

A Progressive Training Strategy for Embodied Vision-Language Models to Mitigate Spatio-Temporal Hallucinations

Xiaoda Yang*
Zhejiang University

Shuai Yang*
The Hong Kong University of Science and Technology

Can Wang
Qingdao University

Jingyang Xue
Qingdao University

Menglan Tang
Qingdao University

Checheng Yu
The University of Hong Kong

Xunzhe Zhou
The University of Hong Kong

Sashuai Zhou
Zhejiang University

Tao Jin
Zhejiang University

Lixin Yang
Shanghai Jiao Tong University

Xiangyu Yue
The Chinese University of Hong Kong

Zhou Zhao†
Zhejiang University

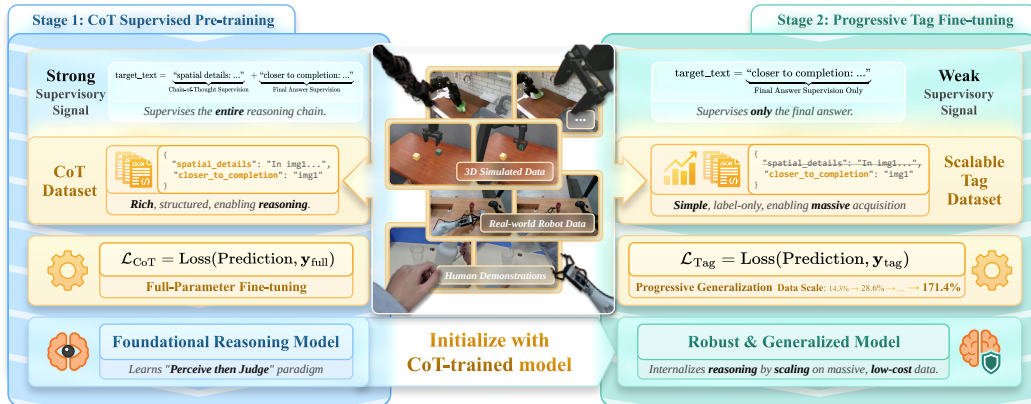


Figure 1: Our progressive training paradigm to mitigate spatio-temporal hallucinations. **Stage 1 (Left):** A CoT-Supervised Pre-training stage instills state-based reasoning by supervising the entire reasoning chain. **Stage 2 (Right):** A Weakly-Supervised Fine-tuning stage scales this ability using a massive, tag-only dataset, showing a positive scaling trend.

Abstract

Vision-Language Models (VLMs) have made significant strides in static image understanding but continue to face critical hurdles in spatiotemporal reasoning. A major bottleneck is “multi-image reasoning hallucination”, where a large performance drop between forward and reverse temporal queries reveals a dependence

*Equal contribution.

†Corresponding author: zhaozhou@zju.edu.cn

on superficial shortcuts instead of state-based understanding. To mitigate this, we first develop a new Chain-of-Thought (CoT) dataset that decomposes intricate reasoning into detailed spatiotemporal steps and definitive judgments. Building on this, we present a progressive training framework: it initiates with supervised pre-training on our CoT dataset to instill logical structures, followed by fine-tuning with scalable weakly-labeled data for broader generalization. Our experiments demonstrate that this approach not only improves backbone accuracy but also reduces the forward-backward performance gap from over 70% to only 6.53%. This shows that the method strengthens dynamic reasoning and reduces the inherent temporal biases of current VLMs.

1 Introduction

Vision-Language Models (VLMs) Zhang et al. [2024a] have achieved milestone progress in understanding static images, robustly aligning language with visual perception and laying a solid foundation for a wide range of downstream tasks. However, a critical vulnerability emerges when these models are deployed in dynamic, task-oriented environments: when handling multi-image comparison tasks, they tend to improperly generalize the temporal-logical priors learned during video pre-training. This leads to erroneous judgments of dynamic spatial relationships and hallucinations, particularly in manipulation scenarios. In this context, while formulated as a simple comparison, this task proxies reward modeling by requiring an understanding of physical state transitions rather than mere visual matching.

To address these spatial hallucinations, this work proposes a progressive training framework guided by Chain-of-Thought (CoT) Wei et al. [2022]. The cornerstone of our approach is a novel, large-scale CoT dataset constructed from videos of daily operational tasks. We first temporally downsample these videos by a factor of ten and then annotate them with detailed reasoning processes. Each data point provides not only the final answer but also a structured analysis of spatial nuances, with this analysis constituting the core supervisory signal. This supervisory approach explicitly teaches the model *how* to reason, thereby establishing a solid foundation for subsequent generalization.

Building on this dataset, we introduce a two-stage, full-parameter fine-tuning paradigm analogous to a student’s learning trajectory: first mastering core concepts, then internalizing knowledge through rigorous practice. The first stage, **CoT-Supervised Pre-training**, fully fine-tunes the VLM (Vision-Language Model) on our CoT dataset. By supervising the complete reasoning chain (both spatial details and the final `closer to completion` label), the model is encouraged to adopt a “perceive-then-judge” cognitive paradigm, thereby building a robust knowledge framework. Subsequently, the curriculum transitions to the second stage: **Weakly-Supervised Fine-tuning**. At this stage, only the final labels are used to fine-tune the model. Unlike the meticulously crafted CoT data, the data for this stage is easy to acquire, as nearly all operational videos inherently contain the `closer to completion` supervisory signal, providing a foundation for massive-scale data augmentation. By removing the explicit reasoning steps, this stage encourages the model to leverage the foundational cognitive paradigm learned in Stage 1 to arrive at its final judgment. This “study-then-practice” progression ensures the model grounds its conclusions in visual evidence, ultimately cultivating an internalized, transferable reasoning ability rather than one based on rote memorization.

Our experiments yield strong results. The CoT-only model achieves an average accuracy of 87.07% on the test set. More importantly, it demonstrates strong robustness against sequence bias, with the performance gap between forward and reverse-ordered tasks being reduced to 6.53%. This mitigates the severe performance disparity in baseline models, indicating that spatial hallucinations are reduced and the model has transitioned from relying on sequential heuristics to more robust state-based temporal reasoning. Our main contributions are:

- **Large-scale Chain-of-Thought (CoT) dataset:** We construct a novel dataset, STCR-CoT, comprising over 30 million fine-grained spatial annotations derived from real-world operational tasks.
- **Low-cost and scalable data engine:** We propose a data generation engine that leverages temporal progress relationships in video. By swapping frame orders and applying multi-scale sampling, it generates massive amounts of weakly-supervised (tag-only) data at a minimal cost while reducing annotation noise through consistency filtering.

- **Progressive training paradigm:** We introduce a training paradigm that transitions from strong supervision on our CoT dataset to weak supervision on the scalable, tag-only data. Through this, we empirically observe a positive scaling trend, showing that model performance generally improves as the volume of weakly-supervised data increases.
- **Effective mitigation of temporal bias and a novel reward modeling approach:** Through extensive experiments, we demonstrate that our paradigm is effective. It mitigates sequence bias in multi-image spatial reasoning by reducing the forward-reverse performance gap to 6.53%, and the resulting model can also accurately assess task progression.

2 Related Work

2.1 Benchmarking Spatio-Temporal Reasoning in Vision-Language Models

Vision-Language Models (VLMs) have evolved from understanding static images to evaluating complex, dynamic scenes. Early benchmarks like CLEVR Johnson et al. [2017] focused on compositional geometry in synthetic images, while subsequent work like GQA Hudson and Manning [2019] extended this to complex real-world imagery. Focus later shifted towards 3D and dynamic environments. Datasets like ScanNet Dai et al. [2017] and Matterport3D Chang et al. [2017] pioneered 3D understanding, while interactive platforms such as Habitat Savva et al. [2019] and AI2-THOR Kolve et al. [2017] marked a shift to dynamic tasks requiring navigation and decision-making. Building on this, research emerged to endow VLMs with native 3D perception, leading to models like 3D-LLM, LL3DA, and Chat-Scene. This trend was further advanced by models including 3D-LLaVA and Video-3D LLM, which integrate 3D perception with LLM reasoning. Despite these advances, diagnostic tools like the Perception Test Patraucean et al. [2023] indicate that even state-of-the-art models often lack a deep temporal understanding of events.

