Visuospatial Cognitive Assistant[*]

Qi Feng[1]
[1]Kyoto University
feng.qi.45n@st.kyoto-u.ac.jp
https://huggingface.co/nkkbr/ViCA

Abstract

Video-based spatial cognition is vital for robotics and embodied AI but challenges current Vision-Language Models (VLMs). This paper makes two key contributions. First, we introduce **ViCA**[1]**-322K**, a diverse dataset of 322,003 QA pairs from real-world indoor videos (ARKitScenes, ScanNet, ScanNet++), offering supervision for 3D metadata-grounded queries and video-based complex reasoning. Second, we develop **ViCA-7B**, fine-tuned on ViCA-322K, which achieves new state-of-the-art on all eight VSI-Bench tasks, outperforming existing models, including larger ones (e.g., +26.1 on Absolute Distance). For interpretability, we present **ViCA-Thinking-2.68K**, a dataset with explicit reasoning chains, and fine-tune **ViCA-7B** to create **ViCA-7B-Thinking**, a model that articulates its spatial reasoning. Our work highlights the importance of targeted data and suggests paths for improved temporal-spatial modeling. We release all resources to foster research in robust visuospatial intelligence.

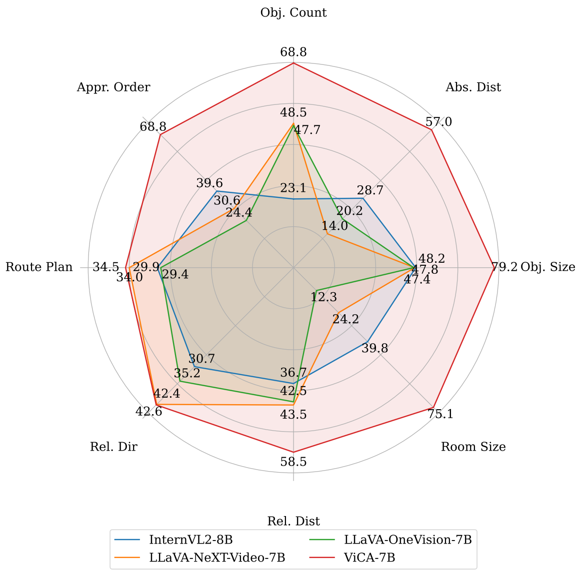


Figure 1: **Performance comparison of ViCA-7B and existing 7B/8B-scale models across eight spatial reasoning tasks in VSI-Bench.** ViCA-7B consistently outperforms all baselines, demonstrating superior visuospatial understanding capabilities.

1 Introduction

Large Vision-Language Models (VLMs) have demonstrated remarkable proficiency in multimodal understanding and generation (Bai et al., 2025; Team et al., 2024; Liu et al., 2023; Zhu et al., 2025; Wang et al., 2024), advancing tasks like visual question answering and image captioning. However, a critical frontier, particularly for robotics, augmented reality, and embodied AI, is sophisticated video-based spatial cognition—the ability to perceive, reason about, and interact with the three-dimensional (3D) structure and dynamics of environments from video.

Current VLMs often struggle with tasks demanding nuanced spatial understanding beyond simple object recognition. Challenges include comprehending fine-grained spatial relationships, tracking appearance and geometric changes, precisely estimating sizes and distances, and reasoning about navigability or functional affordances. This gap is partly due to a scarcity of large-scale, diverse datasets offering rich, targeted supervision for these intricate spatial skills. Existing datasets LLaVA-Instruct-150K(Liu et al., 2023), COCO Captions(Chen et al., 2015), MINT-1T(Awadalla et al., 2024), OmniCorpus(Li et al., 2024b) and Infinity-MM(Gu et al., 2024) may focus on static images, offer limited spatial query diversity, or lack the precise 3D ground truth crucial for robust geometric understanding. Consequently, models may fail to generalize, and their internal reasoning processes often remain opaque, hindering trust in critical applications.

[*]This is a technical report and a draft version. Work in progress.
[1]**ViCA**: **Vi**suospatial **C**ognitive **A**ssistant

To advance video-based spatial cognition, we argue for a concerted effort in two key directions: (1) developing comprehensive datasets that cover a wide spectrum of spatial reasoning tasks, grounded in rich, real-world video data; and (2) creating models that can effectively leverage such data for superior performance while offering insights into their reasoning pathways.

In this work, we address these challenges. Our contributions can be summarized as follows:

- We introduce **ViCA-322K**[2], a large-scale, diverse dataset for video-based spatial cognition, containing 322,003 high-quality question-answer pairs. Derived from three real-world indoor video datasets (ARKitScenes, ScanNet, ScanNet++), it spans direct spatial queries grounded in 3D metadata and complex reasoning tasks based on video observations.

- We develop **ViCA-7B**by fine-tuning a state-of-the-art VLM on ViCA-322K. ViCA-7B achieves new state-of-the-art performance, comprehensively and significantly outperforming current closed-source and open-source models across all eight tasks on the VSI-Bench benchmark, notably surpassing larger 72B open-source models with its 7B parameter scale.

- We construct **ViCA-Thinking-2.68K**[3], a novel dataset incorporating explicit reasoning chains for spatial cognitive tasks. Using this, we fine-tune ViCA-7B to create **ViCA-7B-Thinking**, a model that can articulate its step-by-step reasoning process, thereby enhancing model interpretability.

- We publicly release our ViCA-7B and ViCA-7B-Thinking models, evaluation code, results, fine-tuning scripts, and training logs, aiming to foster further research, ensure reproducibility, and contribute to the broader community's efforts in advancing visuospatial intelligence.

2 Related Work

Large Vision-Language Models (VLMs) The landscape of AI has been significantly reshaped by the advent of Large Language Models (LLMs) and their subsequent extension to multimodal domains, giving rise to VLMs. Models like GPT-4V (OpenAI, 2023), Gemini (Team et al., 2023), LLaVA (Liu et al., 2023), Flamingo (Alayrac et al., 2022), and BLIP (Li et al., 2022, 2023a; Xue et al., 2024) have demonstrated impressive abilities in zero-shot and few-shot learning across a variety of vision-language tasks. Many recent efforts have focused on extending these capabilities to video understanding, with models such as Video-LLaMA (Zhang et al., 2023), VideoChat (Li et al., 2023b), and LLaVA-NeXT-Video (Zhang et al., 2024a) aiming to process and reason about temporal sequences. While these models exhibit strong general visual understanding, our work focuses on enhancing their proficiency in a specific, challenging domain: nuanced video-based spatial cognition, which often requires more than recognizing objects and actions over time.

Models and Benchmarks for Spatial Cognition Reasoning about 3D space from visual input is a long-standing challenge in computer vision and robotics. While traditional approaches often relied on explicit 3D reconstruction (Kerbl et al., 2023) and geometric algorithms (Pautrat et al., 2023), recent VLM-based methods aim for end-to-end spatial understanding. The VSI-Bench (Yang et al., 2024), which we use for evaluation, represents a significant step towards standardized assessment of such capabilities in real-world video settings. Prior benchmarks Video-mme (Fu et al., 2024), Egoschema (Mangalam et al., 2023), Mvbench (Li et al., 2024a) and Video-Bench (Ning et al., 2023) might touch upon spatial aspects (e.g., object localization within an image or relative positioning), but VSI-Bench specifically targets a comprehensive suite of video-based spatial memory and reasoning tasks. Our ViCA-7B model, fine-tuned on ViCA-322K, aims to directly address the types of sophisticated spatial reasoning VSI-Bench evaluates, demonstrating how targeted data can significantly enhance VLM performance in this domain.

3 Dataset: ViCA-322K

In this section, we present ViCA-322K, a large-scale dataset specifically designed to foster research in video-based spatial cognition. We begin by detailing the foundational video sources and the overall composition of our dataset. We then describe the two main categories of question-answer pairs: Base Data, grounded in precise 3D metadata,

[2] Available at https://huggingface.co/datasets/nkkbr/ViCA-322K

[3] Available at https://huggingface.co/datasets/nkkbr/ViCA-thinking-2.68k

and Complex Spatial Reasoning data, requiring deeper video interpretation. Finally, we provide comprehensive statistics to illustrate the scale and diversity of ViCA-322K(see Figure 3).

3.1 Video Sources and Data Composition

The selection of video sources plays a crucial role in video-based spatial understanding. Following the setup of VSI-Bench, we utilize videos from ARKitScenes (Baruch et al., 2021), ScanNet (Dai et al., 2017), and ScanNet++ (Yeshwanth et al., 2023). We exclusively use the training splits of these datasets and conduct strict filtering to ensure that no videos overlap with those used in VSI-Bench. All videos depict real-world indoor environments, covering a diverse range of room types and layouts, and are accompanied by high-quality annotations.

The resulting dataset can be broadly categorized into two types: (1) data derived from metadata, such as 3D oriented bounding boxes, providing precise supervision; and (2) data requiring complex spatial reasoning based solely on video observations.

