini-3.1-Pro [17]	0.3453	0.1882
Open-source General Models		
Qwen2.5-VL-7B-Instruct [4]	0.2344	0.0819
Qwen3-VL-8B-Instruct [3]	0.3344	0.1522
Open-source Spatial Intelligence Models		
CamBrain-S-7B [42]	0.1824	0.0674
RoboBrain2.0-7B [32]	0.2873	0.1098
RoboBrain2.5-8B [31]	0.3384	0.1339
SenseNova-SI-8B [8]	0.3418	0.1555
(**Ours**) `ChainSpace-Pipeline-8B`	**0.3510**	**0.1687**

Table 2: Paired comparison against the same baseline (Qwen3-VL-8B-Instruct [3]) on `ChainSpace-Bench`. Each cell reports the mean paired difference Δ, the 95% paired-bootstrap confidence interval, and the paired-bootstrap p-value in the format Δ [CI] (p). **Bold** indicates statistically significant change ($p < 0.05$).

Metric	(**Ours**) `ChainSpace-Pipeline`	SenseNova-SI
(*Chain-level metrics*) Full-chain Exact Match	**+0.0153** [0.0054, 0.0261] (0.0044)	+0.0038 [-0.0100, 0.0176] (0.6380)
(*Chain-level metrics*) Avg. Chain-Aware Score	**+0.0165** [0.0079, 0.0256] (<0.0001)	+0.0033 [-0.0086, 0.0154] (0.5980)
(*Turn-level metrics*) Turn-1 Accuracy	**+0.0261** [0.0146, 0.0383] (<0.0001)	+0.0092 [-0.0146, 0.0330] (0.4720)
(*Turn-level metrics*) Turn-2 Accuracy	+0.0107 [-0.0038, 0.0253] (0.1692)	**-0.0215** [-0.0376, -0.0054] (0.0116)
(*Turn-level metrics*) Turn-3 Accuracy	+0.0130 [-0.0077, 0.0345] (0.2416)	**+0.0345** [0.0084, 0.0606] (0.0108)

Table 3: Oracle-history intervention on ChainSpace-Bench. "Pred" denotes normal inference using model-generated previous answers, while "Gold" replaces previous answers with ground-truth answers during inference. Relative gains are computed as $(\text{Gold} - \text{Pred})/\text{Pred}$. The text in red and blue denotes the best and second-best results of each setting.

Model	Setting	Avg. Acc.	Q2 Acc.	Q3 Acc.	Chain-Aware
Qwen3-VL-8B [3]	Pred	0.3344	0.3574	0.3497	0.1522
	Gold	0.3563	0.3635	0.4095	0.1589
	Gain	+6.55%	+1.71%	+17.10%	+4.40%
SenseNova-SI [8]	Pred	0.3418	0.3359	**0.3842**	0.1555
	Gold	0.3635	0.3528	0.4325	0.1628
	Gain	+6.35%	**+5.03%**	+12.57%	+4.69%
(**Ours**) `ChainSpace-Pipeline-50k`	Pred	**0.3510**	**0.3681**	0.3627	**0.1687**
	Gold	**0.3873**	**0.3781**	**0.4617**	**0.1806**
	Gain	**+10.34%**	+2.72%	**+27.30%**	**+7.05%**

Effect of Chained Supervision. To isolate the effect of chained supervision, we construct an *Independent-QA* control by converting each ChainSpace chain into independent samples, where the preceding ground-truth answer is explicitly inserted into the subsequent question. The control uses the same backbone, videos, underlying questions and answers, number of QA turns, supervision content, and training configuration as ChainSpace-Pipeline, with the cross-round dependency structure being the only difference. As shown in Table 4, Independent-QA achieves 0.3236 average question accuracy and 0.1482 Chain-Aware Score on ChainSpace-Bench, both below the untuned backbone, whereas ChainSpace-Pipeline improves them to 0.3510 and 0.1687. This controlled comparison shows that

Table 4: Controlled comparison of different training paradigms on ChainSpace-Bench and external spatial reasoning benchmarks. All settings use the same backbone. Independent-QA uses the same videos, questions, answers, supervision content, and training configuration as ChainSpace-Pipeline, while removing cross-round dependencies. The red and blue denote the best and second-best results, respectively.

Models	**ChainSpace-Bench**		**VSI-Bench [41]**	**MMSI-Bench [43]**	**MindCube [35]**
Metric	**Avg. Q. Acc**	**Avg. Chain-Aware Score**	**MRA, Acc**	**Acc**	**Acc**
Qwen3-VL-8B-Instruct [3]	0.3344	0.1522	57.9	31.1	29.4
Independent-QA Control (8B)	0.3236	0.1482	58.1	30.4	30.2
ChainSpace-Pipeline-8B (Ours)	**0.3510**	**0.1687**	**58.8**	**31.8**	**31.1**

simply training on the same videos and QA supervision does not yield the improvement. Preserving cross-round dependencies is essential. Moreover, ChainSpace-Pipeline consistently outperforms the matched Independent-QA control on all three external spatial reasoning benchmarks. These results further demonstrate that the benefit of chained supervision extends beyond ChainSpace-Bench and transfers to external spatial reasoning tasks.