2.2 Task Progress Estimation and Reward Model

Evaluating task progress is crucial, leading to various methods for constructing general-purpose reward models. One approach learns progress representations via time-contrastive learning, where works like Ma et al. [2022] leverage the temporal order of video frames as a self-supervised signal. Another approach utilizes the zero-shot capabilities of VLMs, using prompts to score or compare states Team et al. [2024]. Furthermore, other studies measure progress via semantic similarity, mapping states and goals to a shared embedding space Reuss et al. [2023], or by synthesizing target states for comparison Bruce et al. [2024]. These signals frequently provide feedback for Reinforcement Learning Baumli et al. [2024].

2.3 Chain-of-Thought for Multimodal Reasoning

Chain-of-Thought (CoT) enhances LLM reasoning by prompting models to generate intermediate steps before the final answer Wei et al. [2022]. This method effectively handles arithmetic, commonsense, and symbolic reasoning. Recently, CoT has been extended to the multimodal domain for complex visual tasks. Most multimodal CoT applications operate at inference-time using structured prompts to guide generation Zhang et al. [2024b], Rose et al. [2024].

3 Dataset

A key vulnerability of vision-language models (VLMs) is their strong sensitivity to input order during spatio-temporal reasoning. This sensitivity manifests as severe hallucinations, particularly when the logical progression of events contradicts the visual sequence presented to the model. For instance, when tasked to determine which of two images is closer to task completion, merely reversing the input order can cause a large failure in reasoning, a phenomenon termed “multi-image reasoning hallucination”. A key factor is supervisory misalignment: while modern VLMs possess the latent capacity for spatial reasoning, standard image-caption training encourages them to rely on superficial temporal shortcuts (e.g., file order) rather than state-change analysis. These paradigms typically rely exclusively on sparse, end-to-end labels. Such weak supervision is insufficient to counteract the strong “sequence bias” inherent from large-scale pre-training. Consequently, the model defaults to an

erroneous prior: associating task progression with the second image by default, rather than reasoning from visual evidence. To break this vicious cycle, a dataset with a novel supervisory structure is crucial. Our goal is to replace this inadequate sparse supervision with a dense, structured signal that encourages the model to follow a verifiable reasoning process, thereby decoupling its judgment from superficial positional cues.

3.1 Preprocessing and Sampling

To construct reliable training samples, our data processing pipeline begins with the refinement of the AgiBot AgiBot-World-Contributors et al. [2025] real-world robot video corpus. First, we employ a multi-granularity sampling method, applying 10x temporal downsampling to filter out visual noise (such as camera shake) and ensure significant state changes between frames. This process encourages the model to move beyond simple pattern matching. Second, we utilize multi-granularity sliding windows (with intervals of 5, 10, 12 frames, etc.) to extract image pairs. This multi-granularity design creates a natural difficulty curriculum: smaller windows test the model’s fine-grained visual discrimination capabilities, whereas larger windows demand high-level logical inference. Building upon this, we perform an important optimization step. To reduce potential data noise and improve training efficiency, we programmatically trim redundant operations from the end of many videos that are irrelevant to the core task (e.g., the robotic arm resetting). This refinement process keeps the selected image pairs focused on task progression, thereby laying a solid foundation for subsequent annotation.

The structured CoT annotations are generated by a dedicated data engine rather than by direct manual labeling at scale, which would be infeasible for 34.7M samples. For each sampled pair, we use a structured prompt to guide Seed1.6-Pro to produce pseudo-labels for fine-grained spatial differences between high-frequency frames, including the intermediate spatial details used by our first-stage supervision. We then apply a bidirectional consistency filter: the same video segment is queried in both forward and reverse order, and samples that lead a strong model to inconsistent judgments are removed. Finally, we randomly sample 10,000 generated CoT examples for cross-human inspection, checking both the final task-progress judgment and the spatial reasoning chain. This quality-control pipeline is used before injecting CoT annotations into Stage 1 training. To reduce train-test coupling, we assign source episodes to a single split before sliding-window expansion; forward/reverse variants and overlapping windows are therefore kept within the same split rather than shared across training and evaluation. Although Seed1.6-Pro is used as the CoT teacher, evaluation is conducted with held-out task-progress labels and includes Seed and non-Seed model families, so the Seed1.6 baseline is used as a reference point rather than as the source of the test labels.

3.2 Forward and Inverse Contrasting for Decoupling Order Bias

We organize the data by pairing forward and inverse instances of the same example. Since the visual and semantic context is largely held constant, the model must focus on the single variable of input order: a shallow heuristic learned from one sample (e.g., “select the second image”) is immediately contradicted by its inverse counterpart. This counterbalancing discourages reliance on positional priors and repeatedly guides the model to inspect visual state changes and identify a judgment criterion valid under both sequences. In practice, this makes each paired example serve as a local consistency check: the correct answer changes only when the queried order changes, while the underlying evidence still has to be explained by object states, contact relations, and task progress. Through this design, we encourage a more general state-comparison capability rather than shortcuts tailored to a specific ordering pattern.

3.3 Data Statistics

The large-scale Spatio-Temporal Causal Reasoning Chain-of-Thought (STCR-CoT) dataset constructed in this study comprises a total of 34.7 million filtered samples. We have implemented a balanced design of forward and reverse samples at a global level. As illustrated in Figure 2(a), the dataset ensures that forward and reverse sequences each account for 50% through meticulous sequence construction. This global balancing design directly targets the issue of temporal bias.

The dataset is composed of two parts: (1) Chain-of-Thought (CoT) Samples: These provide strong supervision signals. Each sample includes an image pair, a final answer label, and a structured

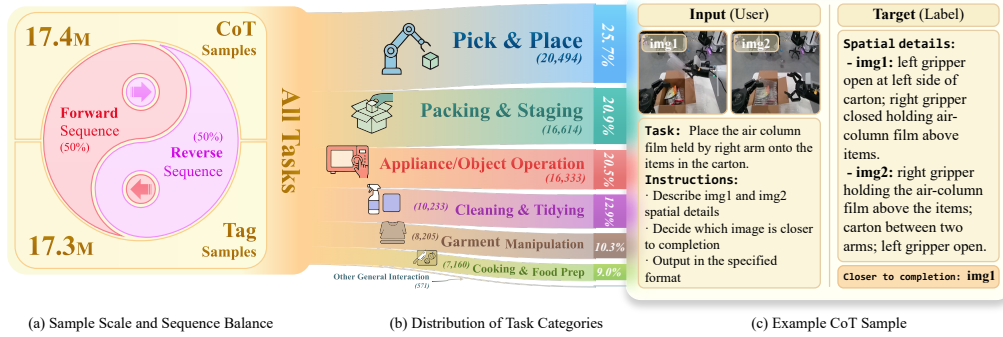


Figure 2: **Statistical analysis of the STCR-CoT dataset.** (a) **Sample Scale and Sequence Balance:** The dataset reaches a total scale of 34.7 million samples and features a globally balanced design, with forward and reverse sequences each constituting 50% to eliminate temporal bias. (b) **Distribution of Task Categories:** The distribution of task categories is diverse, covering 7 major categories including “Pick & Place” and “Packing & Staging”, across 79,610 unique task instances. (c) **Window Size Distribution:** A progressive learning curriculum is established using a multi-scale sliding window strategy (5-16 frames), with difficulty advancing from fine-grained visual discrimination (smaller windows) to macro-level logical inference (larger windows).

reasoning chain with spatial details. (2) Tag Samples: These offer weak supervision signals. The data is obtained from numerous manipulation videos, with each sample containing only an image pair and its corresponding binary classification label.

The dataset covers 79,610 unique task instances, exhibiting a well-diversified category distribution as shown in Figure 2(b): Pick and Place (25.7%), Packing and Staging (20.9%), Appliance/Object Manipulation (20.5%), Cleaning and Tidying (12.9%), Garment Manipulation (10.3%), Cooking and Food Preparation (9.0%), and others.