3.2 Base Data

For the Base Data, we select six spatial reasoning tasks—object count, object relative distance, object size estimation, object absolute distance, object appearance order, and room size—for which precise answers can be generated directly from 3D oriented bounding box annotations.

We deliberately omit two related tasks, object relative direction and route planning, with the objective of encouraging models to transfer learned spatial understanding to novel tasks. This design ensures that the model learns generalizable spatial reasoning abilities, rather than task-specific pattern matching.

3.3 Complex Spatial Reasoning

In contrast to Base Data, which is constructed from structured metadata, the Complex Spatial Reasoning subset is designed to probe deeper spatial understanding through natural language questions that require interpreting complex inter-object relations observed in videos.

This subset is further divided into two categories: *general* and *specific*.

The *general* category aims to develop holistic spatial awareness by prompting the model to describe and reason about the overall structure of the environment. In contrast, the *specific* category targets fine-grained spatial reasoning, requiring precise understanding of local object relationships.

By combining these two levels of spatial reasoning, we aim to comprehensively evaluate and enhance both the global and local spatial understanding capabilities of multimodal models.

3.3.1 General

The *general* category consists of two types of questions.

First, we construct medium-length multi-turn dialogues between a human and a GPT agent, focusing on discussing the spatial relationships among objects observed in the video. Multi-turn interactions are intended to refine the model's reasoning through iterative clarification and elaboration. During training, the loss is computed only on the GPT agent's responses.

Second, we ask about objects important for daily activities within the scene, prompting the model to consider whether such objects should be repositioned for better visibility or accessibility.

To generate the question-answer pairs, we employ GPT-4o-mini (OpenAI, 2024). During generation, only metadata—including object names, coordinates, sizes, and orientations extracted from 3D bounding boxes—is provided. No visual frames or images are supplied.

3.3.2 Specific

The *specific* category includes four types of tasks:

1. Given a video, determine whether it is possible to add a piece of furniture to the space. If so, specify the type, size, and anticipated improvement to the environment.

2. Evaluate the number of people the space can accommodate, propose potential usage scenarios, and describe in detail the spatial relationships between users and surrounding objects during usage.

3. Assess whether the space is accessible to a wheelchair user by choosing one of three options: (i) accessible, (ii) inaccessible, (iii) potentially accessible with minor modifications. Subsequently, provide a detailed description of the user's goals, navigation path, and interactions within the space.

4. Imagine being situated at a random location within the space and describe the surrounding

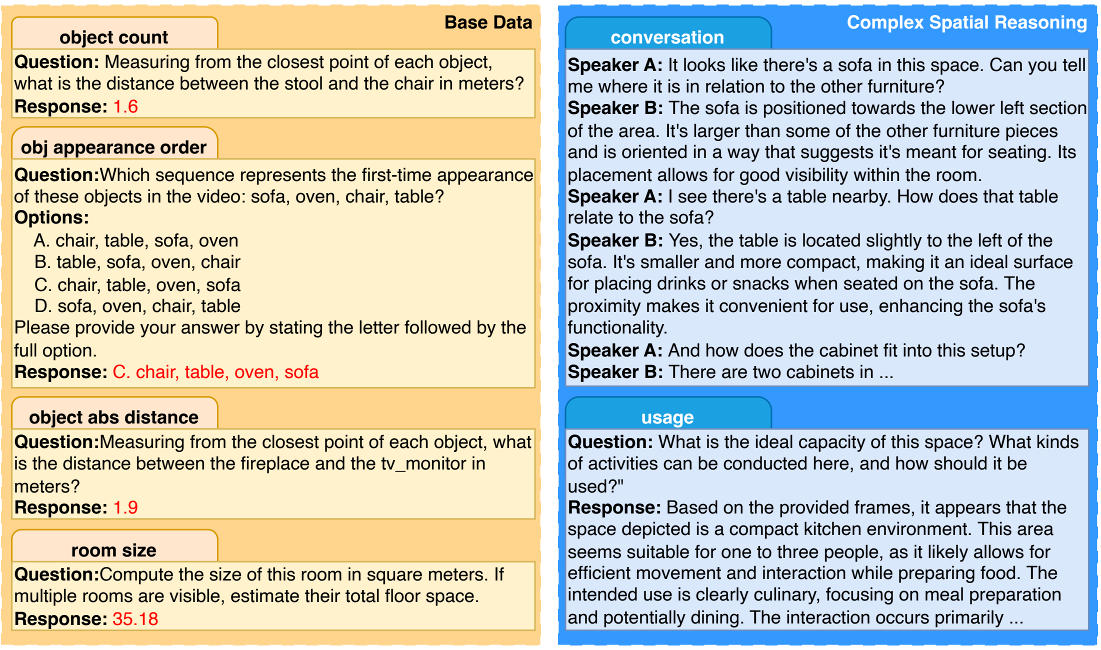


Figure 2: **An illustrative example from our ViCA-322K dataset, showcasing both Base Data and Complex Spatial Reasoning subsets.** The **Base Data** questions (left) are grounded in 3D metadata such as object counts and distances, allowing for precise supervision. The **Complex Spatial Reasoning** questions (right) require deeper interpretive understanding from the video alone, including multi-turn spatial dialogues and functional scene understanding. Together, these tasks promote comprehensive evaluation of visuospatial cognitive capabilities.

environment, focusing on the spatial relations between observed objects. To guide the model, we provide three exemplar answers prior to the task.

We generate the specific question-answer pairs using GPT-4o-mini, conditioning on 16 frames uniformly sampled from each video, along with carefully designed prompts.

To encourage linguistic diversity, we design 10 alternative phrasings for each question type and randomly sample one phrasing during the generation process for both the Base Data and Complex Spatial Reasoning subsets.

3.4 Data Statistics

The ViCA-322K dataset comprises a total of 322,003 question-answer pairs, collected from ARKitScenes, ScanNet, and ScanNet++. As shown in Figure 2, each dataset contributes both **Base Data** and **Complex Spatial Reasoning** samples.[4]

[4]In ARKitScenes, we include an additional experimental split: Triangular Positional Relationship, `triangular_positional_relationship.json` (21,707 entries). Each question asks the model to describe the side lengths and angles of a triangle formed by three specified objects.

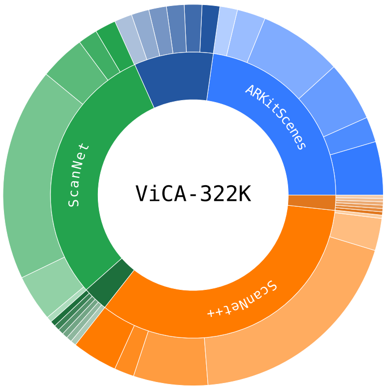
The Base Data subset consists of 281,359 questions derived from structured 3D annotations, covering six spatial reasoning tasks: object count, relative and absolute distances, object size estimation, appearance order, and room size. These questions enable precise, metadata-supervised training.

The Complex Spatial Reasoning subset comprises 40,644 questions, evenly distributed across six functionally grounded categories, including multi-turn conversations, object usage, accessibility assessment, and holistic spatial descriptions. These questions are designed to probe deeper, language-grounded visuospatial understanding based solely on video observations.

Overall, the dataset offers a balanced and diverse benchmark, combining structured spatial perception with rich, context-dependent reasoning. This composition is intended to support the development and evaluation of multimodal models with generalizable visuospatial cognitive capabilities.

4 Experiments

This section presents a comprehensive experimental evaluation of our proposed ViCA framework. We first describe the setup for our experiments, de-



ArkitScenes			
object count	13616	conversation	4495
object relative distance	6416	important daily necessities	4495
object size estimation	15217	furniture	4495
object abs distance	21302	usage	4495
obj appearance order	7230	wheelchair user	4495
room size	4494	spatial description	4495
ScanNet			
object count	5356	conversation	1425
object relative distance	4965	important daily necessities	1425
object size estimation	11658	furniture	1425
object abs distance	54240	usage	1425
obj appearance order	11860	wheelchair user	1425
room size	1425	spatial description	1425
ScanNet++			
object count	11580	conversation	854
object relative distance	5052	important daily necessities	854
object size estimation	18968	furniture	854
object abs distance	57219	usage	854
obj appearance order	8200	wheelchair user	854
room size	854	spatial description	854

Figure 3: **Hierarchical composition of the ViCA-322K dataset.** The inner ring denotes the proportional distribution across three primary sources—*ARKitScenes*, *ScanNet*, and *ScanNet++*—while the outer ring illustrates the fine-grained breakdown into diverse spatial reasoning task types. This structured diversity is designed to provide broad coverage of visuospatial semantics, offering inductive bias that facilitates robust generalization in multimodal learning.

tailing the training process for our ViCA models and the VSI-Bench benchmark used for evaluation. We then report the overall performance of ViCA-7B against leading open-source and proprietary models. Following this, we conduct ablation studies to investigate the effects of training data scale and the specific impact of our Complex Spatial Reasoning (CSR) data. Finally, we discuss additional probing experiments to gain deeper insights into model capabilities and current limitations in visuospatial reasoning.