Table 5: Comparison between proprietary models, open-source general models, and open-source spatial intelligence models across five spatial intelligence benchmarks. We report our 50K and 150K settings to highlight data efficiency. MindCube* denotes MindCube-Tiny.

Models	**VSI-Bench [41]**	**MMSI-Bench [43]**	**MindCube* [35]**	**ViewSpatial [20]**	**EmbSpatial [16]**
Metric	**MRA, Acc**	**Acc**	**Acc**	**Acc**	**Acc**
Random Choice	34.0	25.0	33.0	26.3	-
Proprietary Models					
Seed-1.6-2025-06-15 [6]	49.9	38.3	48.7	43.8	-
Gemini-2.5-Pro-2025-06 [13]	53.5	38.0	57.6	46.0	78.9
Grok-4-2025-07-09 [40]	47.9	37.8	63.5	43.2	75.7
GPT-5-2025-08-07 [26]	55.0	41.8	56.3	45.5	81.6
Gemini-3-Pro-Preview [17]	52.5	45.2	70.8	50.3	-
Open-source General Models					
Bagel-7B-MoT [15]	31.4	31.0	34.7	41.3	73.1
Qwen2.5-VL-7B-Instruct [4]	32.3	26.8	36.0	36.8	-
Qwen3-VL-8B-Instruct [3]	57.9	31.1	29.4	42.2	77.7
InternVL3-8B [49]	42.1	28.0	41.5	38.6	76.4
Open-source Spatial Intelligence Models					
SpatialLadder-3B [21]	44.8	27.4	43.4	39.8	-
Spatial-MLLM-4B [37]	46.3	26.1	33.4	34.6	-
SpaceR-7B [28]	41.5	27.4	37.9	35.8	66.9
ViLaSR-7B [39]	44.6	30.2	35.1	35.7	67.3
SenseNova-SI-100K$_{\text{Qwen3-VL-8B}}$ [8]	56.4	31.4	30.9	42.8	79.1
SenseNova-SI-200K$_{\text{Qwen3-VL-8B}}$ [8]	58.7	30.9	29.4	43.5	79.1
Ours					
Qwen3-VL-8B$_{\text{Ours-50K}}$	58.8(+1.5%)	31.8(+2.3%)	31.1(+5.8%)	43.0(+1.9%)	78.1(+0.5%)
InternVL3-8B$_{\text{Ours-150K}}$	43.6(+3.6%)	28.4(+1.4%)	44.5(+7.2%)	46.8(+21.2%)	76.2(-0.2%)

Efficient and generalizable spatial intelligence learning. Table 5 evaluates the generalization ability of models trained with `ChainSpace-Pipeline`. **(1)** *The learned capability transfers beyond* `ChainSpace-Bench`. Our models remain competitive and often improve over their corresponding untuned backbones on multiple external spatial intelligence benchmarks. **(2)** *Our supervision is data-efficient.* With only 50K simulator-generated chained samples, our Qwen3-VL-8B-Instruct model achieves performance comparable to SenseNova-SI-200K [8], despite using substantially less training data. Since `ChainSpace-Pipeline` is built from purely simulator-generated data and remains distributionally distinct from all evaluation benchmarks, these results provide additional evidence that our design principle contributes to spatial intelligence learning beyond the training distribution. Moreover, compared with prior approaches, our method remains comparable with much smaller training data, showing that better-organized supervision can be more effective than simply increasing independent QA samples. **(3)** *The gain is not limited to a single backbone.* Consistent improvements on both Qwen3-VL-8B-Instruct [3] and InternVL3-8B [49] suggest that the proposed chained supervision principle is broadly useful for training spatial intelligence.

Table 6: End-to-end embodied navigation performance on NavBench [29]. All models are evaluated without NavBench-specific training. Red and blue denote the best and second-best results among open-source models, respectively. Exec. Avg. denotes the average of SR and SPL across the three difficulty levels.

Model	Easy		Medium		Hard		Exec. Avg.
	SR	SPL	SR	SPL	SR	SPL	
GPT-4o [25]	67.36	54.31	41.67	35.71	27.78	21.15	41.33
Qwen2.5-VL-7B-Instruct [4]	41.67	32.55	22.92	17.43	10.42	5.67	21.78
Qwen3-VL-8B-Instruct [3]	50.69	38.04	**42.36**	**34.05**	26.39	20.87	35.40
SenseNova-SI-8B [8]	46.53	33.48	25.00	17.64	22.22	15.32	26.70
ChainSpace-Pipeline-8B (Ours)	**53.47**	**39.65**	**42.36**	32.66	**29.86**	**22.82**	**36.80**

End-to-End Embodied Navigation. To evaluate whether the spatial capabilities learned from ChainSpace transfer beyond QA-style reasoning, we further evaluate ChainSpace-Pipeline-8B on NavBench [29], an end-to-end embodied navigation benchmark requiring sequential perception, planning, and action. We evaluate 432 episodes, including 144 episodes at each of the easy, medium, and hard difficulty levels, under the same NavBench protocol without any NavBench-specific training. We additionally report the official results of GPT-4o [25] and Qwen2.5-VL-7B-Instruct [4] as closed- and open-source references. As shown in Table 6, ChainSpace-Pipeline-8B achieves the best overall navigation performance among the evaluated open-source models, improving the execution average from 35.40 to 36.80 over the Qwen3-VL-8B-Instruct [3] backbone. The improvement is particularly clear on the hard split, where the success rate increases from 26.39 to 29.86 and SPL from 20.87 to 22.82. Notably, these gains are obtained using only 50K simulator-generated chained samples, whereas SenseNova-SI-8B is trained with approximately 8M synthetic and real-world spatial samples. These results provide evidence that the spatial capability learned through chained supervision transfers to sequential embodied decision making, while also highlighting the data efficiency of organizing supervision around cross-round spatial dependencies.