Furthermore, a progressive difficulty curriculum is constructed using a multi-scale sliding window sampling strategy (5–16 frames), as depicted in Figure 2(c). Samples from smaller windows (5–9 frames) focus on fine-grained visual discrimination, while those from larger windows (12–16 frames) emphasize macro-level reasoning. The distribution, where the number of samples increases with the window size, provides a systematic learning path for the model.

In summary, with its substantial scale (34.7M samples), global forward-reverse balancing design, diverse task coverage (7 major categories, 79K unique tasks), and multi-scale curriculum planning, the STCR-CoT dataset provides a strong data foundation for training Vision-Language Models with robust spatio-temporal reasoning capabilities. This holistic balancing design helps the model learn reasoning patterns that are less dependent on the input sequence order.

4 Method

In end-to-end training, large models tend to learn superficial statistical shortcuts rather than a structured cognitive process that proceeds from precise perception to logical judgment. As a result, their decision-making lacks a solid evidential foundation and causes “hallucinations”. To address this challenge, we propose a progressive training paradigm based on Chain-of-Thought (CoT). This paradigm emulates the cognitive progression of a student who first intensively studies core knowledge and then engages in extensive practice for deep comprehension and internalization.

4.1 CoT-Driven Spatial Perception Pre-training: From Shortcut Learning to Structural Grounding

The key vulnerability of Vision-Language Models (VLMs) in spatiotemporal reasoning lies in **Shortcut Learning**. When optimized via sparse end-to-end labels, models tend to converge on “statistical coincidences”—such as the sequential order of input frames—rather than learning visual state transitions. To address this, Stage 1 of our paradigm treats Chain-of-Thought (CoT) as more than a data format: it serves as a **Structural Bottleneck** that reshapes the model’s optimization trajectory.

The “Perceive-then-Judge” Cognitive Scaffold. In this phase, we treat the model as a novice required to master a “core textbook” of spatial relationships. By extending the supervisory signal from the sparse final judgment to the dense intermediate reasoning chain (the `spatial details`), we guide the model toward a more structured internal computation. Instead of learning a direct, high-entropy mapping from pixels to answers, the model is encouraged to follow a **unidirectional information flow** within its computational graph:

1. **The Perception Constraint:** The model must first compress raw visual features into verifiable spatial attributes (e.g., object contact, relative distance, and orientation).
2. **The Logical Derivation:** The final judgment must be derived exclusively from these extracted attributes, acting as a secondary inference layer.

This mechanism acts as a **cognitive scaffold**. We use this term operationally: the training target requires intermediate spatial attributes before the final task-progress judgment, rather than only a direct class label. In terms of the model’s training dynamics, this dense supervision encourages the gradients for the “judgment” logic to depend on the extracted visual evidence. Consequently, the easier path for loss minimization is less likely to bypass visual perception.

Countering the Incentive for Shortcut Discovery. A critical risk in large-scale training is **Incentive Conflict**: if a model can achieve low training loss by exploiting a brittle correlation (e.g., associating the second image position with task completion), it may ignore complex state comparisons. Our progressive strategy mitigates this by front-loading the “rigorous reasoning” objective. By the time the model encounters larger, weakly-labeled datasets in Stage 2, its representation has already been optimized under the perceive-then-judge target. This **structural initialization** makes the model more resistant to erroneous order priors present in raw video data.

4.2 Scaling Trend for a Scalable Training Paradigm

After the model masters fine-grained reasoning ability through pre-training, it transitions to a weakly-supervised fine-tuning stage. This stage simulates a student moving from teacher-guided study to independent, extensive practice. We initialize this stage with the model from the first phase and fine-tune it on a large-scale, diverse database containing only final-answer labels (e.g., labels indicating “which state is closer to task completion”). By removing supervision on the intermediate steps, we encourage the model to rely on the reasoning pathways learned during pre-training. This entire approach is strategically designed to leverage a crucial advantage for scalability: data requiring only final labels is abundant and easy to acquire. Our experiments reveal a clear positive trend: as the volume of weakly-supervised data increases, model performance generally improves. This finding demonstrates the scaling potential of our training paradigm.

If the CoT pre-training stage is akin to studying a “textbook with answers and detailed explanations” to build a robust cognitive “framework” that is harder to bypass with shortcuts, then this stage is analogous to practicing with a “massive problem bank with answers only” to achieve deeper comprehension. Throughout this process, our progressive fine-tuning strategy plays a guiding role. It encourages the model to apply and consolidate this internal structure in every judgment, thereby efficiently “distilling” structured CoT supervision and generalizing it onto the inexpensive, abundant weakly-supervised data.

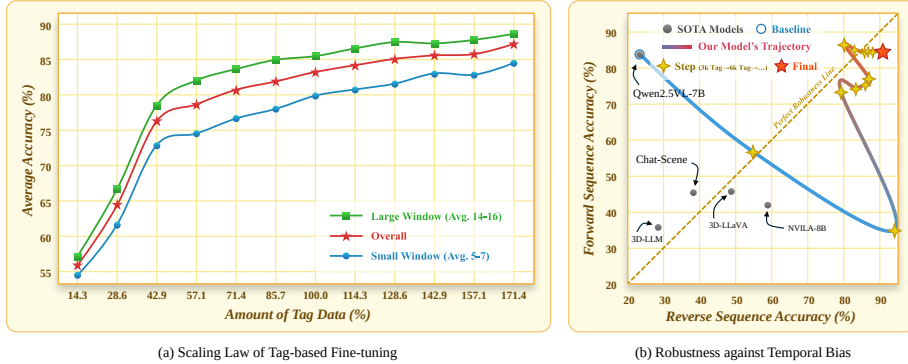


Figure 3: **Main results of our training paradigm.** (a) **Scaling Trend of Tag-based Fine-tuning:** Model accuracy exhibits a clear positive correlation with the amount of weakly-supervised data. This shows that our paradigm effectively benefits from scaling up in this setting. (b) **Robustness against Temporal Bias:** Our model’s trajectory (purple/blue line) demonstrates significantly improved robustness, moving closer to the perfect robustness diagonal compared to the baselines.

5 Experiments

5.1 Evaluation metrics

We introduce a large-scale evaluation system for daily-life operations, comprising 35,660 independent test episodes across 29 high-level task categories. The benchmark spans supermarket and retail scenarios (over 7,400 instances, e.g., picking items, packing, and barcode scanning), daily household activities (over 10,000 instances, e.g., dishwashing, ironing, sorting, and folding), kitchen and cooking tasks, and warehousing/logistics operations (over 5,800 instances). These categories cover multi-object interaction, fine-grained manipulation, procedural reasoning, and industrial task-progress judgment, providing a broad testbed for forward/reverse temporal robustness. We also use MMSI-Bench Yang et al. [2026], a 1,000-question multi-image VQA benchmark with step-by-step reasoning annotations, to evaluate whether the trained model transfers to broader spatial reasoning.

5.2 Implementation details

In the TAG (label-only) phase, we utilize the TAG weakly supervised dataset, sampled to varying proportions of the total CoT dataset volume ($\{\frac{i}{7} \mid i \in \{1, 2, \dots\}\}$). The entire training process uniformly employs the AdamW optimizer Loshchilov and Hutter [2019], with a carefully configured learning rate scheduling strategy: we set an initial learning rate of $2e-7$ and use a Cosine Learning Rate Scheduler Loshchilov and Hutter [2017] to smoothly adjust the learning rate. To optimize training efficiency and effectively handle the large-scale model and data, the batch size is set to 8, combined with a 2-step Gradient Accumulation Goyal et al. [2018] strategy, which is computationally equivalent to a larger effective batch size. Furthermore, to ensure training stability and accelerate convergence, a 1000-step warmup phase is implemented. To improve training speed and reduce memory footprint, BF16 mixed precision is enabled throughout the entire training process.

Stage 1 uses teacher-forced autoregressive supervision over a fixed response schema consisting of spatial details followed by the final closer to completion answer. Stage 2 removes the rationale tokens and supervises only the final binary progress label. All evaluated models use the same image order, prompt template, and deterministic decoding protocol for forward and reverse queries.