4.1 Setup

4.1.1 Training

We fine-tune the current state-of-the-art open-source vision-language model, LLaVA-Video-7B-Qwen2 (Zhang et al., 2024b), using our ViCA-322K dataset. This model is built upon the checkpoint of `LLaVA-OneVision (SI)`, which has already demonstrated strong video understanding capabilities. To explore data scalability, we trained several variants of the ViCA model on different training splits. The most comprehensive version was fine-tuned on 8 H100 (80GB) GPUs for 55 hours using DeepSpeed ZeRO-3 Offload (Rajbhandari et al., 2020). We additionally trained ViCA-Base-7B using only the Base portion of ViCA-322K (i.e., excluding the Complex Spatial Reasoning data).

4.1.2 Evaluation

We evaluate all models on the newly released VSI-Bench, a high-quality benchmark specifically designed to assess multimodal models' visuospatial reasoning and memory capabilities using real-world indoor videos. This benchmark presents challenges that demand not only visual perception and language comprehension, but also reasoning over temporal order, spatial geometry, and egocentric-to-allocentric transformations.

During evaluation, we strictly follow the prompts provided by VSI-Bench, and do not adopt the `time_instruction` feature used in LLaVA-Video examples. Each video is uniformly sampled into 64 frames. In addition, we evaluate two extended variants: one with the inclusion of `time_instruction`, and another using a higher frame sampling rate of 128.

4.2 Overall Results

Table 1 summarizes the main experimental results. Our model, ViCA-7B, achieves the highest average performance across all eight tasks, outperforming not only open-source models—including larger 72B-scale models—but also several proprietary systems. In the Numerical Answer category, ViCA-7B surpasses all other models. Notably, on the Abs. Dist. task, it outperforms the second-best result by an impressive margin of 26.1 percentage points. This success is largely attributed to our precisely annotated data and the inclusion of Complex Spa-

Method	Average	Numerical Answer				Multiple-Choice Answer			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models (API)									
GPT-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models									
InternVL2-8B	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
VILA-1.5-8B	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LLaVA-NeXT-Video-7B	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
ViCA-Base-7B (**ours**)	55.4	65.6	51.3	74.9	67.1	52.0	32.6	28.4	**70.9(+22.3)**
ViCA-7B (**ours**)	**60.6(+15.2)**	**68.8(+12.6)**	**57.0(+26.1)**	**79.2(+15.1)**	**75.1(+20.7)**	**58.5(+7.2)**	**42.6**(+0.2)	**34.5**(+0.5)	68.8

Table 1: **Comparison of different models on VSI-Bench.** Our ViCA-7B achieves the best performance across most metrics. Gray shading indicates the best overall performance among all models, including 72B-scale and proprietary models, while **bold font** indicates the best performance among open-source models with 7B/8B scale. The numbers in parentheses (e.g., +26.1) represent the improvement margins over the best-performing model excluding ViCA-7B and ViCA-Base-7B. For **Rel. Dir.** and **Route Plan**, the improvements are computed relative to the best scores among other open-source models with 7B/8B scale.

tial Reasoning (CSR) tasks.

In the Multiple-Choice Answer category, while Gemini-1.5 Pro (Gemini Team, 2024) performs slightly better on Rel. Dir. and Route Plan, our models (ViCA-7B and ViCA-Base-7B) still deliver the best results among all 7B/8B-scale open-source models. Importantly, we did not provide fine-tuning data specific to these two tasks. The observed improvements suggest that the model successfully transfers its learned spatial understanding from other tasks. On the Appr. Order task, ViCA-7B outperforms all other models by over 20 percentage points, which we attribute to our use of YOLO(Redmon et al., 2016)-based precise temporal detection for object first appearances during data preparation.

4.3 Data Scale vs. Performance

A critical question is: how much data is sufficient for a model to effectively learn these spatial reasoning tasks? Could our data volume be excessive, causing overfitting, or is there still headroom for improvement? To investigate this, we randomly shuffle the training data and save model checkpoints at every 5% increment of the dataset. We then measure model performance at each checkpoint (see Figure 4).

The results show that model performance increases significantly from 5% to 60% of the training data, indicating that early-stage data plays a pivotal role in building core spatial understanding. However, after surpassing 80%, performance gains plateau and slightly decrease between 95% and 100%. This suggests that our dataset size is well-matched to the model's capacity (7B parameters), and observing performance saturation is crucial to verifying the sufficiency of our data scale relative to model architecture.

4.4 Effect of Complex Spatial Reasoning Data on Scaling Behavior

To investigate the impact of our Complex Spatial Reasoning (CSR) data, we conducted a comparative scaling analysis. We trained two main configurations: one using the full ViCA-322K dataset (322K samples, which includes CSR data) and another using only its Base Data subset (281K samples, excluding CSR data). Both models were incrementally fine-tuned, with performance evaluated on VSI-Bench at intervals corresponding to 5% to 100% of their respective training datasets.

The results, illustrated in Figure 4, demonstrate that while both models benefit from increasing amounts of training data, the inclusion of CSR data leads to markedly superior performance. The model trained on the full dataset consistently achieves higher average scores across all data scales, particularly as more training data is utilized (i.e., in higher-resource regimes). Culminating at 100% data usage, the full model achieves an average score of 60.56 on VSI-Bench, significantly outperforming the base-only model's score of 55.39 by +5.17 points.

These findings strongly suggest that the CSR data component provides crucial inductive biases and richer structural supervision. This, in turn, enhances the model's ability to generalize and tackle complex visuospatial tasks. The substantial performance gap underscores the critical role of diverse and high-quality supervision, such as that provided by our CSR tasks, in advancing spatial cognition capabilities for instruction-tuned models. We therefore advocate for the broader incorporation of such complex reasoning supervision in the development of multimodal datasets to unlock the full potential of large-scale vision-language models.

4.5 Additional Probing Experiments

We further explore whether model performance can be enhanced through means other than large-scale data curation and high-resource fine-tuning.

4.5.1 Adding Time Instructions to Prompts

We introduced explicit timing information to the prompt as follows:

> *The video lasts for 148.80 seconds, and 64 frames are uniformly sampled from it. These frames are located at 0.00s, 2.33s, 4.70s, ..., 146.40s, 148.77s. Please answer the following questions related to this video.*

This exposes the model not only to the 64 frames but also to their precise sampling timestamps. Evaluation results (see Appendix Table 5) reveal that including time instructions does not improve performance. In fact, the lengthy `time_instruction` text (listing 64 timestamps) slightly degrades model performance. This suggests that the model fails to leverage temporal alignment cues effectively, and instead, the verbosity may interfere with its original reasoning flow.

It is important to note that both LLaVA-Video-7B-Qwen2 and ViCA-7B were trained with such time instructions. Hence, the inability to utilize temporal information reflects a limitation in how well current architectures can extract structured time-based spatial understanding.

4.5.2 Increasing Input Frames

We also tested whether providing more visual content helps. While training and inference for LLaVA-Video-7B-Qwen2 and ViCA-7B were consistently conducted using 64 frames, we experimented with increasing the input to 128 frames (see Appendix Table 6).

Surprisingly, model performance did not improve. The model fails to benefit from the additional visual inputs, likely because it has been conditioned to expect a fixed number of visual tokens. Notably, ViCA-7B was fine-tuned from LLaVA-Video-7B-Qwen2, which itself inherits training from LLaVA-OneVision, exposed to both single/multi-image and video inputs. However, during our training, the number of visual tokens per example was fixed (13440). Increasing to 128 frames doubles the visual token count to 26880, but the model's context window (32768 tokens) still accommodates both input and output sequences. Despite no overflow, performance dropped.

This indicates that simply feeding the model more frames or more detailed temporal metadata does not lead to better spatial understanding. More sophisticated strategies may be required to align temporal information with spatial reasoning effectively.

5 Probing Model Reasoning via Structured Thought-Response Format

Although instruction-tuned models have demonstrated strong performance on video-based spatial reasoning tasks, their outputs are typically concise and discrete—either numerical values or multiple-choice letters. While effective for evaluation, such brevity often obscures the model's actual reasoning trajectory and hinders interpretability. To better understand and potentially enhance the model's internal reasoning, we introduce a structured fine-tuning format in which the model must explicitly articulate its reasoning process before providing a final answer.

5.1 Dataset Construction: Vica-Thinking-2.68K

We utilize the current state-of-the-art vision-language model, *Gemini 2.5 Pro Preview 03-25*, to construct our **Vica-Thinking-2.68K** dataset. A distinctive feature of *Gemini 2.5 Pro Preview 03-25* is that, prior to generating a final answer, it internally produces a *Thoughts* section, which contains rich and valuable step-by-step reasoning. However, the Gemini API does not expose these Thoughts to users, limiting direct access to this critical intermediate output.