Table 7: Ranking robustness under different weight settings of our *Chain-Aware Metric* (Eq. (4)) on `ChainSpace-Bench`. We vary the weights of the first-order terms (α_t), second-order terms (β_{ij}), and the third-order term (γ), and report the resulting normalized Avg. Chain-Aware scores for representative methods in Table 1. The text in red and blue denotes the best and second-best results.

Setting	$\alpha_1, \alpha_2, \alpha_3$	$\beta_{12}, \beta_{23}, \beta_{13}$	γ	Qwen3-VL-8B-Instruct [3]	RoboBrain2.5 [31]	SenseNova-SI [8]	Ours
W_1	(0.5,0.5,0.5)	(2.0,1.5,1.25)	3.0	0.1522	0.1339	0.1555	**0.1687**
W_2	(1.0,1.0,1.0)	(0.0,0.0,0.0)	0.0	0.3344	0.3384	0.3418	**0.3510**
W_3	(0.0,0.0,0.0)	(1.0,1.0,1.0)	2.0	0.1161	0.0925	0.1189	**0.1327**
W_4	(1.0,1.0,1.0)	(0.0,0.0,0.0)	2.0	0.2298	0.2221	0.2357	**0.2459**
W_5	(1.0,1.0,1.0)	(1.0,1.0,1.0)	1.0	0.1979	0.1847	0.2025	**0.2145**

Robustness and interpretability of the Chain-Aware Metric. As shown in Table 7, **(1)** *our Chain-Aware Metric is robust.* The relative ranking among representative methods remains unchanged under all five weight settings. This shows that the main conclusion of our paper does not depend on a particular choice of weights. **(2)** *The metric is interpretable and flexible*. Different weight settings naturally correspond to different evaluation perspectives. For example, W_2 removes all chained terms and reduces the metric to per-question correctness. **(3)** *The default setting W_1 is well aligned with our goal*. It provides a balanced evaluation of local correctness and chained consistency, while explicitly assigning greater importance to more complete logical chains. This matches our core motivation.

7 Conclusion

We presented `ChainSpace`, a paradigm for evaluating and learning spatial intelligence through chained reasoning. By organizing spatial reasoning as a logically dependent multi-round process, we provide both a more faithful benchmark for persistent spatial reasoning and a more informative supervision signal for training. Experiments validate the effectiveness of our design principle.

Acknowledgements and Disclosures

Funding. This work was supported by Li Auto Inc., which provided GPU computing resources and research infrastructure.

Competing Interests. Feng Gu, Xudong Rao, Xuhao Pan, Tao Wei, Zhou Pan, and Kun Zhan are employees of Li Auto Inc. Xiaohan Zhang conducted this work during an internship at Li Auto Inc. The authors declare no other competing interests.

References

[1] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. BEVBert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[2] Amir Atapour-Abarghouei and Toby P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. SIMS-V: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025.

[6] ByteDance Seed Team. Introduction to techniques used in seed1.6. `https://seed.bytedance.com/blog/introduction-to-techniques-used-in-seed1-6`, 2025. Published on June 25, 2025. Accessed: 2026-04-28.

[7] ByteDance Seed Team. Seed2.0 model card. `https://github.com/ByteDance-Seed/Seed2.0`, 2026. Accessed: 2026-04-28.

[8] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling Spatial Intelligence with Multimodal Foundation Models. *arXiv preprint arXiv:2511.13719*, 2025.

[9] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[10] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? an early investigation into training r1-like reasoning large vision-language models. In *arXiv preprint arXiv:2504.11468*, 2025.

[11] Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *Adv. Neural Inform. Process. Syst.*, 2025.

[12] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *Adv. Neural Inform. Process. Syst.*, 2024.

[13] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025.

[14] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *Adv. Neural Inform. Process. Syst.*, 2022.

[15] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[16] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, 2024.

[17] Google. A new era of intelligence with gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3/`, November 2025. Accessed: 2026-05-07.

[18] Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. RoomTour3D: Geometry-aware video-instruction tuning for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27586–27596, June 2025.

[19] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[20] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models, 2025.

[21] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *Int. Conf. Learn. Represent.*, 2026.

[22] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVa: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024.

[23] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

[24] Maxim Maximov, Kevin Galim, and Laura Leal-Taixé. Focus on defocus: Bridging the synthetic to real domain gap for depth estimation. In *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[25] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[26] OpenAI. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/, 2025. Published on August 7, 2025. Accessed: 2026-04-28.

[27] Anonymous or Authors. Tight error propagation bounds for multi-step chain-of-thought reasoning. *OpenReview*, 2026. under review.

[28] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

[29] Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. Navbench: Probing multimodal large language models for embodied navigation. *arXiv preprint arXiv:2506.01031*, 2025.