Our main two-stage results are obtained with full-parameter fine-tuning in both Stage 1 and Stage 2. LoRA Hu et al. [2021] is used only in the preliminary backbone-selection study: we first tune several candidate VLM backbones under comparable small-scale LoRA settings to estimate their potential and optimization behavior, and then select Qwen2.5-VL-7B for the large-scale full-parameter training

Table 1: Evaluation results across different WinSize and Models

Models	WinSize							
	5	6	7	8	9	10	11	≥ 12
Human Perception								
Human Perception	97.3	97.3	98.2	99.1	99.1	100.0	100.0	100.0
Proprietary Models								
GPT-o4mini OpenAI [2025]	70.0	70.3	72.8	67.7	60.0	67.5	67.7	70.3
GPT-o4mini-high OpenAI et al. [2024]	70.5	72.7	71.8	72.3	75.5	77.3	78.1	78.6
GPT-4o OpenAI et al. [2024]	71.1	71.7	76.7	77.8	79.2	79.2	79.2	79.2
GPT-o3 OpenAI [2025]	67.2	61.4	59.6	61.6	59.6	64.0	62.0	66.6
Gemini-2.5-Pro Comanici et al. [2025]	71.3	75.0	72.8	75.0	77.2	80.9	89.7	100.0
Doubao-1.5-pro Team [2025]	64.0	76.0	56.0	46.0	56.9	86.0	90.0	96.0
Seed1.6 Seed [2025]	66.2	66.8	70.6	73.3	83.2	82.9	86.8	98.3
Open-Source Models (2D Perception)								
NVILA-8B Liu et al. [2025]	46.5	52.0	46.0	56.1	50.0	50.0	47.5	55.0
Paligemma Beyer et al. [2024]	43.8	45.1	44.2	45.7	42.3	43.6	42.5	46.1
Qwen2.5-VL-7B Bai et al. [2025]	53.1	55.4	55.0	57.3	55.2	56.1	56.0	56.5
InternVL-8B Chen et al. [2024b]	47.7	50.6	49.4	47.9	47.1	48.7	48.9	51.0
DeepSeek-VL2 Wu et al. [2024]	64.4	66.9	65.2	67.8	67.0	69.1	77.4	81.0
LLaVA-OneVision-7B Li et al. [2024]	54.8	55.4	53.5	49.8	52.1	50.6	52.7	50.4
Open-Source Models (3D Perception)								
3D-LLM Hong et al. [2023]	27.0	28.0	30.5	30.3	33.9	32.5	36.8	37.7
LL3DA Chen et al. [2024a]	36.0	33.9	32.9	34.5	36.9	33.3	39.6	36.0
Chat-Scene Huang et al. [2024]	30.8	32.5	40.7	49.0	41.6	45.3	43.3	51.0
3D-LLaVA Deng et al. [2025]	43.2	46.1	46.4	47.9	49.6	48.8	48.0	47.9
Video-3D LLM Zheng et al. [2025]	50.0	50.2	50.8	51.1	49.5	50.4	51.1	49.3
Our models								
STCR-CoT	68.3	70.2	72.2	73.7	74.5	75.6	76.7	77.8
STCR-Tag	82.4	84.8	86.2	87.1	87.7	88.6	88.6	88.7

run. The final model reported in the main experiments is trained on 256 H20 GPUs for approximately one month.

5.3 Main Results: From Positional Heuristics to State-Based Grounding

The empirical results summarized in Table 1 reveal a large gap between current VLMs and human-level spatiotemporal reasoning. 2D-perception models generally outperform 3D-perception models, suggesting that current 3D representations do not yet translate spatial geometry into high-level task-progress semantics. The positive correlation between WinSize and accuracy further indicates that larger temporal windows provide clearer progress cues, while smaller windows require more fine-grained visual discrimination.

The impact of our paradigm is clearest on Qwen2.5-VL-7B. Stage 1 (STCR-CoT) anchors the model in explicit spatial attributes and already makes it competitive with strong proprietary baselines. Stage 2 (STCR-Tag) then shows a positive scaling trend: once the perceive-then-judge scaffold is established, scalable weak labels can further improve reasoning instead of reinforcing shortcut learning. The zero-shot baselines in Table 1 are mainly used to diagnose order bias in existing VLMs, while the controlled supervision comparison is given by the same-backbone ablations in Table 2.

Beyond absolute accuracy, Figure 3(b) shows that proprietary models such as GPT-4o collapse under sequence reversal, exposing a strong sequence bias. Our paradigm reduces the forward-reverse gap to 6.53%, indicating that the model bases its decision on physical state rather than frame order.

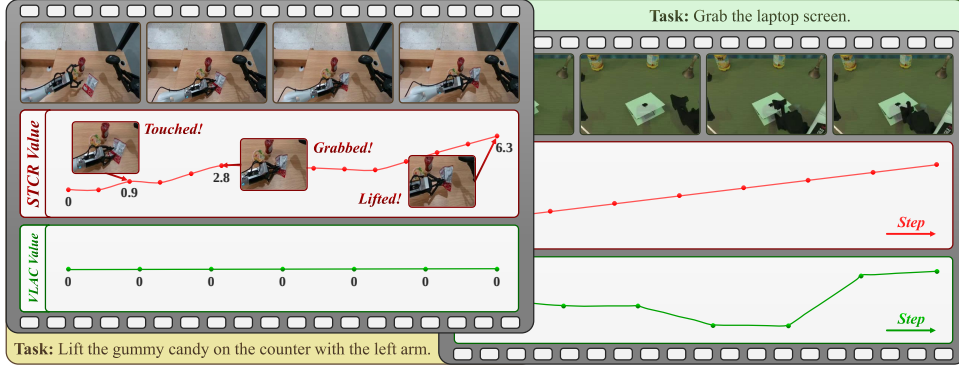


Figure 4: **Validation of Temporal Consistency in Embodied Multi-Image Reasoning.** The figure compares the reward signals from our model (top) against the baseline VLAC (bottom) Zhai et al. [2025].

Finally, the reward-model diagnostic in Figure 4 shows that our model produces a more monotonic task-progress signal than VLAC. This suggests that the learned representation can support temporal credit assignment by mapping physical progress to a continuous reward space.

5.4 Ablation Study: Dissecting the Structural Bottleneck

To diagnose the learning dynamics, we compare our paradigm with classifier-based supervision variants. Table 2 shows that the supervision signal strongly affects accuracy and robustness.

Table 2: Ablation study on supervision strategies

Strategy	Forward (%)	Reverse (%)	Average (%)
Classifier-head only	80.30	9.92	45.11
CoT + Classifier-head	84.43	16.34	50.39
CoT-only (Ours)	83.80	90.33	87.07

Failure of End-to-End Classification. The traditional “Classifier-head only” strategy suffers a catastrophic **Symmetry Collapse**, with a 70.38% forward-reverse gap. This indicates shortcut learning: with only sparse final labels, the model exploits input order instead of grounding the decision in visual evidence.

Gradient Interference in Mixed Supervision. “Mixed supervision” also fails to match CoT-only robustness. Gradients from granular CoT and sparse labels interfere, so the model treats reasoning and classification as parallel targets rather than a single perceive-then-judge hierarchy.

The Power of a Generative Bottleneck. In contrast, “CoT-only” exclusively supervises the reasoning chain, enforcing a unidirectional information flow from perception to judgment. It achieves the best average accuracy (87.07%) and reduces the temporal gap to 6.53%.

6 Conclusion

We present a CoT-based progressive training paradigm for improving spatiotemporal reasoning in VLMs. The results show that temporal hallucination is not only a data-scale problem, but also a supervision-structure problem: sparse labels let models exploit input order, while reasoning-chain

supervision encourages them to connect visual state changes to task progress. Our method reaches 87.07% accuracy and reduces the forward-reverse gap to 6.53%, and external benchmarks and cross-backbone results suggest that this structured intermediate supervision transfers beyond the main evaluation setting.

7 Limitations

Our main experiments use Qwen2.5-VL-7B; Appendix B reports less-tuned InternVL3-8B results. Appendix A shows mixed transfer on external benchmarks, and general 3D measurement remains harder. Residual noise and teacher-bias inheritance may remain in generated CoT annotations despite filtering and human cross-checking. We will release code and data artifacts after anonymity constraints are cleared.

Impact Statement. This work may improve progress estimation and robot-learning feedback, but incorrect progress judgments can still misguide downstream policies.