To address this limitation, we manually collected five high-quality Thoughts samples through the web-based Google AI Studio interface. Specifically, for each video-question pair, we prompted *Gemini 2.5 Pro Preview 03-25* to generate ten candidate answers. From these, we selected the correct

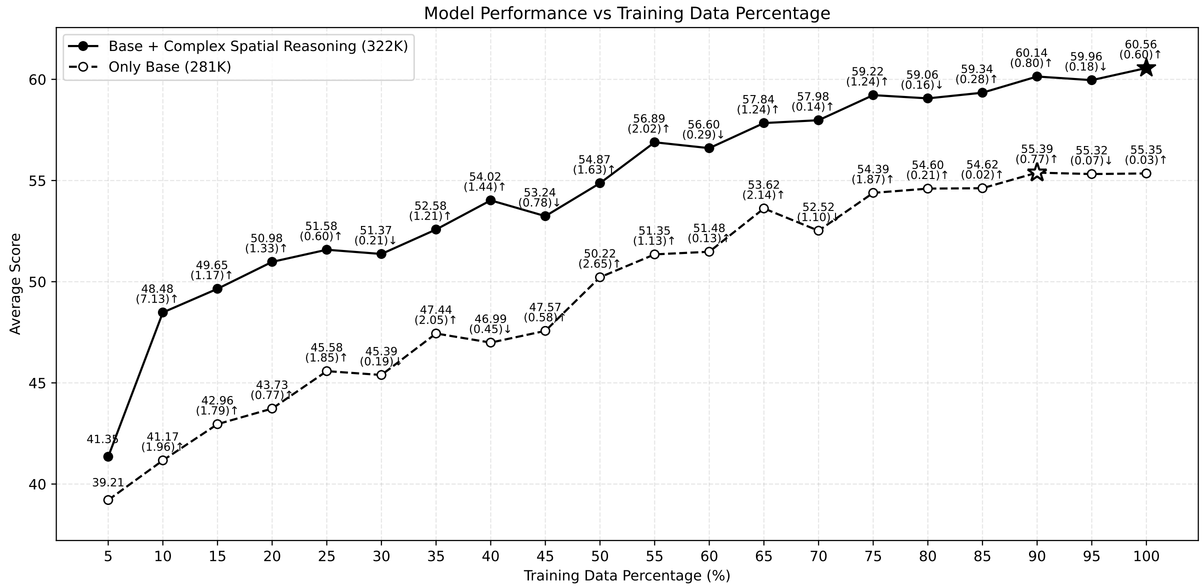


Figure 4: **Impact of Complex Spatial Reasoning on model performance scaling.** Average benchmark scores with respect to training data percentage. The solid line represents the model trained on the full dataset (322K), including Complex Spatial Reasoning data, while the dashed line shows the performance of the model trained only on the base subset (281K). The full model consistently outperforms the base-only variant across all scales. Score changes between successive checkpoints are annotated, and stars indicate the best performance in each setting.

answers—either the correct option for multiple-choice questions or answers with a matching rate (MRA) greater than 0.7 for numeric questions. We then extracted the corresponding Thoughts generated during these successful attempts. To ensure quality, we asked ChatGPT-4o to score these Thoughts, and selected the ones with the highest scores for use in our dataset.

In designing the prompting strategy (Appendix D), we first instructed the model that it must answer the question based on the given video. We explicitly required that a clear and detailed reasoning process be provided before presenting the final answer. Following this instruction, we included the five high-quality Thoughts examples collected as described above. Additionally, we provided a section titled "Guide for Writing 'Thoughts' (Step-by-Step Reasoning for Video QA)", which offered detailed guidance on structuring and articulating the reasoning process. Finally, we presented the specific question along with a strictly defined output format.

5.2 Qualitative Analysis of ViCA-7B-Thinking

Although fine-tuning for explicit reasoning in **ViCA-7B-Thinking** resulted in decrease in quantitative task performance compared to ViCA-7B (detailed in Appendix B), qualitative analysis underscores a significant gain in interpretability. By generating explicit *Thought-Response-Final Answer* outputs, ViCA-7B-Thinking articulates its step-by-step spatial reasoning from video inputs, offering a transparent view into its decision-making process. Representative examples in Appendix F illustrate this capability, showcasing structured and diverse reasoning that adheres to the prescribed format. This enhanced interpretability, despite the accuracy trade-off, is crucial for transparent model evaluation, debugging, and lays groundwork for trustworthy AI in human-in-the-loop applications.

6 Conclusion

We introduced ViCA-322K, a large-scale dataset for diverse spatial understanding, from 3D-grounded queries to complex video reasoning. Our ViCA-7B model, fine-tuned on ViCA-322K, achieved state-of-the-art on all VSI-Bench tasks (e.g., +26.1 on Absolute Distance). We found naive input scaling—such as adding detailed time instructions to prompts or increasing input frames—to be ineffective for enhancing spatial reasoning, highlighting the need for structured temporal-spatial modeling. For interpretability, ViCA-Thinking-2.68K and ViCA-7B-Thinking enable articulated reasoning, offering improved explainability despite slightly lower scores. These contributions lay a foundation for advancing spatially-aware multimodal AI.

Limitations

Despite strong performance, our work highlights several limitations. Current models, including ViCA-7B, still struggle to effectively utilize explicit temporal information or benefit from increased frame density for spatial reasoning, indicating a need for more sophisticated temporal-spatial modeling techniques. While ViCA-7B-Thinking enhances interpretability by articulating reasoning steps, it currently shows a slight decrease in quantitative task performance compared to its direct-answering counterpart, suggesting a potential trade-off or the need for further refinement in training for explicit reasoning. Finally, the ViCA-322K dataset, while diverse, is derived from indoor environments, and the generation of its complex reasoning components relies on advanced LLMs (GPT-4o-mini, Gemini 2.5 Pro), which may introduce their own biases or limitations into the training data.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. 2024. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. 2021. ARKitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Google DeepMind Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. 2024b. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI. 2023. GPT-4V(ision) System Card. https://openai.com/index/gpt-4v-system-card/. Accessed: 2025-05-17.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-\protect\penalty\z@-intelligence/. Accessed: 2025-05-17.

Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. 2023. Gluestick: Robust image matching by sticking points and lines together. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9706–9716.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimization toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. 2023. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: A strong zero-shot video understanding model.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Training Records

We publicly release the training logs(Figure 5, 6) of both **ViCA-7B** and **ViCA-7B-Thinking** to support transparency and reproducibility. During the training of ViCA-7B, we observe a sharp decrease in training loss in the initial phase, followed by a stable and gradually declining trend. This pattern indicates that the model effectively learns from the ViCA-322K dataset and progressively refines its understanding over time.

For ViCA-7B-Thinking, the training loss remains relatively unstable even in the final stages of training. This is due to two factors: (1) the content of our **ViCA-Thinking-2.68K** dataset is significantly more complex and structurally different from that of ViCA-322K used in the previous fine-tuning stage; and (2) the small scale of the dataset introduces inherent variance, and there is still room for further loss reduction if the dataset were to be expanded. These trends reflect the unique challenges of fine-tuning for interpretable reasoning and suggest that additional data could further benefit learning in this setting.

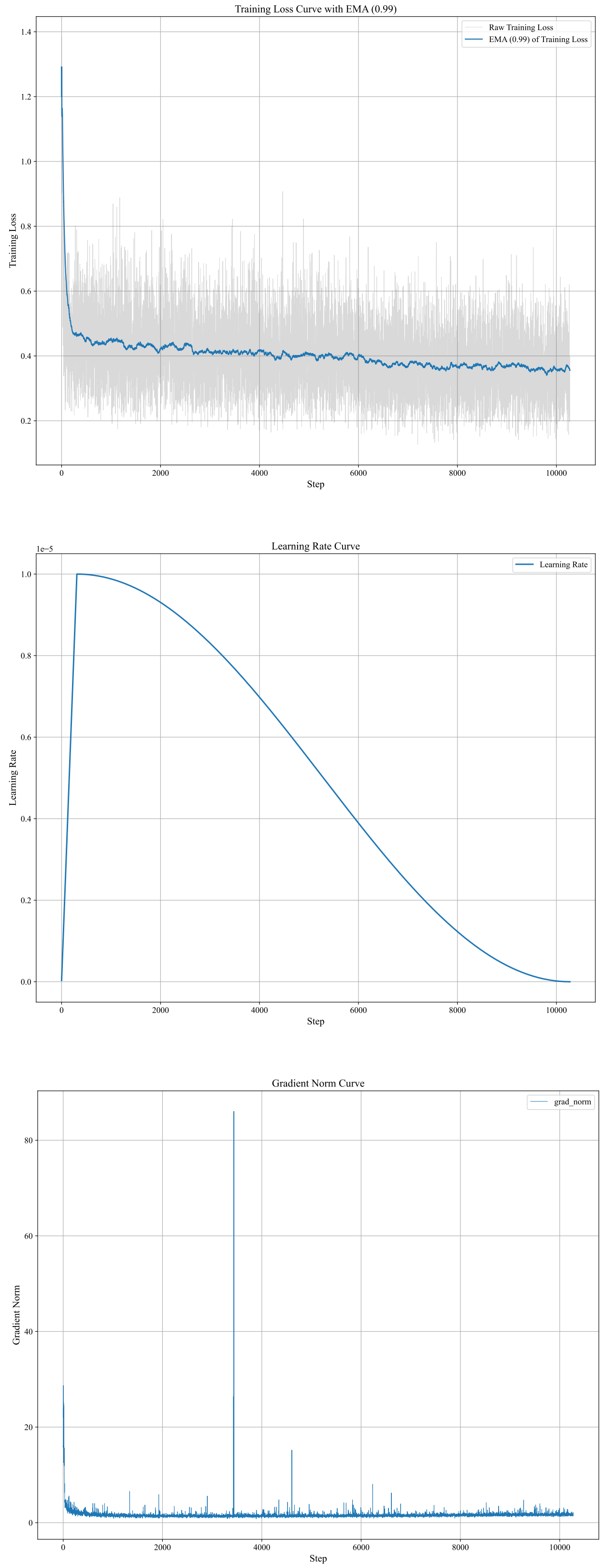


Figure 5: Training curves of ViCA-7B. From top to bottom: Training Loss with EMA, Learning Rate Schedule, and Gradient Norm.