[30] 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 Conf. Comput. Vis. Pattern Recog.*, 2025.

[31] Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

[32] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoxu Lyu, Yijie Xu, Jiayu Shi, Mengfei Du, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Junkai Zhao, Xiaojie Zhang, Shanyu Rong, Huaihai Lyu, Zhengliang Cai, Yankai Fu, Ning Chen, Bolun Zhang, Lingfeng Zhang, Shuyi Zhang, Dong Liu, Xi Feng, Songjing Wang, Xiaodan Liu, Yance Jiao, Mengsi Lyu, Zhuo Chen, Chenrui He, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

[33] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

[34] Vishaal Udandarao, Shyamgopal Karthik, Surabhi S. Nath, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Solving spatial supersensing without spatial supersensing, 2025.

[35] Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

[36] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF Int. Conf. Comput. Vis.*, 2025.

[37] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *Adv. Neural Inform. Process. Syst.*, 2025.

[38] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[39] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025. URL https://arxiv.org/abs/2506.09965.

[40] xAI. Grok 4 model card. https://data.x.ai/2025-08-20-grok-4-model-card.pdf, 2025. Last updated on August 20, 2025. Accessed: 2026-04-28.

[41] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

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

[43] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *Int. Conf. Learn. Represent.*, 2025.

[44] Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. VLA-R1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025.

[45] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.

[46] Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, and Alex Wong. RSA: Resolving scale ambiguities in monocular depth estimators through language descriptions. In *Adv. Neural Inform. Process. Syst.*, 2024.

[47] Jiawei Zhang, Xuan Yang, Taiqi Wang, Yu Yao, Aleksandr Petiushko, and Bo Li. SafeAuto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. In *Forty-second International Conference on Machine Learning*, 2025.

[48] Y. Zhang et al. Turnbench-ms: A benchmark for evaluating multi-turn, multi-step reasoning. *arXiv preprint arXiv:2506.01341*, 2025.

[49] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[50] Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang. Video-MSR: Benchmarking multi-hop spatial reasoning capabilities of mllms, 2026.

A Appendix

A.1 More Details of ChainSpace-Bench

Dataset profile of ChainSpace-Bench. ChainSpace-Bench exhibits substantial diversity in both temporal complexity and answer space. As shown in Fig. 3, the benchmark covers trajectories with varied lengths, with most samples concentrated in a moderate frame range and a long tail of substantially longer videos, indicating that evaluation is not restricted to short or temporally simple cases. In addition, Fig. 4 shows that the A/B/C/D answer distribution remains reasonably balanced across all three rounds, reducing the possibility that models can achieve strong performance by exploiting simple option-frequency bias. Moreover, the benchmark is also well balanced in task-specific aspects. For action prediction, different actions are distributed relatively evenly, with turns

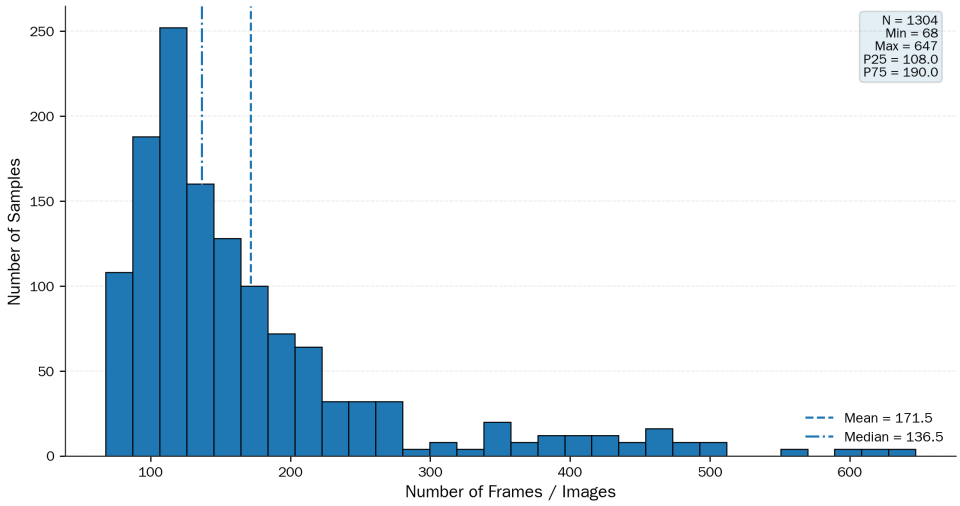


Figure 3: Frame count distribution of `ChainSpace-Bench`. The benchmark covers trajectories with diverse lengths, with most samples concentrated in the moderate range and a long tail of substantially longer videos.

