BEYOND NEEDLE(S) IN THE EMBODIED HAYSTACK: ENVIRONMENT, ARCHITECTURE, AND TRAINING CONSIDERATIONS FOR LONG CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce ∞-THOR, a new framework for long-horizon embodied tasks that advances long-context understanding in embodied AI. ∞-THOR provides: (1) a generation framework for synthesizing scalable, reproducible, and unlimited long-horizon trajectories; (2) a novel embodied QA task, Needle(s) in the Embodied Haystack, where multiple scattered clues across extended trajectories test agents' long-context reasoning ability; and (3) a long-horizon dataset and benchmark suite featuring complex tasks that span hundreds of environment steps, each paired with ground-truth action sequences. To enable this capability, we explore architectural adaptations, including interleaved Goal-State-Action modeling, context extension techniques, and Context Parallelism, to equip LLM-based agents for extreme long-context reasoning and interaction. Experimental results and analyses highlight the challenges posed by our benchmark and provide insights into training strategies and model behaviors under long-horizon conditions. Our work provides a foundation for the next generation of embodied AI systems capable of robust, long-term reasoning and planning.

1 INTRODUCTION

Real-world embodied reasoning is a sequential decision-making problem requiring long-horizon planning, where task success depends on both memorizing and reasoning over multiple events that occur far apart in time. Using large pre-trained vision-language-action (VLA) models as policies for such tasks requires surpassing the key challenge of *long-context* reasoning. We seek to answer questions pertaining to what design choices matter in terms of environments, model architectures, and training methods when using VLA models for long-horizon embodied tasks. To this end, we develop a new framework for long-horizon tasks designed to push the boundaries of long-context understanding in embodied AI.

We introduce ∞-THOR, a new framework for generation, training, and evaluation of long-horizon embodied tasks. Our benchmark uniquely features tasks with a synthetic final goal, which involves multiple objects that appear at distant time steps, requiring multi-step reasoning across over hundreds of steps. Figure 1 illustrates an example: the agent observes the tomato at an early step (t=17) and the counter top much later (t=560). Then, the final task is given at t=670, which requires the agent to place the tomato on the counter top. This setup highlights the challenge of long-horizon dependency, where key objects and locations must be remembered and acted upon after hundreds of steps. Beyond these long-horizon dependencies, our framework also generates low-level robot-arm manipulation action sequences aligned with the trajectories, enabling agents to bridge from high-level reasoning to fine-grained physical execution.

This long-horizon setup introduces a new challenging task, Needle(s) in the Embodied Haystack (NiEH). Unlike the standard Needle in a Haystack task (Liu et al., 2024), which focuses on recalling a single clue in text, NiEH poses two main challenges: (1) multiple scattered clues (Needles) and (2) multi-modal inputs that combine visual and linguistic observations from the environment (Embodiment). This task is designed to evaluate the agent's ability to recall and reason about previously encountered environmental details, such as identifying objects and recalling performed actions.

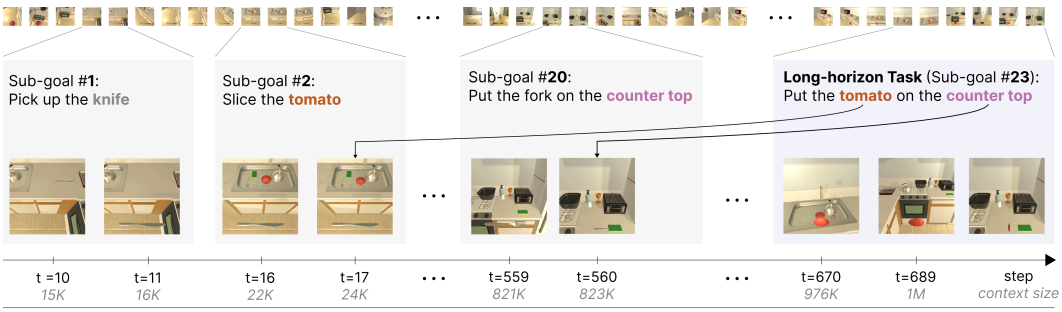


Figure 1: Example of the trajectory and a long-horizon embodied task generated from ∞-THOR. The final goal ("Put the tomato on the counter top" at t=670) requires recalling both the tomato (seen at t=17) and the counter (seen at t=560) to solved the long-horizon task. Context size refers to the input token length when converting the trajectory into the LLM input space.

Going beyond static evaluations such as NiEH, ∞-THOR also provides an interactive evaluation, allowing agents to execute policies and complete long-horizon tasks within a dynamic environment. To support this, we release a trajectory dataset for training, with episodes over 400 steps in the training set and more than 600 steps in the dev and test sets. These trajectories can be used for imitation learning, and our experiments show that access to longer context during training leads to significant performance gains, highlighting the importance of our dataset for long-context embodied reasoning.