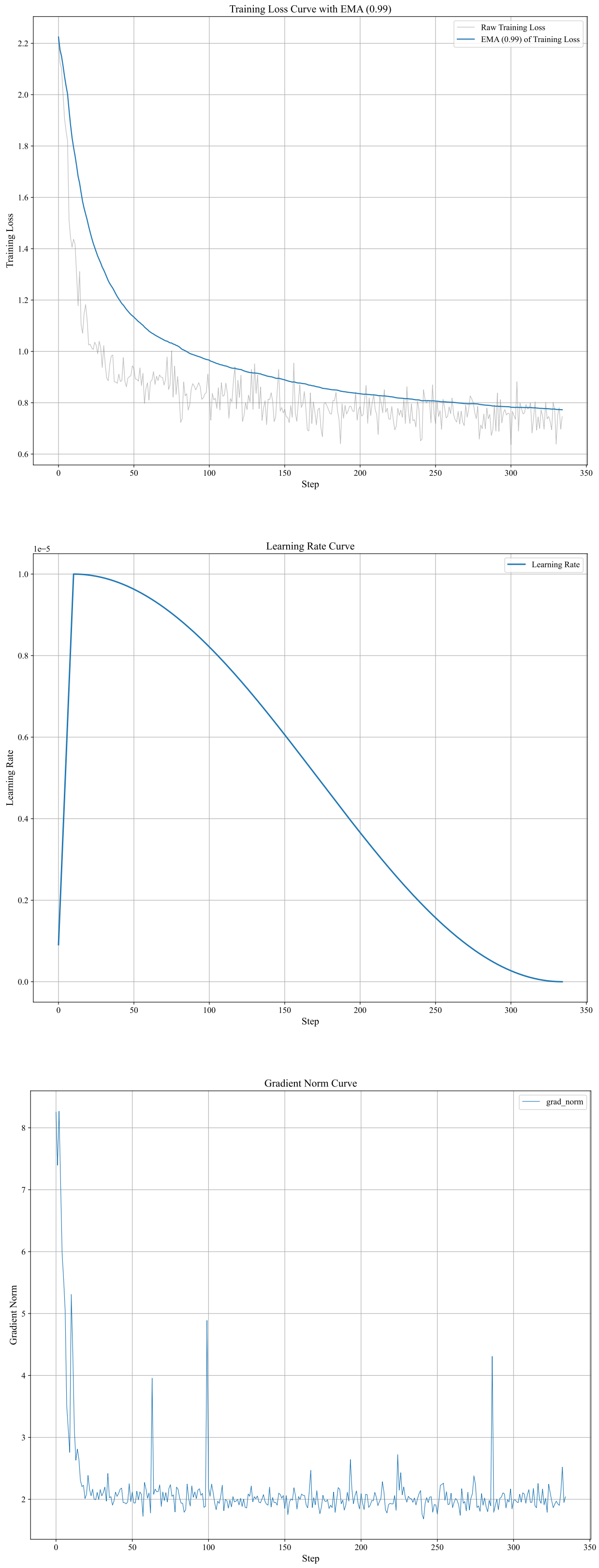


Figure 6: Training curves of ViCA-7B-Thinking. From top to bottom: Training Loss with EMA, Learning Rate Schedule, and Gradient Norm.

B Analysis of Structured Output Behavior during ViCA-Thinking Fine-tuning

To further understand the learning dynamics of structured reasoning supervision, we conduct a series of controlled experiments on the ViCA-Thinking-2.68K dataset. Our goal is to investigate the following three questions:

- **Q1:** How much fine-tuning is required for the model to adopt the desired structured output format?
- **Q2:** At different training stages, what proportion of outputs are syntactically valid and contain a usable final answer?
- **Q3:** How does this structured fine-tuning impact the model's visuospatial reasoning performance, as measured by VSI-Bench?

To answer these questions, we train **ViCA-7B-Thinking** on ViCA-Thinking-2.68K and save checkpoints every 5% of training progress. Each checkpoint is evaluated on the full VSI-Bench test set (5,130 examples).

Structured Output Format. The desired output format consists of three clearly segmented sections:

<|im_start|>think ...
<|im_start|>response ...
<|im_start|>final

We first examine whether the model produces outputs in this structured format. When using the original ViCA-7B (i.e., with 0% ViCA-Thinking data), the model correctly predicts numerical or multiple-choice answers but does not generate any intermediate reasoning or formatted segmentation. After fine-tuning on 5% of the data, we observe partial adoption of the desired structure: some outputs begin with a `<|im_start|>think` segment, while others resemble unstructured predictions.

As shown in Table 2, the adoption of the structured format stabilizes after 15% of training, with over 85% of outputs across all VSI-Bench examples conforming to our defined format. This indicates that even limited exposure to structured reasoning examples is sufficient to induce the desired output behavior.

Syntactic Validity and Final Answer Presence. Not all outputs are valid even when the overall format is correct. Some responses are incomplete (e.g., missing a final answer section), contain malformed LaTeX (e.g., incorrect `\boxed{}` syntax), or omit the final answer altogether. We define a response as *valid* if it includes a properly formatted `final` section with a clearly extractable answer in one of the expected forms.

Table 2 shows the proportion of valid responses across checkpoints. We observe steady improvements in syntactic and semantic correctness, plateauing after 15%-20% of training.

Answer Extraction and Evaluation. To evaluate model performance quantitatively, we extract the content within the following LaTeX-styled segment:

```
\[
\boxed{\text{Your final answer}}
\]
```

While most final answers are numerical or categorical, some contain vague language, ranges, or textual justification. To ensure evaluation consistency, we employ `Qwen2.5-32B-Instruct` as a post-processing model to canonicalize final answers into a numerical or multiple-choice format suitable for VSI-Bench scoring. This conversion process is fast and efficient: using a single NVIDIA H200, we process all 5,130 valid examples from a given checkpoint in under 13 minutes.

Performance Results. The results, presented in Table 3, show a sharp drop in VSI-Bench accuracy after structured fine-tuning. Beyond 10% of training on ViCA-Thinking-2.68K, the average score across all tasks falls below 40 and remains consistently low at higher percentages. These findings align with our main paper observations (see Section 5.2): although structured supervision enhances reasoning transparency, it introduces a measurable tradeoff in token-level task performance under current modeling frameworks.

Training %	Valid Outputs	Direct Answer	Via Thoughts
5	5036	4717	319
10	4434	515	3919
15	4432	0	4432
20	4241	0	4241
25	4438	0	4438
30	3668	0	3668
35	4313	0	4313
40	4424	0	4424
45	4338	0	4338
50	4555	0	4555
55	4524	0	4524
60	4518	0	4518
65	4559	0	4559
70	4357	0	4357
75	3826	0	3826
80	4281	0	4281
85	4414	0	4414
90	4493	0	4493
95	4496	0	4496
100	4491	0	4491

Table 2: Structured output adoption across training progress. Each row shows the number of VSI-Bench test examples producing a valid output—either a direct answer or a structured output with explicit `<|im_start|>think` and `<|im_start|>response` segments—at a given training percentage.