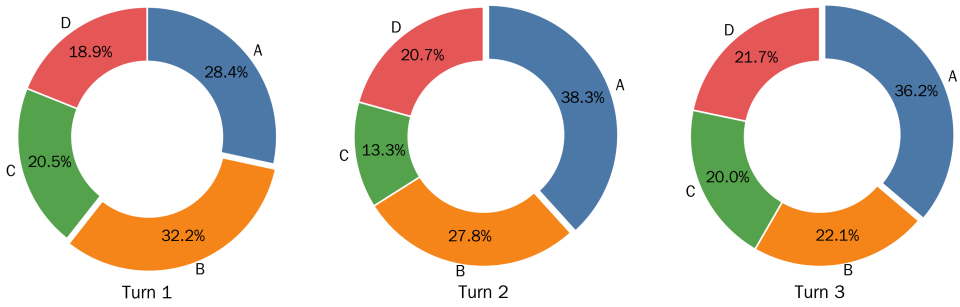


Figure 4: The A/B/C/D answer distribution in `ChainSpace-Bench` by turn. The option labels are reasonably balanced across rounds, reducing the possibility that models can perform well by exploiting simple answer-frequency bias.

toward different directions each accounting for roughly 10% of the answers. For room-identification questions, the answer distribution is led by `Room 1` and `Room 2`, which account for 33.2% and 27.5% respectively, while the remaining rooms are more evenly distributed. For object-location questions, different location categories are all distributed at roughly 2%, suggesting that no single location prior dominates the benchmark. Finally, object categories are also diverse. Although *TV*, *lamp*, and *refrigerator* are the most frequent categories, each accounts for only about 3% of all object mentions. Overall, these statistics suggest that `ChainSpace-Bench` is not dominated by a few trivial priors, but instead provides a relatively diverse and balanced testbed for chained spatial reasoning.

A.2 More Benchmark Results

Table 8: Performance on the enlarged ChainSpace-Bench containing 800 videos and 3,200 chains. The red and blue denote the best and second-best results, respectively.

Model	Avg. Question Accuracy	Chain-Aware Score
Qwen3-VL-8B-Instruct [3]	0.3348	0.1531
SenseNova-SI-8B [8]	0.3427	0.1562
ChainSpace-Pipeline-8B (Ours)	**0.3546**	**0.1690**

Our advantage remains stable. We further expand the benchmark from 326 to **800 videos** and from 1,304 to **3,200 chains**, corresponding to a **145.4% increase** in evaluation scale. As shown

in Table 8, the enlarged benchmark preserves the main performance trend. ChainSpace-Pipeline-8B consistently outperforms both Qwen3-VL-8B-Instruct [3] and SenseNova-SI-8B [8] in terms of average question accuracy and Chain-Aware Score. In particular, the Chain-Aware Score of ChainSpace-Pipeline-8B changes only marginally from 0.1687 on the original benchmark to 0.1690 on the enlarged benchmark. These results indicate that the observed advantage remains stable under a substantially larger evaluation set and is unlikely to be an artifact of the original benchmark scale.

Table 9: Task-level analysis on external spatial reasoning benchmarks. ChainSpace-Pipeline-8B yields larger gains on compositional, viewpoint-conditioned, and planning-related spatial tasks.

Benchmark	Qwen3-VL-8B-Instruct [3]	ChainSpace-Pipeline-8B (Ours)	Representative Task-Level Gains
VSI-Bench [41]	57.90	**58.82**	Route Planning $+3.56$; Absolute Distance $+1.54$; Relative Distance $+1.30$; Appearance Order $+1.18$
MMSI-Bench [43]	31.10	**31.80**	Camera–Camera Relation $+6.41$; Attribute Measurement $+6.25$; Appearance $+2.98$; Multi-step Spatial Reasoning $+2.02$
ViewSpatial-Bench [20]	42.20	**42.96**	Person-perspective Object Orientation $+3.30$; Scene Simulation $+0.85$
3DSRBench (Circ.) [23]	53.30	**54.10**	Viewpoint-to-Object Relation $+2.64$; Next-to Relation $+1.79$; In-front-of Relation $+1.15$

Task-Level Analysis on External Spatial Benchmarks. To better understand where the improvement comes from, we further analyze the task-level performance on four external spatial reasoning benchmarks. As shown in Table 9, the improvements are substantially larger on several higher-order spatial tasks. On VSI-Bench [41], ChainSpace-Pipeline improves 7 of 8 task categories, with the largest gain on route planning ($+3.56$), followed by absolute distance ($+1.54$), relative distance ($+1.30$), and appearance order ($+1.18$). On MMSI-Bench, the largest improvements occur in camera–camera relation ($+6.41$), attribute measurement ($+6.25$), appearance reasoning ($+2.98$), and multi-step spatial reasoning ($+2.02$). Similar trends are observed on ViewSpatial-Bench [20] and 3DSRBench [23], where ChainSpace-Pipeline particularly improves perspective-conditioned object orientation, scene simulation, and viewpoint-to-object relations. Overall, the gains are concentrated on tasks that require relational composition, reference-frame transformation, multi-step reasoning, or planning, which closely aligns with the capabilities targeted by chained spatial supervision.

Table 10: Comparison of Qwen3-VL-8B-Instruct [3] trained with different data source under Chance-Adjusted Accuracy (CAA, %) on the SITE-Bench [36]. The text in red and blue denotes the best and second-best results.