We further investigate various architectural considerations for embodied agents to operate under extreme sequence lengths. We show that interleaved Goal-State-Action modeling—a multimodal, goal-conditioned VLA architecture that jointly models interleaved sequences of goals, states, and actions using a LLM backbone is the most practical approach for this class of problems. Moreover, since standard LLMs are constrained by fixed context windows and cannot natively handle inputs exceeding 1M tokens, we explore long-context extension techniques such as rotary embedding scaling and positional interpolation (Chen et al., 2023; Ding et al., 2024; Peng et al., 2024). Lastly, we demonstrate how to further strengthen long-context reasoning by fine-tuning the model on extended-context inputs using Context Parallelism, a parallel training strategy that allows efficient scaling to very long sequences.

We provide comprehensive experiments and analyses, demonstrating both the challenges posed by our benchmark and the behavior of baseline models under long-horizon settings. We investigate a range of training considerations, including different configurations for fine-tuning and long-context adaptation, and evaluate their impact on model performance.

Our contributions are summarized as follows:

- We introduce ∞-THOR, a new framework for generating, training, and evaluating long-horizon embodied tasks, featuring synthetic final goals that require multi-step reasoning across hundreds of steps.
- We propose a novel embodied QA task, Needle(s) in the Embodied Haystack, requiring agents to recall and reason over multiple scattered clues across extended trajectories.
- We release a large-scale trajectory dataset and an interactive evaluation environment to support both offline imitation learning and online policy execution in long-horizon settings.
- We describe architectural adaptations including interleaved Goal-State-Action modeling, long-context extension and Context Parallelism, tailored for interactive embodied reasoning.
- We present empirical results and analyses, providing insights to the current capabilities and limitations of embodied AI systems on long-horizon tasks.

2 RELATED WORK

Long-horizon Planning in Virtual Environments. AI2THOR (Kolve et al., 2017) provides interactive indoor environments widely used for embodied reasoning research, while ProcTHOR (Deitke

Table 1: Comparison of benchmarks. We use Short (< 50 steps), Medium (50-300 steps), and Long (> 300 steps) to describe task horizon, reflecting the approximate number of environment steps required to complete a task in each benchmark. (Inter. w/ env: Interaction with the environment, Mod: Modality, GT: GT actions; single/multi in the QA set column denotes single- and multi-evidence question type. * indicates the number of annotations newly collected in that work.)

Benchmark / Platform	Task Horizon	Inter. w/ env	Dataset			QA set	
			Mod	Avg steps	GT	single	multi
ProcTHOR (Deitke et al., 2022)	✗	✓	✗	✗	✗	✗	✗
MineDojo (Fan et al., 2022)	Long	✓	✗	✗	✗	✗	✗
Habitat 3.0 (Puig et al., 2023)	Long	✓	✗	✗	✗	✗	✗
VirtualHome (Puig et al., 2018)	Short	✓	multi	11.6	✓	✗	✗
ALFRED (Shridhar et al., 2020)	Medium	✓	multi	50	✓	✗	✗
ALFWorld (Shridhar et al., 2021)	Medium	✓	text	50	✓	✗	✗
BEHAVIOR-100 (Srivastava et al., 2021)	Med/Long	✓	✗	✗	✗	✗	✗
BALROG (Paglieri et al., 2024)	Long	✗	✗	✗	✗	✗	✗
EAI (Li et al., 2024b)	Med/Long	✓	✓	14.6*	✓	✗	✗
EQA (Das et al., 2018)	✗	✗	✗	✗	✗	✓	✗
MM-EGO (Ye et al., 2025)	✗	✗	✗	✗	✗	✓	✗
∞-THOR	∞	✓	multi	627	✓	✓	✓

et al., 2022) extends these capabilities by procedurally generating scalable environments, potentially facilitating longer trajectories. MineDojo (Fan et al., 2022) offers an open-ended platform within Minecraft, explicitly geared toward tasks requiring extensive long-term planning. Additionally, platforms such as VirtualHome (Puig et al., 2018) and Habitat 3.0 (Puig et al., 2023) have demonstrated suitability for tasks involving long-term interactions and complex activity sequences. However, all of these platforms only provide environments and do not include standardized datasets or benchmark suites to support training and evaluation for long-horizon embodied tasks.

Embodied QA and Multimodal Needle in the Haystack Tasks. Embodied QA tasks, such as EmbodiedQA (Das et al., 2018) and MM-EGO (Ye et al., 2025), require agents to answer questions grounded in visual observations with spatial and temporal reasoning, but without active environment interaction during evaluation. Our NiEH task is also related to multimodal Needle in a Haystack (NiH) problems. While early NiH focused on textual recall in long contexts (Liu et al., 2024), recent multimodal extensions add visual inputs (Wang et al., 2024; 2025), though they remain limited to shorter contexts (up to 72K tokens) and lack embodied reasoning or temporal dependencies.