Training %	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Obj. Cnt	65.35	46.85	41.24	34.48	28.97	40.42	36.61	40.23	39.85	37.77	37.28	34.12	35.10	40.24	37.51	36.51	34.36	37.85	36.70	37.32
Abs. Dist.	52.16	24.48	26.47	26.18	26.24	27.53	28.00	27.98	28.25	25.60	26.16	24.96	27.21	28.31	28.89	29.51	28.20	29.63	28.48	29.10
Obj. Size	76.00	55.25	41.18	37.92	42.27	40.05	42.90	46.51	42.82	41.28	48.80	44.25	47.49	46.17	45.55	45.65	48.94	46.34	46.12	46.94
Room Size	63.19	21.14	25.58	30.88	22.29	26.80	36.53	32.60	31.16	29.23	37.52	29.93	35.37	32.33	35.75	35.56	31.15	32.54	33.05	31.68
Rel. Dist.	58.31	45.66	40.28	47.17	44.77	43.27	47.47	46.98	43.01	43.31	45.35	48.16	45.05	48.01	41.84	48.77	49.79	50.61	46.91	50.20
Rel. Dir.	37.95	34.73	37.13	36.41	37.91	34.63	39.06	36.74	36.72	34.89	36.43	38.61	39.18	32.68	36.55	37.83	39.84	40.21	37.86	35.95
Route Plan	30.00	32.32	25.15	30.56	29.10	28.57	26.88	29.53	25.00	23.03	28.46	25.33	24.83	21.85	23.53	26.35	25.79	28.67	30.92	21.95
Appr. Order	63.53	43.67	44.22	57.59	47.52	44.94	44.25	46.69	48.76	48.39	53.89	51.90	56.77	53.71	53.56	53.28	51.57	51.16	54.81	54.88
Average	55.81	38.01	35.16	37.65	34.88	35.78	37.71	38.41	36.95	35.44	39.23	37.16	38.87	37.91	37.90	39.18	38.70	39.63	39.36	38.50

Table 3: VSI-Bench performance of ViCA-7B-Thinking at different training stages (5% to 100% of ViCA-Thinking-2.68K). Each row reports the model's accuracy on a specific task, and the final row shows the average across all eight tasks. While structured supervision enables explainable reasoning (see Section 5.5), it introduces a notable decline in raw accuracy, particularly in early training stages.

Prompt Used for Converting Outputs from ViCA-7B-Thinking

```
You will be given a question and a response from a large language model (LLM) to that question.

Your task is to convert the LLM's response into a specific format I will describe.

There are two types of expected answers:

The correct answer should be a number, which can be either an integer or a decimal.
The correct answer should be a multiple-choice option, such as "A. xxx", "B. xxx", etc.
Here is how you should handle each case:

If the LLM's response is already a number, return it exactly as is, without any modifications.
Example: '2.12' → return '2.12'
If the response is already a multiple-choice option, return it exactly as is, without any
↪  modifications.
Example: 'B. front-right' → return 'B. front-right'
Example: 'C. ' → return 'C. '
If the response includes a number with units, strip off the units and return only the number.
Example: '17.5 square meters' → return '17.5'
If the response includes an estimated range, compute a reasonable estimate (e.g., the average)
↪  and return a single number without any unit.
Example: 'Approximately 10-11 square meters' → return '10.5'
Example: '50-60' → return '55'
If the response contains a textual explanation that indicates the answer cannot be determined
↪  or matched, return 'NULL'.
Example responses that should result in 'NULL':
'The refrigerator appears first, followed by the suitcase, and then the monitor. None of the
↪  options match this sequence.'
'The video does not show a chair, so the relative position cannot be determined.'
'The sequence of actions is Sink -> Washer -> Doorway, which does not match any of the options.'
'The question is flawed.'
In summary, your output must be one of the following three types only:

A number (e.g., '27.5')
A multiple-choice option (e.g., 'B. front-right')
'NULL'
Nothing else.

Now, here is the LLM's response for you to convert:

{repr(item['response'])}

Please output exactly one of the three allowed answer types mentioned above. Do not provide any
↪  explanation.
```

C Evaluation: Effectiveness and Limitations of Structured Reasoning

C.1 Qualitative Evaluation: Improved Interpretability and Coherence

To evaluate the effectiveness of structured reasoning supervision, we qualitatively analyze predictions made by our fine-tuned model, ViCA-7B-Thinking, on a subset of the VSI-Bench benchmark. In contrast to the original ViCA-7B—which is directly supervised to produce concise final answers—ViCA-7B-Thinking is instructed to articulate an explicit reasoning process in a dedicated Thoughts section before arriving at its final response.

Figure 7, 8, 9 and 10 present representative cases that highlight the advantages of this approach. The model begins by identifying relevant visual cues from the input frames, proceeds to infer spatial relationships based on object positioning and scene geometry, and finally synthesizes these observations into a coherent conclusion. This progression mirrors a human-like chain of reasoning, where intermediate judgments are made explicit and traceable. Such outputs provide valuable interpretability: not only is the answer presented, but the why behind the answer becomes accessible to users and researchers.

Compared to the baseline ViCA-7B, which often relies on shortcut correlations learned from instruction tuning, ViCA-7B-Thinking demonstrates a higher degree of spatial awareness and contextual grounding in its verbalized reasoning. While this does not always result in higher factual accuracy (as discussed in Section 5.2), the structured format facilitates fine-grained error analysis, reveals latent model biases, and makes the reasoning process more amenable to human feedback and iterative correction.

C.2 Quantitative Evaluation: Performance Tradeoffs and Insights

To assess the quantitative impact of structured reasoning supervision, we evaluate ViCA-7B-Thinking on VSI-Bench and compare it with the original ViCA-7B. As summarized in Table 4, ViCA-7B-Thinking obtains an average score of 38.5, which is 22.1 points lower than ViCA-7B (60.6). This decline is most pronounced in tasks requiring precise numerical predictions, such as Object Count (-31.5), Absolute Distance (-27.9), and Room Size (-43.4). Even in relatively less regression-heavy tasks such as Appearance Order and Route Planning, we observe noticeable drops (-13.9 and -12.5 respectively).

At first glance, this gap may suggest that structured reasoning negatively affects task performance. However, we argue that the observed degradation reveals a deeper structural limitation of current decoder-only architectures when adapting to multiphase output formats. Specifically, requiring the model to generate both a coherent reasoning process (Thoughts) and a correct final answer in a single sequence may dilute the attention allocation and weaken the tight alignment previously learned between video input patterns and token-level outputs.

Nevertheless, we emphasize that this is not merely a loss in performance, but a constructive signal about the model's internal behavior. The structured format encourages the model to surface its latent reasoning processes, even if doing so disrupts its prior shortcut-driven optimization. As discussed in Section 5.2, ViCA-7B-Thinking produces substantially more interpretable outputs, where the connection between observations and conclusions is made explicit. In practice, such transparency is vital for applications where trust, auditability, and debugging are essential.

Moreover, our qualitative analysis indicates that ViCA-7B-Thinking tends to generalize better on under-specified or ambiguous inputs, and its errors are often more "reasonable" in the sense that they reflect plausible but imperfect reasoning chains. In contrast, ViCA-7B occasionally outputs correct answers without any visible justification—a behavior that may be beneficial for benchmarks but undesirable for safety-critical scenarios.

In summary, while ViCA-7B-Thinking underperforms its counterpart in raw accuracy, it offers a valuable shift toward explainable spatial cognition. We view this as an important step toward the broader goal of building multimodal models that can reason in human-like, interpretable ways—paving the way for future architectures capable of reconciling explicit reasoning with precise prediction.

C.3 Summary and Takeaways

Our investigation into structured reasoning supervision via ViCA-7B-Thinking reveals a nuanced tradeoff between interpretability and task-specific accuracy. While performance on standard evaluation metrics declines significantly—particularly

	ViCA-7B	ViCA-7B-Thinking
Numerical Answer		
Object Count	**68.8**	37.3 (-31.5)
Absolute Distance	**57.0**	29.1 (-27.9)
Object Size	**79.2**	46.9 (-32.3)
Room Size	**75.1**	31.7 (-43.4)
Multiple-Choice Answer		
Relative Distance	**58.5**	50.2 (-8.3)
Relative Direction	**42.6**	36.0 (-6.6)
Route Planning	**34.5**	22.0 (-12.5)
Approach Order	**68.8**	54.9 (-13.9)
Average	**60.6**	38.5 (-22.1)

Table 4: Comparison between ViCA-7B and ViCA-7B-Thinking on spatial reasoning benchmarks. The values in parentheses denote the performance difference from ViCA-7B. Bold highlights indicate the higher score between the two models for each metric.

on numerically grounded tasks—the model gains the ability to explicitly articulate its internal reasoning trajectory in a human-readable form. This enhancement is not merely cosmetic: it surfaces previously opaque decision-making processes and facilitates qualitative analysis, error attribution, and human-in-the-loop verification.

These findings underscore a key challenge in current decoder-only multimodal architectures: the difficulty of jointly optimizing for factual precision and reasoning coherence within a single generation stream. The observed performance degradation should therefore be interpreted not as a regression in modeling capability, but as evidence of structural friction in aligning generation objectives with cognitive transparency.

We believe that structured reasoning formats, despite their current limitations, represent a critical direction for the future of multimodal AI. To this end, we advocate for further research into hybrid generation schemes, modular decoding strategies, and task-specific reasoning heads that can disentangle explanation from prediction. Ultimately, our results suggest that interpretability and performance need not be at odds—if explicitly modeled, both may be improved in tandem.

D Prompt Template for the ViCA-Thinking-2.68K Dataset

In this section, we provide the complete Prompt Template for the ViCA-Thinking-2.68K dataset.