Model	Overall	Cnt./Exist.	Obj. Loc.	3D Info.	MV/XImg.	Spatial Rel.	Move/Nav.
Baseline	43.54	55.99	49.11	56.00	23.05	66.13	21.32
SenseNova-SI-150K [8]	45.35	57.36	51.91	**58.73**	24.75	68.45	19.51
Ours-50K	43.94	56.97	47.07	56.78	23.52	66.79	**23.74**
Ours-50K+SenseNova-SI-100K [8]	**45.51**	**58.74**	**58.85**	58.34	**25.78**	**68.61**	18.90

Abbreviations: Cnt./Exist. = counting & existence; Obj. Loc. = object localization & positioning; 3D Info. = 3D information understanding; MV/XImg. = multi-view & cross-image reasoning; Spatial Rel. = spatial relationship reasoning; Move/Nav. = movement prediction & navigation.

Results on SITE-Bench [36] and 3DSRBench [23]. We further evaluate our method on two additional spatial intelligence benchmarks that are distributionally distinct from our simulator-generated training data. SITE-Bench [36] is a broad spatial intelligence benchmark covering multiple sub-skills, including counting and existence, object localization, 3D information understanding, multi-view reasoning, spatial relations, and movement/navigation. 3DSRBench [23] focuses on 3D spatial reasoning under several transformed evaluation settings that test robustness to viewpoint and orientation changes.

Table 10 shows that our chained supervision transfers well to the real-world benchmark with diverse spatial sub-tasks. Although our model is trained only on simulator-generated data, it remains

Table 11: Performance comparison on 3DSRBench [23] under different evaluation settings. Bold indicates the better result within each model group.

Model	Method	Vanilla	Flip	Circ	Flip+Circ
Qwen3-VL-8B-Instruct [3]	Baseline	**0.566**	0.498	0.533	0.475
	Ours-50K	0.564	**0.499**	**0.541**	**0.483**
InternVL3-8B [49]	Baseline	0.524	0.444	0.396	0.324
	Ours-150K	**0.541**	**0.466**	**0.447**	**0.384**

Abbreviations: *Vanilla* denotes the original evaluation setting; *Flip* applies image flipping; *Circ* applies circular rotation; *Flip+Circ* applies both transformations together.

competitive with SenseNova-SI [8], which is trained on substantially more real-data supervision. More importantly, our method shows a clear advantage on *Move/Nav.*, where **Ours-50K** achieves the best score. This is particularly consistent with our design, since chained supervision explicitly encourages preserving and updating spatial state across dependent steps. Moreover, combining our data with SenseNova-SI-100K further improves the overall score and achieves the best results on several SITE sub-categories, suggesting that **our supervision can be complementary to existing real-data pipelines**. Table 11 further supports the robustness of our learned spatial reasoning under geometric transformations. For Qwen3-VL-8B [3], our method improves three out of four evaluation settings, especially under *Circ* and *Flip+Circ*. For InternVL3-8B [49], the gains are consistent across all four settings. Overall, these results provide additional evidence that the proposed chained design principle can transfer beyond the simulator-generated training distribution and improve several aspects of spatial reasoning.

A.3 Scaling Results

Scaling Behavior. As shown in Table 12, **(1)** *the chain-level gains emerge early and become reliable from 50K samples onward.* At 25K, both *Full-chain Exact Match* and *Avg. Chain-Aware Score* already improve over the untuned baseline, and from 50K onward both metrics become consistently significant, showing that relatively small amounts of chained supervision are sufficient to induce stable improvement. **(2)** *the early saturation is consistent with our simulator-to-real setting.* Since training is performed on synthetic trajectories while evaluation is conducted on real-world videos, additional synthetic data is naturally subject to a simulator-to-real domain gap, so gains may saturate once the core chained reasoning behavior has been learned [33].

Table 12: Scaling behavior of chained supervision on `ChainSpace-Bench`. We report the paired improvement Δ over the same untuned baseline under different training data scales. Each entry is shown as Δ followed by the paired-bootstrap p-value in parentheses. * indicates statistical significance ($p < 0.05$ and 95% CI excluding 0). The text in red and blue denotes the best and second-best results.

Metric	25K	50K	75K	125K
Full-chain Exact Match	+0.0038 (0.3900)	**+0.0153*** (0.0044)	+0.0130* (0.0220)	**+0.0153*** (0.0153)
Avg. Chain-Aware Score	+0.0038 (0.2740)	**+0.0165*** (<0.0001)	+0.0141* (0.0032)	+0.0141* (0.0048)

A.4 More Qualitative Results

Qualitative comparison. Fig. 5 shows representative examples on `ChainSpace-Bench`. SenseNova-SI [8] answers some individual turns correctly, such as the room sequence or earliest object, but fails on intermediate or later questions that require using previously established spatial constraints. In contrast, our model answers all rounds correctly, suggesting more coherent use of accumulated spatial state.

Robustness to counterfactual spatial queries. Fig. 6 tests whether the model maintains a spatial state beyond the originally queried relation. While both models answer the original route-planning chain correctly, changing the queried room pair in Question (2) causes SenseNova-SI [8] to fail. Our model remains correct under the changed query, suggesting that `ChainSpace-Pipeline` encourages a more complete and query-flexible spatial representation.