Datasets and Benchmarks for Long-horizon Embodied Tasks. Recent work has advanced long-horizon embodied tasks, where agents complete multi-step goals with extended temporal dependencies. ALFRED (Shridhar et al., 2020) and ALFWorld (Shridhar et al., 2021) introduced instruction-following tasks with action annotations and textual grounding, but their horizons remain short (typically < 50 steps). BEHAVIOR-100 (Srivastava et al., 2021) targets household activities requiring extended engagement, though mostly single-task. EAI (Li et al., 2024b) offers a generalized interface to evaluate LLMs for embodied decision making, while our framework emphasizes online evaluation with real-time streaming inputs and focusing on long-horizon reasoning tasks.

Long-context Benchmarks. Outside embodied AI, general benchmarks have addressed challenges in long-context reasoning. Benchmarks, such as LongBench (Bai et al., 2024) and RULER (Hsieh et al., 2024), focus on retrieval or summarization tasks. GSM-∞ (Zhou et al., 2025) extends GSM-8K (Cobbe et al., 2021) to assess mathematical reasoning over extremely long textual inputs. More recently, LMAct (Ruoss et al., 2025) proposed a benchmark for evaluating frontier models' long-context multimodal decision-making on interactive game-based tasks, with up to 1M context lengths.

3 ∞-THOR: AN ENVIRONMENT FOR GENERATING, TRAINING, AND EVALUATING LONG-HORIZON EMBODIED TASKS

∞-THOR features with a generation framework for synthesizing long trajectories to train and evaluate AI agents in long-horizon embodied tasks. We build ∞-THOR upon AI2-THOR (Kolve et al., 2017) simulator, an interactive 3D environment for embodied AI research that supports diverse

scenes, objects, and agent actions. ∞-THOR enables the creation of unlimited trajectories with arbitrarily long, and provides an evaluation setup where agents can interact dynamically with the environment during both training and testing. This supports both offline learning by producing large-scale datasets, and online learning through direct agent-environment interaction.

Each trajectory generated by ∞-THOR consists of multiple task goals, such as "Put a clean sponge on a metal rack" and "Pick up the apple and place it on the microwave", requiring grounded understanding and action to achieve the goal. At the end of each trajectory, the agent is assigned a synthetic long-horizon task that requires reasoning over entities encountered at distant time steps. For the example in Figure 1, the long-horizon task (Sub-goal #23) at step t=689, "Put the tomato on the counter top", depends on observations made far earlier: the tomato at t=17 and the counter top at t=560.

Our generation framework can generate unlimited tasks, the trajectories can be exceptionally long, exceeding 1M context tokens or beyond when the trajectory is processed with LLMs. Moreover, ∞-THOR can generate low-level robot manipulation trajectories which are compatible with the ManipulaTHOR (Ehsani et al., 2021) simulator, supporting both symbolic plans and low-level controls. Successfully completing this task requires the agent to (1) memorize and integrate key environmental information over hundreds of steps, and (2) plan actions based on dependencies that are separated in time, demonstrating the need for long-context reasoning and robust spatio-temporal memory, and (3) execute low-level physical actions through extended manipulation sequences.

3.1 STATIC EVALUATION: NEEDLE(S) IN THE EMBODIED HAYSTACK

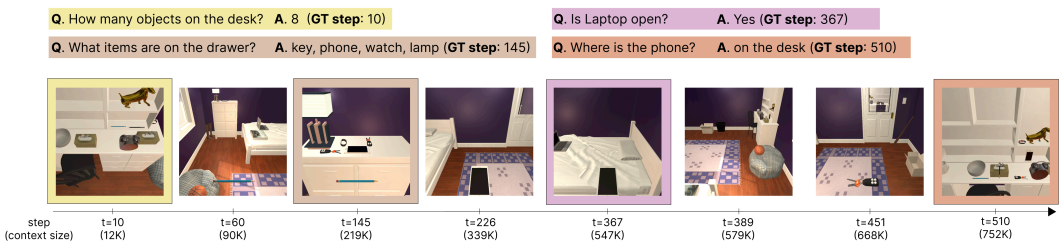
We first introduce a novel task in the form of a static evaluation: Needle(s) in the Embodied Haystack (NiEH). NiEH is designed to evaluate an agent's ability to recall and reason about environmental states encountered throughout a trajectory. Unlike traditional embodied QA tasks that focus primarily on visual understanding of a single image, NiEH emphasizes reasoning about environmental changes over time, requiring agents to interpret and integrate sequences of multimodal observations.