Prompt Template for the ViCA-Thinking-2.68K Dataset

```
You are an expert AI assistant specializing in visuospatial reasoning. Please read the
↪  following video description carefully and generate 10 high-quality question-answer pairs
↪  that require fine-grained spatial understanding. Each question should be grounded in the
↪  content of the video, with emphasis on ...

You will be given a question. Please answer it based on the provided video.

You are required to first go through a clear reasoning process and explicitly write out your
↪  thought process. While thinking, you should plan your approach and go through a series of
↪  logical deductions.

Below are some examples for your reference:

---

### Example 1

"Here's my thought process for determining the number of cabinets in the room based on the
↪  video:

1.  **Understand the Goal:** The request asks for the number of cabinets *in the room*. The
↪  video shows a person entering a room and panning around.

2.  **Define ""Cabinet"":** I need a working definition of ""cabinet"" for this context.
↪  Generally, a cabinet is a piece of furniture with shelves or drawers, enclosed by doors or
↪  open. This could include kitchen cabinets, bathroom cabinets, dressers, chests of drawers,
↪  storage units, etc. I need to be consistent in applying this definition.

3.  **Analyze the Video Frame by Frame (or segment by segment):**
    *   **0:00 - 0:05:** The camera looks down at the floor (black and white tiles) and feet
    ↪  wearing blue shoe covers. There's a comic book on the floor. The view then transitions
    ↪  to looking forward into a room with wood flooring. No cabinets visible yet, just the
    ↪  doorway area.
    *   **0:05 - 0:10:** The camera pans up and into the room. I see a bed, a window, a yellow
    ↪  hanging lamp, and part of a door.
    *   **0:10 - 0:14:** The camera pans left.
        *   A white **chest of drawers** (dresser) comes into view. It has multiple drawers
        ↪  (looks like 4 or 5) with wooden knobs. This fits the definition of a
        ↪  cabinet/storage unit. **Cabinet 1**.
        *   Next to it, a white **cube storage unit** (like an IKEA Kallax) is visible. It has
        ↪  multiple square compartments, some holding items (like a red suitcase, folded
        ↪  clothes/shoes). This is also a storage unit/cabinet. **Cabinet 2**.
    *   **0:14 - 0:21:** The camera continues panning left, then up towards the ceiling,
    ↪  showing the lamp again, then moves towards the window area. I briefly see the white
    ↪  chest of drawers and cube storage again from a different angle.
    *   **0:21 - 0:32:** The camera focuses on the window area. An air purifier (white,
    ↪  rectangular object) is visible, then a baby crib/bassinet. No new cabinets here.
    *   **0:32 - 0:38:** The camera pans across the room, showing the bed, the crib again.
    *   **0:38 - 0:41:** The camera focuses on the area next to the crib/bassinet, near the
    ↪  radiator. There's a tall, narrow, light **wooden bookshelf/shelving unit**. It has
    ↪  multiple shelves holding books and other items. This counts as a cabinet/storage unit.
    ↪  **Cabinet 3**.
    *   **0:41 - 0:51:** The camera pans back across the room, showing the bed, the white chest
    ↪  of drawers (Cabinet 1), and the cube storage (Cabinet 2) again. It passes a white
    ↪  chest/box on the floor near the door - maybe a toy box? It looks like a low storage
    ↪  chest. Let's count it. **Cabinet 4**. (It's the white box with wooden knobs on top seen
    ↪  clearly around 0:50).
    *   **0:51 - 0:55:** The camera pans into what seems like a different area or the other side
    ↪  of the room/closet area. It shows the cube storage (Cabinet 2) again, and next to it, a
    ↪  taller, light **wooden chest of drawers/dresser**. This one looks different from
    ↪  Cabinet 1 (it's wood, not white). It has multiple drawers. **Cabinet 5**.
    *   **0:55 - 1:05:** The camera pans around this area, showing a fan, a folding screen,
    ↪  jeans hanging on a hook, and the wooden dresser (Cabinet 5) again.
    *   **1:05 - 1:18:** The camera looks towards a slanted ceiling with a skylight/roof window.
    ↪  It shows the top of the cube storage (Cabinet 2) and the wooden dresser (Cabinet 5)
    ↪  again. Some hangers are visible near the window.
```

```
    *   **1:18 - 1:27:** The camera pans back down, showing the white chest of drawers (Cabinet 1) and the door again. The low white chest/box (Cabinet 4) is visible again near the door.
    *   **1:27 - 1:38:** The camera moves back out of the room, showing the doorway and the patterned rug/mat outside the room, then back towards the entrance area with the black and white tiles.
    *   **1:38 - 1:41:** The camera focuses back on the starting point, looking down at the comic book and feet.

4.  **Consolidate and Count:**
    *   Cabinet 1: White chest of drawers/dresser (seen multiple times, e.g., 0:11, 0:45, 1:20).
    *   Cabinet 2: White cube storage unit (seen multiple times, e.g., 0:12, 0:46, 0:52, 1:11).
    *   Cabinet 3: Tall, narrow wooden bookshelf/shelving unit (seen near the radiator/crib, e.g., 0:39).
    *   Cabinet 4: Low white storage chest/box (seen near the door, e.g., 0:50, 1:22).
    *   Cabinet 5: Taller wooden chest of drawers/dresser (seen next to cube storage, e.g., 0:53, 1:06).

5.  **Final Check:** Review the list against the video segments. Did I miss any? Did I double-count? The identified items seem distinct and fit the definition of cabinets/storage furniture. The count is 5.

6.  **Formulate the Answer:** State the final count clearly."

---

### Example 2

1.  **Analyze the Request:** The user wants to know which object (sink, bathtub, bed, or toilet) is nearest to the ""cabinet"" shown in the video. The answer format requires the letter followed by the full option.

2.  **Identify the ""Cabinet"":** The video starts focused on a white, built-in structure that looks like a cabinet or perhaps access to storage/utilities under the eaves of the roof (common in loft conversions). It has double doors. Let's assume this is the ""cabinet"" referred to.

3.  **Track the Camera Movement:**
    *   0:00-0:04: Close-up of the cabinet door opening, looking *into* the bathroom. The camera is positioned *outside* the bathroom, near the cabinet.
    *   0:05-0:07: Camera pans left, moving *into* the bathroom. The sink comes into view on the left wall.
    *   0:07-0:09: Camera continues panning left, showing the shower area next to the sink.
    *   0:09-0:14: Camera pans further left/down, showing the toilet and the area behind it (shower controls, niche).
    *   0:15-0:18: Camera pans back towards the sink and the entrance door. The sink is clearly visible, along with a heated towel rail next to the door.
    *   0:19-0:25: Camera continues panning, focusing on the shower area again.
    *   0:26-0:39: Camera moves to the far end of the bathroom, showing the toilet and the freestanding bathtub next to it.
    *   0:40-0:44: Camera looks up towards the skylight and back towards the entrance door.
    *   0:45-0:52: Camera focuses on the bathtub again.
    *   0:53-0:57: Camera shows a small window near the bathtub.
    *   0:58-1:04: Camera moves back towards the entrance, showing the sink, towel rail, and the view into the bedroom through the open door. The cabinet is visible again near the entrance.
    *   1:05-1:18: More views of the bathtub, toilet, and panning back towards the entrance.

4.  **Assess Proximity:**
    *   **Cabinet Location:** The cabinet is located just outside the main bathroom area, near the entrance door, under a sloped ceiling.
    *   **Sink:** The sink is immediately to the left as you enter the bathroom, making it very close to the entrance and thus relatively close to the cabinet just outside the entrance.
    *   **Toilet:** The toilet is further into the bathroom, past the shower area, at the far end relative to the entrance.
    *   **Bathtub:** The bathtub is at the very far end of the bathroom, next to the toilet, furthest from the entrance and the cabinet.
    *   **Bed:** The bed is in the adjacent room (bedroom), seen through the open doorway. While the doorway is near the cabinet, the bed itself is further inside the bedroom.
```

```
5.  **Compare Distances:**
    *   The cabinet is right by the entrance.
    *   The sink is just inside the entrance.
    *   The toilet and bathtub are at the far end.
    *   The bed is in another room.

6.  **Conclusion:** The sink is the closest object *within the bathroom* to the location of the
↪   cabinet (which is just outside/at the entrance). Comparing the sink to the bed requires
↪   considering the threshold. The cabinet is right next to the door frame leading into the
↪   bedroom. The sink is just inside the bathroom door frame. Visually, the sink appears
↪   slightly closer to the cabinet's location than the bed does, especially considering the bed
↪   is set back within the bedroom. However, the cabinet *is* part of the structure that
↪   separates the bathroom entrance area from the bedroom. The sink is *inside* the bathroom
↪   proper, while the bed is *inside* the bedroom proper. The cabinet sits *between* these two
↪   areas, very near the threshold.

7.  **Re-evaluate Cabinet Definition:** Is the ""cabinet"" the structure seen initially, or
↪   could it refer to something else? No other cabinet is prominently featured. The initial