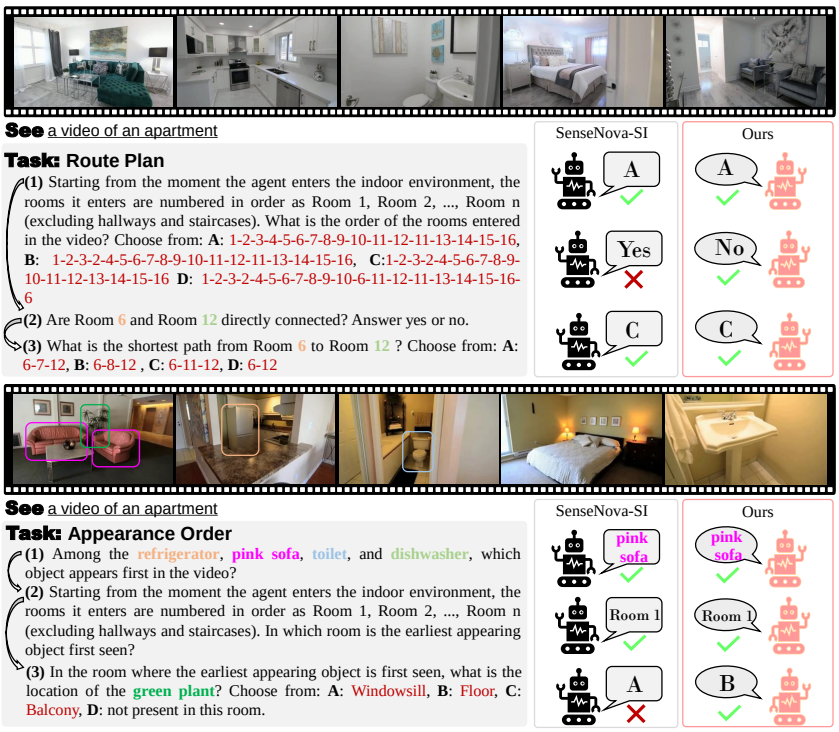


Figure 5: Visualization of qualitative examples on `ChainSpace-Bench`. SenseNova-SI [8] answers some individual turns correctly, whereas our model maintains consistent answers across all rounds.

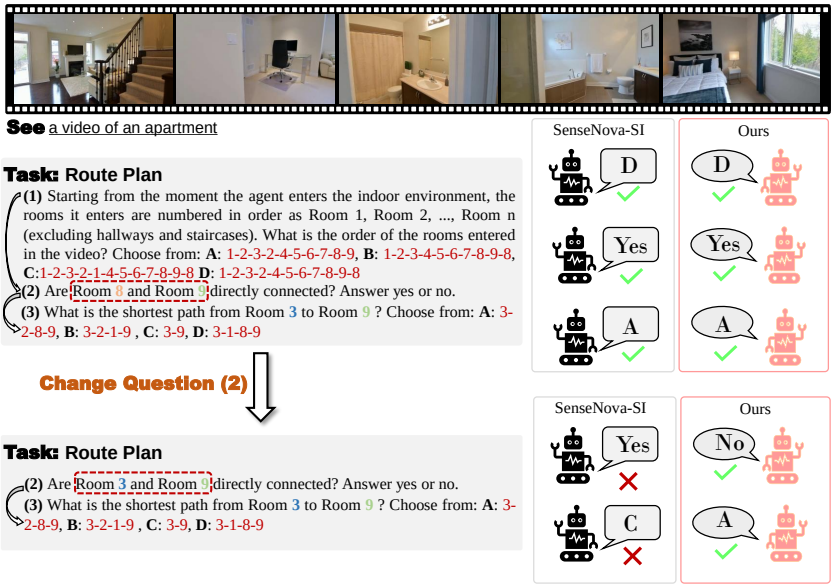
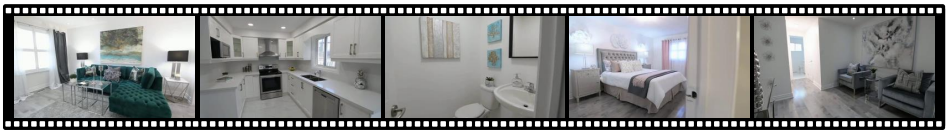


Figure 6: Visualization of counterfactual query examples. With the same video context but a changed room-pair question, our model remains correct while SenseNova-SI [8] fails, suggesting a more complete maintained spatial state.

A.5 Failure Analysis

Fig. 7 shows a long-horizon failure case that still supports the effectiveness of our method. The model fails to recover the full room visitation order in Question (1), likely because long trajectories place a

heavy burden on the long-context memory of large models [38]. However, it still correctly maintains the spatial states of the earlier queried rooms (Room 2 and Room 4), which enables correct answers to the subsequent connectivity and shortest-path questions. This suggests that `ChainSpace-Pipeline` is effective in encouraging the model to preserve spatial states that are relevant for later reasoning, even when the backbone struggles with complete global route memory.



See a video of an apartment

Task: Route Plan

(1) Starting from the moment the agent enters the indoor environment, the rooms it enters are numbered in order as Room 1, Room 2, ..., Room n (excluding hallways and staircases). What is the order of the rooms entered in the video? Choose from: **A**: 1-2-3-2-4-5-6-7-6-8-2-9-10-11-12-13-14-15-16-17-18-17-19-20-21-22-23-24-25-26-27-26-28-26, **B**: 1-2-3-2-4-5-6-7-8-2-9-10-11-12-13-14-15-16-17-18-17-19-20-21-22-23-24-25-26-27-26-28-26, **C**:1-2-3-2-4-5-6-7-6-8-2-9-10-11-12-11-13-14-15-16-17-18-17-19-20-21-22-23-24-25-26-27-26-28-26 **D**: 1-2-3-2-4-5-6-7-6-8-2-9-10-11-12-11-13-14-15-16-17-18-17-19-20-21-22-23-24-25-26-27-26-28-26-1

(2) Are Room 2 and Room 4 directly connected? Answer yes or no.