Figure 2 presents examples of the two NiEH task types. In the single-evidence setting, a question is answerable based on a single observation step; in the multi-evidence setting, multiple temporally distant steps must be combined to answer the question. The NiEH testset includes diverse question types, such as binary ("yes" or "no"), "what"-, "where"-, and "how many"-style questions. These questions span a broad range of difficulty, from simple memory recall (similar to the Needle in a Haystack paradigm) to complex queries that requiring multi-step reasoning across temporally and spatially distributed evidence.

Testset Construction. We first replay the generated trajectories and collect the agent's egocentric views, along with all objects that interact with the agent throughout the trajectories, such as objects that are picked up, moved, or even simply observed. Based on these interactions, we apply a set of rule-based templates to generate QA pairs, such as "Q. What object did you slice? A. {`object name`}" and "Q. Is {`obeject name`} on the desk? A. Yes/No". Then, we sample questions based on the frequency to ensure diversity across object types, and annotate the GT answer steps using the replay logs.

After generating QA pairs and annotating the GT steps, we cross-validate the answerability of each question with GT images using four different multimodal LLMs: LLaVA-OneVision 7B (Li et al., 2024a), Qwen2.5-VL 7B (Bai et al., 2025), Deepseek-VL 7B (Lu et al., 2024), and Pixtral 12B (Agrawal et al., 2024). Since these models are highly capable at standard visual QA, we filter out the questions that none of the four models successfully answer with GT images. At test time, the entire trajectory is treated as a Haystack and then cropped based on the GT image's depth. Full details on templates, generation rules, and the validation scores of the four models are included in the Appendix.

Challenges in Needle(s) in the Embodied Haystack. The NiEH task introduces two key challenges for current models. First, many questions require reasoning over multiple temporally distant events. As shown in Figure 2(b), the agent moves a `dish sponge` from the garbage bin at $t = 24$, then to the counter top, and later places it into a drawer at $t = 751$. A question such as *"Where was the dish sponge before you put it on the counter top?"* requires the model to recall and chain together multiple actions and locations across hundreds of steps. Second, some questions demand



(a) Needle in the Embodied Haystack: Single-evidence question types.

(b) Needles in the Embodied Haystack: Multi-evidence question types.

Figure 2: Example of N(s)iEH task and Ground-truth steps.

aggregating sparse and temporally scattered evidence from long trajectories. In the second example in Figure 2(b), answering *"How many times did you move the credit card?"* requires the model to track and count all relevant actions occurring from the beginning to the end of the episode. These challenges highlight the need for models that can perform robust long-horizon reasoning across both time and modalities in complex embodied environments.

3.2 Constructing Long-horizon Trajectories for Interactive Evaluations

With ∞-THOR's generation framework, we can synthesize long-horizon trajectories to construct training, validation, and test sets for offline learning and evaluation. Our approach builds upon a planner-based method (Kolve et al., 2017), in which we sequentially concatenate multiple single-task demonstrations into a extended trajectory, while maintaining consistency in object states and agent interactions throughout.

Table 2: Dataset Statistics. "low" denotes low-level robot arm manipulation actions.

NiEH testset		Single-clue	Multi-clue
# of question-answer pair		829	474
Trajectory	**Train**	**Dev**	**Test**
# trajectory	2,456	125	225
# avg/max subgoals	14/30	16/24	18/33
# avg/max steps	405/654	613/890	627/952
# avg token length	602K	880K	912K
# max token length	954K	1.2M	1.3M
# avg steps (low)	3,000+	5,000+	5,000+

To generate each trajectory, we first sample a task type from one of seven predefined templates (e.g., pick two objects and place, pick and place with movable receptacle). We then sample objects that are required to perform the task, such as items to be picked up or receptacles to be interacted with. Based on the sampled task and objects, we use a classical task planner that operates on PDDL-defined domains to generate ground-truth action sequences. Only successful rollouts (re-simulated without failure) are retained,

ensuring the reproducibility and reliability of our trajectories. We then concatenate these successful demonstrations to construct long-horizon sequences that span hundreds of steps. For the final goal, the involved objects are sampled exclusively from those seen during the early 20% and the final 20% of the trajectory. This enforces a long-term temporal dependency between two objects that must be jointly referenced to complete the final task. Through this procedure, we generate 2,456/125/225 trajectories for the training, validation, and test sets, respectively.

We use a similar approach for low-level manipulation actions. For the Pick-up action, a trajectory is generated from the robot arm's start point to the object's target point (reversed for the Put action). Each trajectory is executed through fine-grained arm movements $(\Delta x, \Delta y, \Delta z)$, and only successful rollouts are retained. Details on task types and a pseudo-algorithm for the generation process are provided in the Appendix.