References

- Agibot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialu Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mingkang Shi, Modi Shi, Chonghao Sima, Jianheng Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shukai Yang, Maoqing Yao, Jia Zeng, Chi Zhang, Qinglin Zhang, Bin Zhao, Chengyue Zhao, Jiaqi Zhao, and Jianchao Zhu. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems, 2025. URL <https://arxiv.org/abs/2503.06669>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, 2025. URL <https://arxiv.org/abs/2502.13923>.
- Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, Clare Lyle, Hussain Masoom, Kay McKinney, Volodymyr Mnih, Alexander Neitz, Dmitry Nikulin, Fabio Pardo, Jack Parker-Holder, John Quan, Tim Rocktäschel, Himanshu Sahni, Tom Schaul, Yannick Schroecker, Stephen Spencer, Richie Steigerwald, Luyu Wang, and Lei Zhang. Vision-language models as a source of rewards, 2024. URL <https://arxiv.org/abs/2312.09187>.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL <https://arxiv.org/abs/2407.07726>.
- Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=bJbSbJsk0S>.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017. URL <https://arxiv.org/abs/1709.06158>.

- Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26418–26428, 2024a. doi: 10.1109/CVPR52733.2024.02496.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, June 2024b.
- 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. *arXiv preprint arXiv:2507.06261*, 2025.
- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. doi: 10.1109/CVPR.2017.261.
- Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d llms with omni superpoint transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3772–3782, 2025. doi: 10.1109/CVPR52734.2025.00357.
- Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. URL <https://arxiv.org/abs/1706.02677>.
- Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 20482–20494. Curran Associates, Inc., 2023. doi: 10.52202/075280-0900. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/413885e70482b95dcbeeddc1daf39177-Paper-Conference.pdf.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 113991–114017. Curran Associates, Inc., 2024. doi: 10.52202/079017-3620. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cebbd24f1e50bcb63d015611fe0fe767-Paper-Conference.pdf.
- Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.
- Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, volume 2026, pages 35634–35670, 2026.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

- Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL <https://arxiv.org/abs/2408.03326>.
- Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4122–4134, 2025. doi: 10.1109/CVPR52734.2025.00390.
- Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL <https://arxiv.org/abs/1608.03983>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.
- OpenAI. Openai o3 and o4-mini system card. <https://openai.com/index/o3-o4-mini-system-card/>, 2025.
- OpenAI et al. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- Viorica Patraucean et al. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761, 2023.
- Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies, 2023. URL <https://arxiv.org/abs/2304.02532>.
- Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings, 2024. URL <https://arxiv.org/abs/2305.02317>.
- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- ByteDance Seed. Technical introduction to the seed1.6 model series, 2025.
- Doubao Team. Doubao-1.5-pro: A high-performance sparse moe large language model, 2025.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL <https://arxiv.org/abs/2405.12213>.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL <https://arxiv.org/abs/2412.10302>.
- 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10632–10643. IEEE, 2025.
- 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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 157051–157088, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fe5d4bd3e2af823701b5d8f70a3c7602-Paper-Conference.pdf.
- Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A vision-language-action-critic model for robotic real-world reinforcement learning, 2025. URL <https://arxiv.org/abs/2509.15937>.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024a. doi: 10.1109/TPAMI.2024.3369699.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024b. URL <https://arxiv.org/abs/2302.00923>.
- Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8995–9006, 2025. doi: 10.1109/CVPR52734.2025.00841.

A Additional External Benchmark Results

To evaluate whether the learned reasoning ability transfers beyond our task-progress benchmark, we additionally test the trained models on several public spatial and multi-image reasoning benchmarks. These experiments are included as supplementary evidence because they are not the primary optimization target of our method.

Table 3: OmniSpatial evaluation results.

Models	Manip.	Motion	Traffic	Locate	Geo.	Geom.	Ego	Hypo.
Random Choice	24.86	26.30	25.88	23.43	27.27	24.77	22.55	25.78
GPT-3.5-turbo	38.38	29.19	38.35	28.76	36.91	24.00	42.16	35.90
GPT-4-turbo	42.97	37.40	41.18	28.95	40.00	26.32	31.37	35.42
STCR-Tag	42.86	42.86	41.67	53.33	66.67	38.46	75.00	66.67

Table 4: VSI-Bench evaluation results.

Models	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Rel. Dist.	Rel. Dir.	Appr. Order
LongVILA-8B	21.6	29.1	9.1	16.7	29.6	30.7	25.5
InternVL2-2B	26.5	25.7	24.0	20.0	32.1	44.1	6.3
VILA-1.5-8B	28.9	17.4	21.8	50.3	32.1	34.8	24.8
STCR-Tag	24.8	24.1	21.9	21.9	47.1	31.4	29.4

Table 5: CV-Bench evaluation results.

Models	ADE20K	Omni3D	Depth	CV-Bench Tong et al. [2024]
Qwen2.5-VL-7B	68.56	75.17	70.83	75.07
STCR-CoT	65.09	73.00	73.33	72.34
STCR-Tag	67.30	76.42	79.17	75.27

The results show that STCR-Tag transfers to several external spatial reasoning settings without direct optimization on those benchmarks. Gains are clearest on OmniSpatial Jia et al. [2026] and MMSI-Bench, while VSI-Bench Yang et al. [2025] remains mixed: the model improves relative-distance and appearance-order categories but does not outperform all baselines on the overall score. This suggests that our progressive training primarily strengthens relational spatial judgment and task-progress estimation, while general 3D scene measurement remains an open direction.

B Cross-Backbone Generalization

To test whether the training paradigm is tied to Qwen2.5-VL-7B, we additionally apply the full Stage 1 plus Stage 2 procedure to InternVL3-8B. Due to the limited turnaround time of the revision experiments, this reproduction is not as exhaustively tuned as the Qwen-based run. Nevertheless, it already yields consistent gains across all window sizes.

These results indicate that the progressive training recipe is not specific to Qwen2.5-VL-7B. Even with a less optimized reproduction budget, InternVL3-8B benefits from the same CoT-to-tag progression, supporting the generality of the training paradigm across VLM backbones.

C Discussion

Why did we choose to fine-tune Qwen2.5-VL?

We decided to choose Qwen2.5-VL-7B as the backbone model based on a thorough analysis of its untuned performance and key phenomena observed in subsequent fine-tuning experiments. First,

Table 6: MMSI-Bench evaluation results on compact categories.

Models	Cam.-Cam.	Obj.-Obj.	Reg.-Reg.	Cam.-Obj.	Obj.-Reg.	Cam.-Reg.
Qwen2.5-VL-7B	24.7	24.5	24.7	25.6	29.4	26.5
STCR-CoT	25.0	30.2	24.7	25.3	34.1	30.2
STCR-Tag	27.9	33.0	32.1	33.7	36.5	32.1
Models	Meas.	Appr.	Cam.	Obj.	MSR	Avg.
Qwen2.5-VL-7B	25.0	18.2	20.3	39.5	25.8	25.9
STCR-CoT	28.6	25.8	20.3	35.5	26.3	28.3
STCR-Tag	29.7	24.2	16.2	34.2	27.3	29.6

Table 7: Cross-backbone evaluation with InternVL3-8B under the progressive training pipeline.

Models	5	6	7	8	9	10	11	≥ 12
InternVL3-8B	47.7	50.6	49.4	47.9	47.1	48.7	48.9	51.0
InternVL3-8B-Full	53.9	63.5	62.5	61.1	60.4	61.8	62.9	59.9

Table 8 clearly shows Qwen2.5-VL-7B’s exceptional performance in forward tasks without any fine-tuning. For instance, it achieved an average accuracy of up to 83.73% across different window sizes, demonstrating its strong foundational capabilities. However, the table also revealed a significant performance drop in reverse tests, with an average accuracy of only 23.21%. This indicates the model’s bias but also highlights its substantial potential for improvement. Subsequently, we conducted LoRA fine-tuning experiments based on COT (Chain of Thought) on a small dataset. We observed a significant improvement in Qwen2.5-VL-7B’s robustness and generalization ability, despite its initial bias. Table 9 presents the specific manifestations and quantitative results of these performance enhancements. Collectively, these findings—Qwen2.5-VL-7B’s strong forward capabilities in its untuned state and the significant optimization space and potential revealed through fine-tuning experiments—ultimately led us to select Qwen2.5-VL-7B as the foundational model for our research.