(3) What is the shortest path from Room 2 to Room 4 ? Choose from: **A**: 2-3-4, **B**: 2-1-3-4 , **C**: 2-1-4, **D**: 2-4

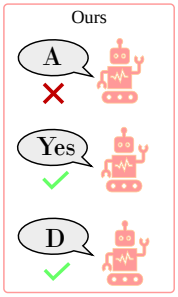


Figure 7: Visualization of a representative failure case. The model confuses the complete room order in a many-room trajectory, while preserving correct local spatial relations for later reasoning.

A.6 Discussions

(1) Why we do not perform ablations such as reducing the chain from 3 rounds to 2 rounds, or shuffling the question order within each chain. The reason is that these manipulations do not constitute clean controlled-variable ablations in our setting. Our benchmark and training pipeline are built on chained logical dependency, *i.e.*, later questions are defined with respect to the spatial constraints established by earlier ones. As a result, changing the number of rounds or permuting the question order does not merely remove one design factor while keeping the task unchanged. Instead, it changes the task semantics and the data distribution themselves. For example, removing one round may collapse a multi-step reasoning chain into a different and often easier task, while shuffling the order can break the intended dependency structure and even make some questions ill-posed. Therefore, such experiments would not isolate the effect of chained supervision in a fair way. For this reason, we choose to support our central claim through results in Table 1, Table 2, and Table 3. In particular, the oracle-history intervention preserves the original chained task semantics while directly probing whether later-round predictions benefit from correct preceding spatial states.

(2) Why we evaluate the proposed principle only under full-parameter supervised fine-tuning (SFT), rather than further introducing reinforcement learning (RL). The main reason is that the goal of this paper is to study whether chained supervision improves the model's *base spatial intelligence*. In prior practice, improvements in base capability are typically established primarily through supervised learning, while RL is more often used as a subsequent policy refinement stage [10]. Under this perspective, SFT provides the cleanest setting for testing whether the proposed chained design principle itself yields a stronger spatial reasoning capability. A second reason is that RL is less straightforward in our setting. If RL is applied directly to model responses, the reward must largely be designed around whether the answer sequence preserves cross-question logical consistency. However, logical consistency alone does not guarantee genuine spatial intelligence. A model may produce answers that are mutually compatible as a reasoning chain, yet still be misaligned with the actual spatial structure in the video. Therefore, we choose to first isolate and validate the chained design principle through supervised learning. We believe this provides a cleaner foundation for future work on RL, where the same principle could be combined with more carefully grounded rewards to further refine spatial reasoning policies.

(3) Why multi-round QA instead of asking all questions at once. We use a multi-round QA format because our goal is to evaluate spatial reasoning as a process of progressively establishing

and reusing spatial states, rather than as a static joint-QA problem. If all questions are presented simultaneously, the model can see the entire question set before answering, including future queries that may reveal which objects, rooms, or relations are important. This may allow the model to answer each question independently or use later questions as hints to narrow down earlier reasoning, without explicitly preserving an intermediate spatial state. In contrast, the multi-round format exposes the dependency structure more directly, *i.e.*, each later question is conditioned on the spatial constraints established in previous turns, making it possible to evaluate whether the model can maintain and reuse them across the chain. This design also enables task-preserving analyses such as oracle-history intervention, where previous model answers can be replaced by ground-truth answers to test whether later predictions benefit from correct preceding spatial states. Therefore, the multi-round design is not a superficial prompting choice, but an essential part of our benchmark for evaluating persistent spatial reasoning.

(4) Discussion on the simulator-based data pipelines and Limitations. The main contribution of this work is the *chained design principle* rather than simulation itself. The results in the manuscript already show that this principle is effective even in the simulator-based setting. However, the simulated training data inevitably introduces a simulator-to-real domain gap when evaluating on real-world videos and existing spatial intelligence benchmarks [33]. Therefore, the gains on benchmarks should not be interpreted as the upper bound of chained supervision, but rather as evidence that the proposed principle can transfer beyond the synthetic training distribution. We expect larger and more consistent improvements when the chained construction principle can be applied to large-scale real-world trajectories, where the visual distribution, object arrangements, and navigation patterns are closer to practical deployment scenarios.

Meanwhile, the results in Table 10 also suggest that our simulator-generated chained data can serve as an effective complement to existing real-data pipelines. Even though the training data are fully synthetic, when combined with existing real-data supervision, they further improve performance. This indicates that the structured signal provided by our pipeline captures information that is not fully covered by current independent-QA real datasets, and therefore offers a useful supplementary source of supervision for spatial intelligence learning.

Impact Statement

This paper presents work whose goal is to advance the field of Spatial Intelligence. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section A.6

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The dataset and code are currently undergoing the company's approval process and will be released upon approval.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 6.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section A.6

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 6.1

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.