4 Architectures for Long-Horizon Vision-Language-Action Models

Embodied agents must effectively interact with complex, dynamic environments, necessitating capabilities to interpret multimodal inputs (vision, language) and produce coherent sequences of actions. Developing such Vision-Language-Action (VLA) models is particularly challenging due to the need for seamless integration of perceptual understanding, linguistic reasoning, and action prediction. Existing VLA models either use separate encoders for vision, language, and action modules (Shridhar et al., 2020), or focus on short-horizon in constrained environments (Brohan et al., 2023; Kim et al., 2024), where decisions depend only on the most recent observation and a single instruction. While recent multimodal LLMs like LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), and Llama 3.2 (AI, 2024) show strong multimodal reasoning abilities, they primarily operate on static inputs (e.g., single or few images) and lack the dynamic interactivity and memory needed for long-horizon embodied tasks involving continuous vision-language-action sequences. Moreover, many state-of-the-art models are only accessible via proprietary APIs, making them impractical for real-time, controllable embodied settings and managing long-term memory states.

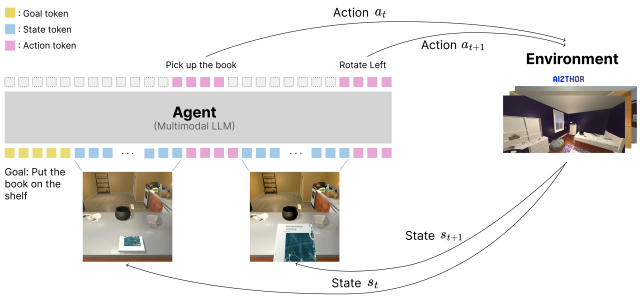


Figure 3: Agent–environment interaction through interleaved Goal-State-Action modeling.

We explore the potential of a multimodal LLM as a unified model for VLA modeling, utilizing an interleaved input structure of goal, state (visual observations), and action tokens, as illustrated in Figure 3. This interleaved multimodal input allows the model to process vision, language, and actions concurrently, thereby facilitating more coherent, real-time interaction modeling. Specifically, our goal-conditioned agent uses a multimodal LLM backbone trained to predict subsequent actions autoregressively, conditioned on sequences of goal and state tokens. At each timestep t, the environment provides a new visual observation s_t, which is encoded as state tokens and appended to the existing token stream. The model then autoregressively predicts the next action a_t conditioned on the full history of goals, states, and previously taken actions. This action is executed in the environment (e.g., "Pick up the book"), which leads to the next observation s_{t+1}, continuing the perception-action loop. This interactive sequence is repeated over hundreds of steps, allowing the model to reason over temporally distant information while maintaining grounded behavior in dynamic settings. By leveraging this interleaved Goal–State–Action modeling, our architecture supports coherent decision-making across long-horizon embodied tasks.

Context Extension. Given the limitations in context length of most LLMs, using off-the-shelf models is insufficient for processing long inputs such as those exceeding 1M tokens. We explore various long-context extension techniques that allows the model to generalize to longer input sequences without retraining from scratch. Specifically, we consider: **Linear Interpolation**(Chen

et al., 2023): Rescales input positions to fit within the pretrained RoPE range by linearly interpolating positional indices; **Dynamic Scaling**(Chen et al., 2023): Adapts RoPE frequencies at runtime based on the input sequence length, using a linear rescaling to maintain consistent positional encoding behavior across varying lengths; **YaRN**(Peng et al., 2024): Dynamically interpolates attention frequencies during inference, balancing between pretrained and extrapolated positional regimes; **LongRoPE**(Ding et al., 2024): Augments RoPE with specially designed extrapolation functions, enabling robust generalization to long sequences without degrading attention quality. We apply these techniques during fine-tuning, at inference time, or both.

Context Parallelism. To further enhance the model's ability to reason over long contexts, it is crucial to fine-tune on extended context inputs. However, the quadratic complexity of the attention mechanism makes it computationally infeasible to train directly over long sequences. To address this challenge, we employ Context Parallelism, a parallel training technique designed for efficient long-context modeling.

Context Parallelism leverages Ring Attention (Liu et al., 2023a), a novel parallel implementation of the attention layer. In Ring Attention, key-value (KV) shards are cyclically shuffled across devices, and partial attention scores are computed iteratively. This process is repeated until all KV shards have been incorporated on each device, ensuring complete attention coverage without the full memory cost of standard attention. By combining Context Parallelism with our dataset of extended long-context inputs, we are able to scale fine-tuning to substantially longer sequences, unlocking improved long-horizon reasoning capabilities.