Table 8: Model accuracy on forward and reverse passes across different window sizes

Models	WinSize									
	5	6	7	8	9	10	11	≥12	Avg.	
2D Perception										
NVILA-8B (forward)	30.3	43.4	36.3	50.5	43.4	44.4	40.4	46.4	41.89	
NVILA-8B (reverse)	62.6	60.6	55.6	61.6	56.6	55.6	54.6	63.6	58.85	
Paligemma (forward)	85.19	86.57	86	86.19	83.33	83.47	79.39	83.77	84.24	
Paligemma (reverse)	2.47	3.7	2.48	5.13	1.34	3.65	5.69	8.33	4.1	
Qwen2.5-VL-7B (forward)	83.8	78.8	82.5	82.1	83.8	87.1	86.3	85.4	83.73	
Qwen2.5-VL-7B (reverse)	23.8	22.5	24.7	22.5	23.3	21.7	27.1	20.1	23.21	
3D Perception										
3D-LLM (forward)	29.3	33.7	35.2	33.6	37.9	36.8	40.2	39.7	35.8	
3D-LLM (reverse)	24.7	22.3	25.8	27.1	29.9	28.3	33.5	35.8	28.43	
Chat-Scene (forward)	41.4	43.4	51.3	40.7	38.7	42.8	46.6	58.1	45.38	
Chat-Scene (reverse)	20.3	21.6	30.1	57.4	44.5	47.9	40.1	43.9	38.23	
3D-LLaVA (forward)	38.4	44.4	44.4	46.3	48.8	48.8	47.9	47.3	45.79	
3D-LLaVA (reverse)	48	47.7	48.3	49.5	50.3	48.8	48	48.5	48.64	

Table 9: Performance comparison of LoRA fine-tuned models across various window sizes.

Model\WinSize	5	6	7	8	9	10	11	≥ 12
Baseline								
NVILA-8B (Forward)	45.45	41.94	72.73	50.00	45.00	43.33	61.11	48.28
NVILA-8B (Backward)	63.64	41.94	36.36	35.00	65.00	53.33	44.44	48.28
NVILA-8B (Average)	54.55	41.94	54.55	42.50	55.00	48.33	52.78	48.28
Paligemma (Forward)	77.78	94.12	90.91	66.67	92.86	88.24	100	69.23
Paligemma (Backward)	76.92	85.71	69.23	66.67	100	68.75	88.89	72.22
Paligemma (Average)	77.35	89.92	80.07	66.67	96.43	78.50	94.45	70.73
InternVL3-8B (Forward)	44.44	47.06	63.64	33.33	69.23	87.50	44.44	58.33
InternVL3-8B (Backward)	38.46	64.29	36.36	27.27	57.14	21.43	66.67	47.06
InternVL3-8B (Average)	41.45	55.68	50.00	30.30	63.19	54.47	55.56	52.70
DeepSeek-VL2 (Forward)	57.14	50.00	81.81	54.54	57.14	71.42	66.67	64.71
DeepSeek-VL2 (Backward)	44.44	52.94	45.45	33.33	53.84	50.00	55.55	58.33
DeepSeek-VL2 (Average)	50.79	51.47	63.63	43.94	55.49	60.71	61.11	61.52
LLaVA-OneVision-7B (Forward)	61.54	57.14	45.45	54.54	42.86	57.14	55.56	70.59
LLaVA-OneVision-7B (Backward)	22.22	35.29	18.18	66.67	53.84	62.50	44.44	28.57
LLaVA-OneVision-7B (Average)	41.88	46.22	31.82	60.61	48.35	59.82	50.00	49.58
Qwen2.5-VL-7B (Forward)	77.78	80.65	63.64	66.67	69.23	56.25	77.78	69.23
Qwen2.5-VL-7B (Backward)	61.54	57.14	45.45	36.37	57.14	50.00	55.56	58.33
Qwen2.5-VL-7B (Average)	69.66	68.90	54.55	51.52	63.19	53.13	66.67	63.78

D Potential Dense Reward Model and Safety Boundary for Imitation

D.1 Dense reward model

In fact, by leveraging preference learning, our model can be used to supervise the real-time, online training of an embodied reward model. Unlike approaches that rely on sparse success signals or hand-crafted reward functions, our model provides dense, semantically-rich feedback online as the reinforcement learning (RL) policy interacts with the environment. Specifically, it assesses whether the policy’s action proposal at each timestep contributes positively toward the task objective. By defining a visual reward function $r_\phi(\cdot)$ over a single-frame observation o , we can directly leverage pairwise preference supervision $y \in \{0, 1\}$ from a Vision-Language Model (VLM), indicating that one image is closer to task completion than another, to train a vision-based reward model via preference learning.

We model the robot policy rollout process as a Markov Decision Process (MDP) and use a neural network r_ϕ , parameterized by ϕ , to predict the reward $r_\phi(s_t, a_t)$ for the current state-action pair $o_t = (s_t, a_t)$. Adopting a Bradley-Terry style probabilistic model, we then convert the reward function $r_\phi(\cdot)$ into a probability of pairwise comparison between two observations:

$$P_\psi[\tau^0 \succ \tau^1] = \frac{\exp(\sum_{t \in \tau^0} r_\psi(o_t))}{\exp(\sum_{t \in \tau^0} r_\psi(o_t)) + \exp(\sum_{t \in \tau^1} r_\psi(o_t))} \quad (1)$$

Periodically, every K timesteps, we feed the image pairs accumulated in the replay buffer into our Vision-Language Model (VLM) to obtain K preference labels, which are then stored in our preference buffer, D . The pairwise preference data in this buffer is subsequently used to supervise the reward model by minimizing the cross-entropy loss on the discrepancy between the preferences predicted by the reward model and those provided by the VLM.

$$\mathcal{L}_{\text{pref}}(\psi) = -\mathbb{E}_{(\tau^0, \tau^1, y)} [y \log P_\psi[\tau^0 \succ \tau^1] + (1 - y) \log P_\psi[\tau^1 \succ \tau^0]] \quad (2)$$

This enables the online update of the reward model. Once the reward model $r_\phi(\cdot)$ is updated, it is used to re-calculate the instantaneous reward for every transition in the replay buffer. The robot’s manipulation policy is then improved using these updated cumulative rewards via a standard RL loss, such as PPO Schulman et al. [2017], facilitating the post-training of an RL policy from scratch or an imitation policy, such as a VLA. Through multi-round and multi-task policy rollouts, we ultimately obtain a general-purpose reward model whose preferences are aligned with our pretrained VLM. This model, in turn, helps the robot’s manipulation policy to self-improve during its execution process.

D.2 Safety boundary for imitation policy (VLA)

A Vision-Language Model’s (VLM) preference output on consecutive frames from a robot’s manipulation process can provide a criterion for both safety protection and error recovery for imitation policies, such as the VLA. This gives open-loop imitation policies a real-time detection and recovery signal. Traditional imitation policies often struggle to recover from failure states during real-world deployment due to compounding errors, which can even lead to physical self-damage from large action errors. Intuitively, when a VLM provides a long sequence of negative preference feedback during a robot policy’s rollout, it indicates that the policy is progressively deviating from the task objective or entering an out-of-control state. By setting a tolerable threshold for the maximum length of a cumulative negative preference sequence, the agent’s erroneous execution can be terminated prematurely if this limit is exceeded. This simple criterion—triggering a policy interruption or reset—can reduce task failure from excessive cumulative error in long-horizon tasks and help detect potentially unsafe behavior during execution.

E More results

Table 10 presents a quantitative analysis of how augmenting the weakly supervised tag dataset affects model performance. A clear observation from the results is a consistent improvement in performance across both forward and reverse evaluation metrics. This trend, where model efficacy scales with the volume of tag data, provides additional evidence for the scalability of our proposed paradigm.