5 Experiments

5.1 Static Evaluation: Needle(s) in the Embodied Haystack

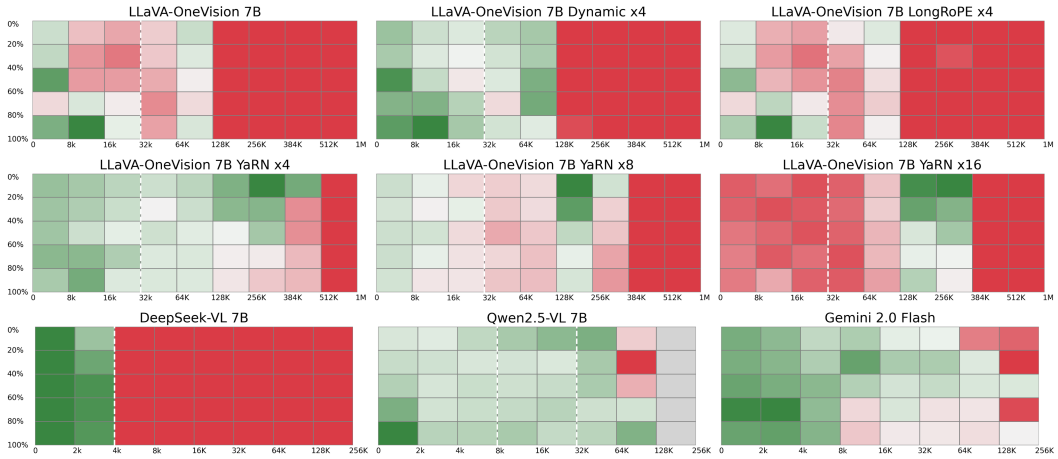
We first evaluate model performance on the Needle in the Embodied Haystack (NiEH) and Needle**s** in the Embodied Haystack (NsiEH) tasks, which test an agent's ability to retrieve and reason over sparse evidence scattered throughout long embodied trajectories.

Building a Embodied Haystack. Unlike the traditional Needle in the Haystack setup, which inserts a target sentence into a long text corpus like a book, we use the entire embodied trajectory as the input context. To simulate different reasoning depths, we crop the input sequence either from the beginning or the end based on the GT image's position. In the NsiEH task, where multiple evidences are scattered throughout the trajectory, we fill the context with intermediate steps in between the GT steps keeping their temporal order.

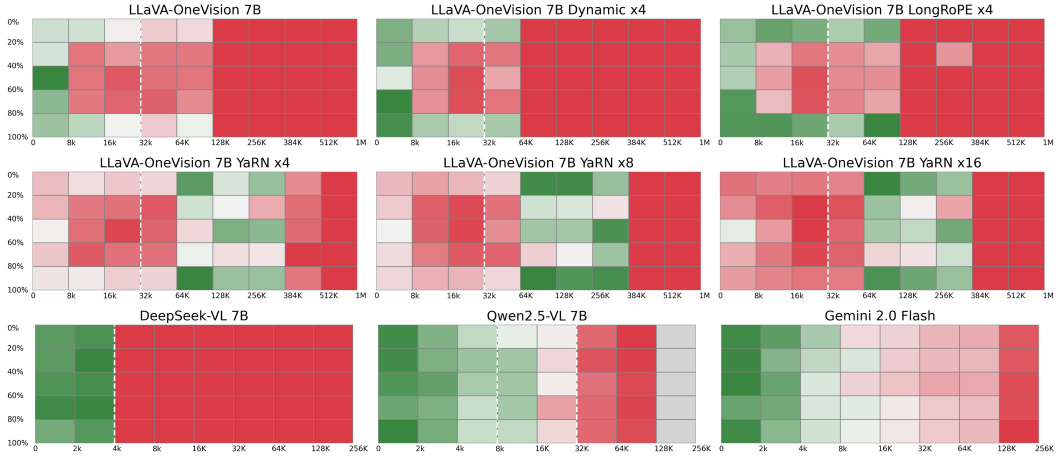
Results. Figure 4 presents the performance of LLaVA-OneVision 7B (Li et al., 2024a) model across various context extension methods. Linear Interpolation, Dynamic Scaling, and LongRoPE scaling all struggled with very long contexts beyond 128K tokens (the results of Linear Interpolation are excluded from Figure since it fails at all examples). YaRN consistently outperformed other methods, effectively handling contexts above 384K tokens, likely due to its architectural alignment with LLaVA-OneVision's Qwen2 LM backbone, which employs RoPE and YaRN scaling during pretraining. YaRN performed best at moderate scaling factors (e.g., x4), however, further scaling to x8 and x16 did not yield additional gains. In particular, x16 slightly improved performance in the 256K–384K token range but led to degradation notably at shorter context sizes (¡64K), suggesting that excessive scaling may introduce instability and negatively impact performance. Overall, all methods fail beyond 512K tokens, highlighting a need for improved long-context methods.

Figure 4 also includes additional results from DeepSeek-VL 7B (Lu et al., 2024), Qwen2.5-VL 7B (Bai et al., 2025), and Gemini 2.0 Flash[1] (Google DeepMind, 2024). Each model processes images into tokens differently—DeepSeek (576 tokens/image), Qwen2.5 (121 tokens/image), and Gemini 2.0 Flash (258 tokens/image)—which consequently impacts maximum context lengths when transforming N(s)iEH sequences into tokenized inputs (DeepSeek-VL: 512K, Qwen2.5-VL: 128K, Gemini 2.0 Flash: 256K). DeepSeek-VL struggles beyond its 4K pretrained context limit. Qwen2.5-VL maintains stronger performance up to approximately 64K tokens, benefiting from its smaller per-image tokenization, but performance notably declines on NsiEH at longer contexts. Gemini 2.0

[1] We used the `gemini-2.0-flash-001` version for all experiments.



(a) Results of Needle in the Embodied Haystack (NiEH).



(b) Results of Needles in the Embodied Haystack (NsiEH).

Figure 4: Results of Needle(s) in the Embodied Haystack. The white dashed line denotes the maximum input context length of the model. Qwen2.5-VL was pre-trained initially with an 8K token context window and incrementally scaled up to 32K tokens in subsequent stages (Bai et al., 2025). The gray area indicates contexts not applicable (N/A) due to the model's smaller image token size limiting sequences to under 128K tokens. Context Parallelism is applied to all experiments with the context size over 384K.

Flash performs robustly up to 8K tokens but degrades beyond 128K, especially on NsiEH tasks, highlighting room for improvement in complex long-range multimodal reasoning.

Single vs Multi-evidence Reasoning. Comparing NiEH to the more challenging NsiEH task, we observe a significant performance drop in the multi-evidence setting. This is especially pronounced for mid-depth questions involving sparse or distant evidence (e.g., *"Where was the Mug before you put it on the CounterTop?"*) or questions requiring the aggregation of multiple clues (e.g., *"How many times did you move the Apple?"*), as shown in Figure 2(b). These results demonstrate that our NiEH and NsiEH tasks pose a substantial challenge to current long-context models and success requires both fine-grained temporal memory and multi-evidence reasoning across extended interactions.

5.2 Interactive Evaluation in ∞-Thor

To measure agent performance on our long-horizon test set, we conduct an interactive online evaluation using the AI2THOR for high-level planning tasks and the ManipulaTHOR for low-level manipulation tasks.

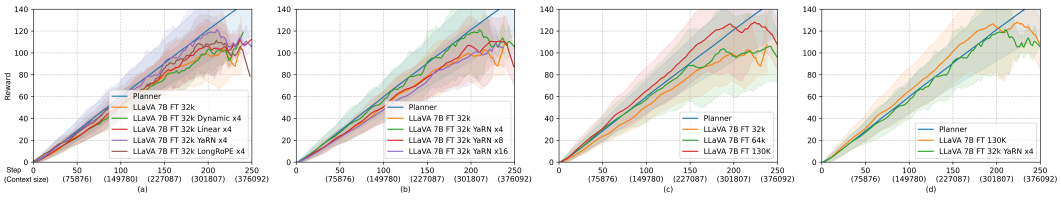


Figure 5: Agent's reward across different experimental configurations for high-level planning tasks. We compare (a) context extension methods at fixed scaling (x4), (b) varying YaRN scaling factors, and (c) fine-tuning with different context lengths using Context Parallelism. (d) summarizes the most effective strategies, highlighting that exposure to longer contexts during training significantly improves performance. Non-planner models cannot generate valid actions after around 250 steps (≈376K in context size). More configurations are in Figure 7 in Appendix.

High-level Planning Task. We fine-tune the LLaVA-OneVision 7B model on our training set while freezing the vision encoder. We used 8 H100 GPUs with both tensor parallelism (Shoeybi et al., 2020) and pipeline parallelism (Huang et al., 2019) for the 32K context size, while Context Parallelism is employed for training on larger context sizes (64K and 130K). Additional training specifics are available in the Appendix. Figure 5 presents the accumulated rewards over time across six experimental configurations. The Planner trajectory serves as the performance upper bound.

In Figure 5(a), at a fixed scaling factor of x4, YaRN consistently achieves the highest performance, closely matching the Planner. Figure 5(b) shows that scaling YaRN further to x8 or x16 yields little additional gain. In Figure 5(c), fine-tuning with Context Parallelism extends context lengths to 64K and 130K tokens, enabling the model to learn much longer sequences (86 vs. 22 steps with 32K). Additional scaling at evaluation after fine-tuning (Figures 7(d,e) in Appendix) provides no improvement and may even degrade shorter-context performance. Overall, fine-tuning on long-trajectory datasets proves most effective (Figure 5(d)), while YaRN at x4 offers strong results when large-scale training data are unavailable.