As shown in Table 11, the model, after being trained on Qwen2.5-VL-7B, demonstrates improved performance on the MMSI-Bench Yang et al. [2026] benchmark. Following CoT training, our model (CoT) shows an increase in average accuracy (Avg.) from 25.9% to 28.3%, which is further boosted to 29.6% after weakly supervised fine-tuning with Tag. This result gives STCR-Tag the best average score among the tested models in Table 11. Furthermore, we observe that the performance gains are concentrated in the key dimensions of spatial perception and reasoning. Specifically, appearance attribute perception (Appr.) increased from 18.2% to 24.2%, indicating an enhanced fine-grained recognition of intrinsic object features. The accuracy of reasoning about object-object relationships improved from 24.5% to 33.0%, an 7.1% increase, showing a strengthened ability to understand and judge the relative spatial relationships between multiple objects. Moreover, reasoning about object-region relationships rose from 29.4% to 36.5%, a 7.1% improvement, demonstrating a more precise understanding of object localization and relationships within spatial areas.

Table 10: Evaluation results of varying Tag-to-CoT ratios across windows (Forward).

Tag vs. CoT	5	6	7	8	9	10
14.3%	56.30%	55.96%	56.72%	56.33%	56.72%	56.90%
28.6%	28.49%	29.93%	31.34%	33.69%	34.93%	36.49%
42.9%	66.63%	70.15%	71.47%	73.48%	74.06%	74.95%
57.1%	67.17%	70.52%	71.68%	73.39%	74.08%	74.88%
71.4%	66.23%	69.70%	72.54%	74.50%	75.04%	77.34%
85.7%	68.68%	72.07%	74.13%	75.49%	77.04%	79.03%
100%	80.21%	82.81%	84.90%	85.99%	86.94%	87.44%
114.3%	78.45%	81.31%	83.28%	84.97%	85.71%	86.78%
128.6%	79.70%	81.98%	83.95%	84.48%	84.80%	86.13%
142.9%	79.06%	81.70%	83.19%	84.28%	84.80%	86.26%
157.1%	79.27%	81.82%	83.26%	83.87%	84.60%	85.35%
171.4%	77.28%	80.81%	82.64%	83.93%	84.66%	90.81%

Tag vs. CoT	11	12	13	14	15	16	Overall
14.3%	57.10%	56.67%	55.99%	55.53%	57.76%	55.14%	56.32%
28.6%	37.68%	38.38%	38.56%	38.32%	37.88%	38.49%	34.63%
42.9%	74.59%	75.20%	75.55%	75.00%	75.00%	75.59%	72.99%
57.1%	76.09%	76.81%	76.43%	76.93%	77.13%	76.83%	73.72%
71.4%	78.75%	78.55%	78.77%	77.34%	78.75%	78.55%	75.04%
85.7%	79.41%	79.54%	80.55%	79.92%	81.06%	80.83%	76.54%
100%	88.09%	88.44%	88.87%	88.20%	88.87%	88.87%	86.04%
114.3%	87.34%	87.69%	88.01%	88.13%	88.32%	88.02%	85.05%
128.6%	86.60%	86.94%	87.53%	87.07%	87.77%	87.16%	84.06%
142.9%	86.10%	86.09%	86.27%	86.43%	86.31%	86.39%	84.31%
157.1%	85.90%	86.16%	86.16%	86.15%	86.18%	85.97%	84.15%
171.4%	85.63%	86.30%	86.27%	86.27%	86.18%	85.92%	83.80%

Table 10: Evaluation results of varying Tag-to-CoT ratios across windows (Reverse).

Tag vs. CoT	5	6	7	8	9	10
14.3%	51.72%	52.07%	53.98%	54.73%	54.17%	55.47%
28.6%	92.73%	93.18%	93.65%	94.35%	93.91%	94.73%
42.9%	74.27%	75.91%	78.16%	79.07%	80.53%	80.65%
57.1%	76.82%	79.63%	81.23%	82.58%	84.28%	85.29%
71.4%	82.52%	83.77%	84.80%	85.61%	86.35%	87.35%
85.7%	82.61%	84.53%	85.87%	86.74%	87.38%	88.16%
100%	75.25%	77.17%	78.66%	79.68%	81.49%	81.86%
114.3%	78.50%	80.43%	81.84%	82.94%	83.99%	84.49%
128.6%	81.75%	83.69%	85.01%	86.19%	86.50%	87.13%
142.9%	82.87%	85.05%	85.81%	87.07%	86.97%	87.82%
157.1%	82.26%	84.43%	85.61%	86.71%	87.67%	88.41%
171.4%	87.47%	88.79%	89.85%	90.35%	85.73%	91.43%

Tag vs. CoT	11	12	13	14	15	16	Overall
14.3%	56.53%	56.92%	57.63%	57.21%	59.16%	59.37%	55.23%
28.6%	94.71%	94.52%	94.62%	95.08%	95.10%	95.05%	94.15%
42.9%	81.38%	81.54%	81.88%	81.89%	81.63%	81.68%	79.40%
57.1%	85.27%	86.27%	86.92%	86.89%	86.71%	87.01%	83.42%
71.4%	87.76%	88.01%	87.87%	87.83%	87.64%	87.73%	86.09%
85.7%	88.89%	88.51%	89.21%	89.14%	88.83%	89.20%	87.02%
100%	82.34%	82.50%	82.50%	82.75%	82.30%	82.30%	80.28%
114.3%	84.77%	84.97%	84.91%	85.20%	84.44%	84.58%	83.04%
128.6%	87.20%	87.44%	87.82%	87.66%	87.88%	87.57%	86.69%
142.9%	87.96%	88.22%	88.22%	87.97%	88.22%	88.22%	87.13%
157.1%	88.96%	89.79%	89.17%	89.67%	89.45%	88.82%	87.27%
171.4%	91.55%	91.85%	91.42%	91.23%	90.99%	90.68%	90.33%

Table 10: Evaluation results of varying Tag-to-CoT ratios across windows (Average).

Tag vs. CoT	5	6	7	8	9	10
14.3%	54.01%	54.01%	55.35%	55.53%	55.45%	56.19%
28.6%	60.61%	61.55%	62.49%	64.02%	64.42%	65.61%
42.9%	70.45%	73.03%	74.82%	76.28%	77.29%	77.80%
57.1%	71.99%	75.07%	76.46%	77.98%	79.18%	80.09%
71.4%	74.37%	76.74%	78.67%	80.05%	80.97%	82.35%
85.7%	75.65%	78.30%	80.00%	81.11%	82.21%	83.59%
100%	77.73%	79.99%	81.78%	82.84%	84.22%	85.65%
114.3%	78.48%	80.87%	82.56%	83.96%	84.85%	85.64%
128.6%	79.22%	81.75%	83.43%	85.07%	86.50%	86.63%
142.9%	80.97%	83.37%	84.50%	85.68%	85.89%	87.04%
157.1%	80.77%	83.13%	84.44%	85.29%	86.13%	86.88%
171.4%	82.38%	84.80%	86.24%	87.14%	87.73%	88.58%

Tag vs. CoT	11	12	13	14	15	16	Overall
14.3%	56.82%	56.79%	56.81%	56.37%	57.46%	57.25%	55.78%
28.6%	66.20%	66.45%	66.59%	66.70%	66.49%	66.77%	64.39%
42.9%	77.98%	78.37%	78.72%	78.44%	78.34%	78.64%	76.20%
57.1%	80.68%	81.54%	81.67%	81.91%	81.92%	81.92%	78.57%
71.4%	83.26%	83.28%	83.60%	83.71%	83.53%	83.35%	80.56%
85.7%	84.15%	84.03%	84.88%	84.53%	84.94%	85.01%	81.78%
100%	85.22%	85.47%	85.83%	85.47%	85.14%	85.59%	83.16%
114.3%	86.06%	86.33%	86.46%	86.67%	86.38%	86.30%	84.05%
128.6%	86.90%	87.19%	87.40%	87.87%	87.17%	87.01%	84.93%
142.9%	87.43%	87.16%	87.24%	87.38%	88.06%	86.99%	85.50%
157.1%	87.43%	87.98%	87.66%	87.91%	87.91%	87.39%	85.64%
171.4%	88.59%	89.08%	88.84%	88.75%	88.59%	88.30%	87.07%

Table 11: MMSI-Bench evaluation results (positional relationships).