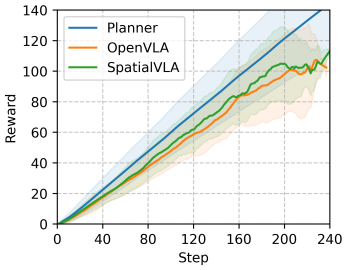


Figure 6: Agent's reward with low-level manipulation tasks. Low-level VLA models are used to control the robot arm for Pick-up and Put actions.

Low-level Manipulation Task. Figure 6 shows agent rewards on low-level manipulation tasks using OpenVLA-7B (Kim et al., 2024) and SpatialVLA-4B (Qu et al., 2025). Ego-centric views are provided as image input, and the models generate the next action $(\Delta x, \Delta y, \Delta z)$ to control the robot arm. Both models perform below the Planner due to differences in robot arm configuration and out-of-distribution inputs, though SpatialVLA achieves slightly higher and more stable rewards than OpenVLA over long sequences. We provide more results and analysis in Appendix C.

6 Conclusion

We presented ∞-THOR, a new framework for long-horizon embodied tasks designed to advance long-context understanding in embodied AI. Our framework enables scalable synthesis of long, complex trajectories paired with high- and low-level action sequences, and supports both offline training and online interaction with the environment. As part of this framework, we introduced a novel embodied QA benchmark, Needle(s) in the Embodied Haystack, that challenges agents to reason over sparse, temporally distant visual evidence embedded within extended trajectories. To equip models for this setting, we explored architectural adaptations including interleaved Goal–State–Action modeling, context extension techniques such as YaRN and LongRoPE, and efficient fine-tuning via Context Parallelism. Our experiments demonstrate that exposure to longer contexts during training significantly improves model performance, and the limitation of existing context extension techniques struggle with long-context reasoning. We hope our framework and benchmark encourage further research into models capable of robust long-horizon reasoning under realistic, interactive environments.

REPRODUCIBILITY STATEMENT

Experiments in this work were conducted using the publicly available simulators AI2-THOR and ManipulaTHOR, which allow for the deterministic replaying of all trajectories provided by ∞-THOR. We provide the complete source code, along with the generated datasets and configurations used for each experiment, and this will be made available in a public repository upon publication. Furthermore, all VLM models used in our experiments are open-source and available on Hugging Face. For the proprietary model Gemini Flash, we have declared the specific version of Gemini Flash used in the paper, and our experiments can be reproduced using the Google Cloud APIs. While the initial generation of trajectories involves some randomness in the sampling of target objects and destination positions, the resulting trajectories themselves are fully reproducible within the simulator environments. The detailed data processing steps and the experimental setup are further described in the Appendix.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024. Accessed: 2025-05-11.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023. URL https://arxiv.org/abs/2212.06817.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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, 2022. URL https://arxiv.org/abs/2206.06994.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4495–4504, 2021. URL https://api.semanticscholar.org/CorpusID:233346916.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge, 2022. URL https://arxiv.org/abs/2206.08853.

Google DeepMind. Gemini: Our largest and most capable ai models. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message, 2024. Accessed: 2025-06-20.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models, 2024. URL https://arxiv.org/abs/2308.16137.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL https://arxiv.org/abs/2404.06654.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism, 2019. URL https://arxiv.org/abs/1811.06965.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention, 2024. URL https://arxiv.org/abs/2407.02490.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

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, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS 2024*, 2024b.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL https://arxiv.org/abs/2403.05525.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. Balrog: Benchmarking agentic llm/vlm reasoning on games. In *pre-print*, 2024. URL https://example.com/BALROG.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs, 2018. URL https://arxiv.org/abs/1806.07011.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023. URL https://arxiv.org/abs/2310.13724.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Anian Ruoss, Fabio Pardo, Harris Chan, Bonnie Li, Volodymyr Mnih, and Tim Genewein. Lmact: A benchmark for in-context imitation learning with long multimodal demonstrations, 2025. URL https://arxiv.org/abs/2412.01441.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Xinyi Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.03768.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C. Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments, 2021. URL https://arxiv.org/abs/2108.03332.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.166/.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle in a multimodal haystack, 2024. URL https://arxiv.org/abs/2406.07230.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=67sSPPAZiG.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity?, 2025. URL https://arxiv.org/abs/2502.05252.

Deyao Zhu, Kan Chen He, Junnan Zhao, Wayne Wu, and Xinchao Chen. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.