Models	Cam.- Cam.	Obj.-Obj.	Reg.-Reg.	Cam.- Obj.	Obj.-Reg.	Cam.- Reg.
Baseline						
InternVL3-78B	34.4	23.4	32.1	12.8	37.6	26.5
InternVL2.5-78B	23.7	22.3	39.5	29.1	31.8	42.2
LLaVA-OneVision-72B	43.0	31.9	33.3	30.2	37.6	38.6
InternVL2.5-38B	18.3	22.3	35.8	22.1	38.8	34.9
Qwen2.5-VL-32B	24.7	26.6	29.6	22.1	32.9	31.3
InternVL2.5-26B	24.7	19.1	29.6	33.7	31.8	37.3
InternVL3-14B	19.4	24.5	24.7	23.3	37.6	24.1
Llama-3.2-11B-Vision	25.8	30.8	32.0	25.6	21.2	25.9
InternVL3-9B	18.3	25.5	32.1	29.1	31.8	22.9
InternVL3-8B	25.8	31.9	37.0	25.6	35.3	28.9
InternVL2.5-8B	32.3	27.7	29.6	32.6	24.7	32.5
NVILA-8B	17.2	29.8	24.7	30.2	22.4	34.9
Qwen2.5-VL-7B	24.7	24.5	24.7	25.6	29.4	26.5
LLaVA-OneVision-7B	20.4	33.0	29.6	29.1	25.9	30.1
InternVL2.5-4B	31.2	23.4	21.0	31.4	34.1	25.3
Qwen2.5-VL-3B	26.9	27.7	30.9	29.1	28.2	34.9
InternVL3-2B	26.9	25.5	29.6	31.4	28.2	27.7
InternVL2.5-2B	28.0	27.7	24.7	37.2	29.4	36.1
InternVL3-1B	24.7	35.1	22.2	30.2	29.4	30.1
InternVL2.5-1B	23.7	26.6	24.7	25.6	31.8	25.3
DeepSeek-VL2	23.7	31.9	22.2	36.0	30.6	22.9
DeepSeek-VL2-Small	24.7	28.7	18.5	33.7	38.8	27.7
DeepSeek-VL2-Tiny	29.0	27.7	21.0	23.3	17.6	31.3
Our models						
STCR-CoT	25.0	30.2	24.7	25.3	34.1	30.2
STCR-Tag	27.9	33.0	32.1	33.7	36.5	32.1

Table 11: MMSI-Bench evaluation results (attribute, motion, and summary).

Models	Meas.	Appr.	Cam.	Obj.	MSR	Avg.
Baseline						
InternVL3-78B	37.5	19.7	28.4	31.6	29.3	28.5
InternVL2.5-78B	35.9	19.7	17.6	26.3	27.3	28.5
LLaVA-OneVision-72B	28.1	19.7	13.5	32.9	15.7	28.4
InternVL2.5-38B	37.5	25.8	14.9	38.2	25.3	27.9
Qwen2.5-VL-32B	31.2	24.2	18.9	35.5	27.8	27.7
InternVL2.5-26B	35.9	30.3	10.8	31.6	26.8	28.0
InternVL3-14B	31.2	22.7	24.3	31.6	29.3	26.8
Llama-3.2-11B-Vision	20.3	19.7	25.6	28.9	19.2	25.4
InternVL3-9B	29.7	24.2	16.2	38.2	26.8	26.7
InternVL3-8B	23.4	24.2	16.2	32.9	14.6	25.7
InternVL2.5-8B	26.6	27.3	16.2	31.6	30.3	28.7
NVILA-8B	34.4	25.8	25.7	34.2	29.8	28.1
Qwen2.5-VL-7B	25.0	18.2	20.3	39.5	25.8	25.9
LLaVA-OneVision-7B	29.7	25.8	18.9	34.2	11.6	24.5
InternVL2.5-4B	23.4	24.2	13.5	31.6	26.8	26.3
Qwen2.5-VL-3B	31.2	16.7	17.6	27.6	23.2	26.5
InternVL3-2B	26.6	22.7	12.2	23.7	23.7	25.3
InternVL2.5-2B	43.8	15.2	21.6	31.6	26.8	29.0
InternVL3-1B	32.8	28.8	17.6	19.7	26.3	27.0
InternVL2.5-1B	31.2	30.3	17.6	25.0	26.3	26.1
DeepSeek-VL2	28.1	15.2	28.4	26.3	28.3	27.1
DeepSeek-VL2-Small	28.1	33.3	24.3	25.0	29.8	28.6
DeepSeek-VL2-Tiny	14.1	24.2	14.9	25.0	27.3	24.0
Our models						
STCR-CoT	28.6	25.8	20.3	35.5	26.3	28.3
STCR-Tag	29.7	24.2	16.2	34.2	27.3	29.6

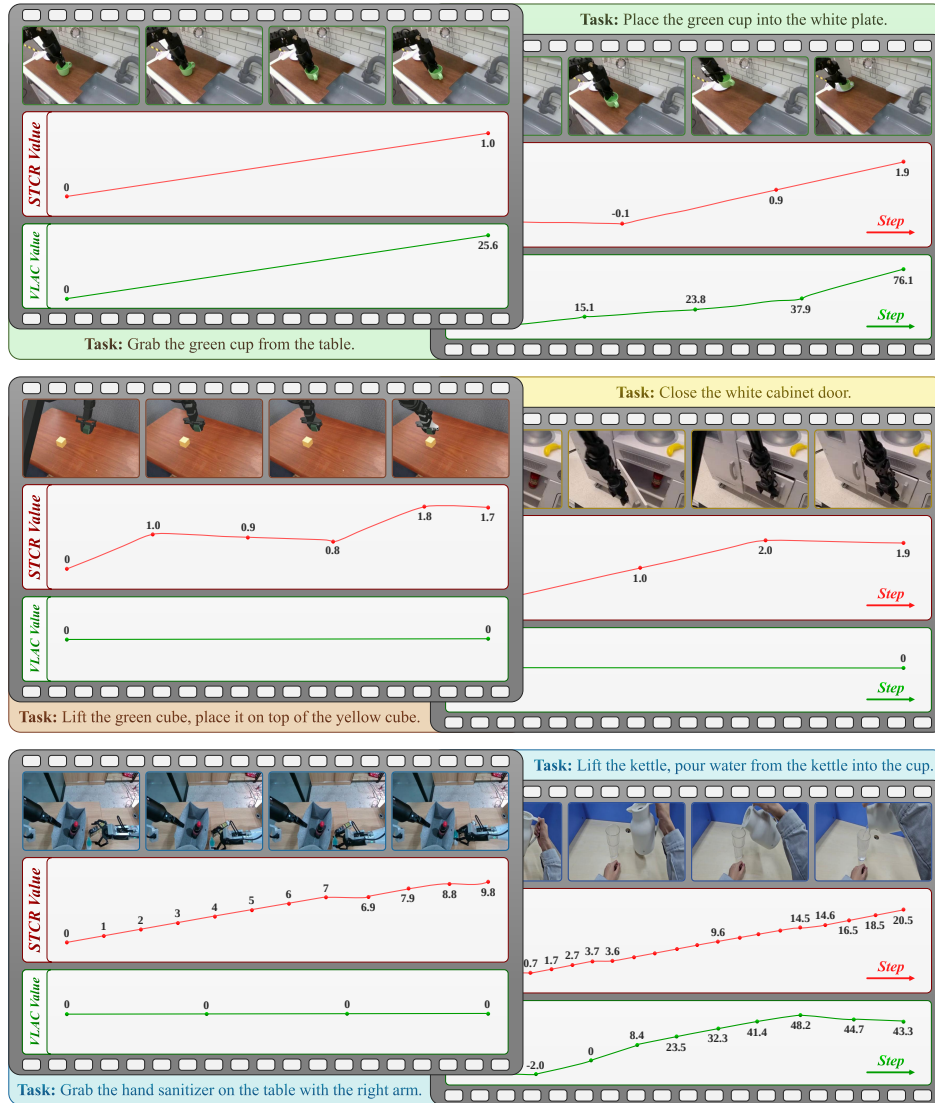


Figure 5: Comparison with VLAC.