↪   structure seems the most likely candidate. It appears to be built into the wall/eaves space
↪   *between* the bedroom and the bathroom entrance area.

8.  **Refine Proximity Comparison:**
    *   The cabinet is structurally part of the wall separating the bedroom/hallway from the
    ↪   bathroom entrance.
    *   The bed is in the bedroom, adjacent to this wall/cabinet.
    *   The sink is in the bathroom, adjacent to the opposite side of the doorway near the
    ↪   cabinet.
    *   The toilet and bathtub are much further away.

    Looking closely at 0:00-0:04 and 1:02-1:04, the cabinet seems to be *in* the bedroom space,
    ↪   right next to the bathroom door. The camera starts outside the bathroom, looking
    ↪   *through* the cabinet opening into the bathroom. Then it enters the bathroom. When
    ↪   looking back out (1:02-1:04), the sink is to the left of the door, the towel rail is
    ↪   between the sink and the door, and the bed is visible through the door. The cabinet is
    ↪   part of the wall structure the door is set into. Therefore, the objects closest to the
    ↪   *physical structure* of the cabinet are the bed (on the bedroom side) and the
    ↪   sink/towel rail (on the bathroom side, near the door).

    Let's measure visually from the *center* of the cabinet doors:
    *   Distance to sink: Requires passing through the doorway and moving slightly left.
    *   Distance to bed: Seems very close, just on the other side of the wall/cabinet structure.
    *   Distance to toilet/bathtub: Clearly much further.

    The framing around the cabinet opening (0:02-0:03) suggests it might be built into the wall
    ↪   separating the bedroom and bathroom. The perspective at 1:03 shows the bathroom door
    ↪   opening *into* the bedroom. The cabinet is on the wall to the right of the bathroom
    ↪   door (when viewed from the bedroom). The sink is inside the bathroom, to the left of
    ↪   the door frame. The bed is in the bedroom, likely against the wall opposite the
    ↪   bathroom door or the wall adjacent to the cabinet.

    Consider the path: From the cabinet -> doorway -> sink. From the cabinet -> turn around ->
    ↪   bed. Given the cabinet is built into the wall *between* the rooms, both the bed and the
    ↪   sink are nearby, but likely on opposite sides of the wall/doorway.

    Let's re-examine the very beginning. The camera seems positioned in the bedroom, looking
    ↪   towards the bathroom *through* the open cabinet doors, which are set into the wall. The
    ↪   bathroom is reflected. Then the camera pulls back (0:01-0:03) revealing the cabinet is
    ↪   built into the wall of the bedroom, facing the bathroom door. The camera then enters
    ↪   the bathroom (0:04 onwards).

    Okay, new interpretation:
    *   The camera starts *in the bedroom*.
    *   The cabinet is built into the bedroom wall, adjacent to the bathroom door.
    *   The reflection shows the bathroom.
    *   The camera moves towards and enters the bathroom.
    *   The sink is immediately inside the bathroom.
    *   The bed is in the bedroom where the camera started, near the cabinet.

    Comparing distance from the physical cabinet structure:
```

```
    *   Bed: In the same room, likely very close.
    *   Sink: In the adjacent room, just through the doorway.
    *   Toilet/Bathtub: Far away in the bathroom.

    Therefore, the bed is likely the nearest object to the physical cabinet structure.

9.  **Final Check:** The video shows the cabinet is in the bedroom space, right next to the
↪   bathroom door frame (visible from 0:01-0:03 and 1:02-1:04). The bed is also in the bedroom
↪   (seen clearly through the door at 1:02-1:04). The sink is just inside the bathroom. Given
↪   the cabinet and bed are in the same room, and the cabinet is adjacent to the door, the bed
↪   is most likely the closest of the options listed.

10. **Format the Answer:** The letter is C, and the full option is ""bed""."

---

### Example 3

"The user wants to identify the first appearance time of four specific categories (refrigerator,
↪   chair, tv_monitor, oven) in the video and determine their sequence.

1.  **Refrigerator:** Appears early, visible through the doorway around 0:01-0:02, more clearly
↪   seen as the camera moves forward around 0:06-0:07. It's the large stainless steel appliance
↪   on the left around 03:08. Let's pinpoint the first *clear* view. It's partially visible
↪   from the start through the doorway, but the first clear, unobstructed view seems to be
↪   around 03:08 as the camera pans left. However, it is visible much earlier as part of the
↪   kitchen setup, seen beyond the dining table. Let's re-evaluate. Ah, wait, the *stainless
↪   steel* one is seen clearly at 03:08, but there's a *different* fridge/freezer integrated
↪   into the cabinetry earlier? No, the large stainless steel one is the main one. Looking
↪   again at the start, through the doorway, you see the kitchen area. A tall, light-coloured
↪   appliance (possibly integrated fridge/freezer) is visible against the back wall from 0:00.
↪   Let's call the first visible instance of *any* refrigerator. Yes, a tall unit, likely a
↪   refrigerator/freezer, is visible in the kitchen background from the very start (0:00).

2.  **Chair:** Chairs are visible almost immediately. The dining chairs around the table are
↪   clearly seen starting around 0:05-0:06. There's also an armchair near the window visible
↪   from 0:08-0:09. A computer chair appears later around 1:31. The dining chairs are the first
↪   type seen.

3.  **TV Monitor:** An iMac computer monitor is visible on a desk around 1:31-1:32. Is there
↪   another TV earlier? Yes, there appears to be a screen, maybe a TV or large monitor, mounted
↪   on the wall or sitting high up near the red LED strip, visible from around 0:07-0:08. Let's
↪   confirm. Yes, at 0:07, a black rectangular screen is visible above the kitchen counter area,
↪   framed by red light. This seems to be the first instance. The iMac appears much later.

4.  **Oven:** An oven (part of a range cooker, looks like an AGA or similar) is visible in the
↪   kitchen area. It's black and noticeable from around 0:07 when the kitchen area comes into
↪   view. It's more clearly seen later, around 03:03 and especially 03:40. The first appearance
↪   is around 0:07.

**Sequence Determination:**

*   **Refrigerator:** Visible in the background from 0:00.
*   **Chair:** Dining chairs clearly visible from ~0:06.
*   **TV Monitor:** Screen near red light visible from ~0:07.
*   **Oven:** Range cooker visible from ~0:07.

Okay, refining the times:
*   **Refrigerator:** 0:00 (visible through doorway in kitchen)
*   **Chair:** 0:06 (dining chairs)
*   **TV Monitor:** 0:07 (screen near red light)
*   **Oven:** 0:07 (black range cooker)

The sequence is: Refrigerator, Chair, TV Monitor/Oven (appear almost simultaneously).

Let's check the options against this sequence: Refrigerator, Chair, TV Monitor, Oven.

*   A. chair, tv_monitor, refrigerator, oven - Incorrect (Refrigerator is first)
*   B. refrigerator, chair, tv_monitor, oven - Matches the observed sequence.
*   C. tv_monitor, oven, chair, refrigerator - Incorrect (Refrigerator is first)
```

```
*   D. oven, refrigerator, tv_monitor, chair - Incorrect (Refrigerator is first)

Double-checking the very first frames for any chairs. The view starts looking through a doorway.
↪  No chairs are clearly visible *before* the dining chairs appear around 0:06. The
↪  refrigerator is visible *through* the doorway in the kitchen area right from 0:00. The TV
↪  Monitor and Oven become visible around 0:07 as the camera view encompasses more of the
↪  kitchen/dining area.

Therefore, the sequence is Refrigerator (0:00), Chair (0:06), TV Monitor (0:07), Oven (0:07).
↪  Option B aligns with this."

---

### Example 4

"Here's my thought process for estimating the sofa's length:

1.  **Understand the Goal:** The request asks for the length of the *largest side* of the sofa
↪  in centimeters. This means I need to identify the longest dimension of the sofa visible in
↪  the video and estimate its measurement.

2.  **Analyze the Video:**
    *   The video starts with a close-up of the floor, showing a movie poster printout
    ↪  (""Blackmail"") and someone's feet wearing blue shoe covers. This doesn't help with the
    ↪  sofa measurement yet.
    *   The camera pans up and moves through a doorway into a living room (0:05 - 0:07).
    *   The camera pans around the living room, showing a sofa against one wall (0:08 - 0:21).
    ↪  This is the primary view of the sofa.
    *   The camera continues panning, showing the window, another chair, the TV area, furniture,
    ↪  and eventually returns to the doorway area.

3.  **Identify the Sofa and its Largest Side:** The sofa is the long, dark grey piece of
↪  furniture pushed against the light grey wall. Its largest side is clearly its length,
↪  running parallel to the wall.

4.  **Find Reference Objects:** To estimate the sofa's length, I need objects with known or
↪  easily estimated dimensions within the same frame or perspective. Potential reference
↪  objects:
    *   **Floorboards:** The floor appears to be laminate or wood plank flooring. Standard
    ↪  plank widths are often around 12-19 cm (5-7.5 inches).
    *   **Doorway:** Standard interior doorways are often around 80 cm wide. However, the
    ↪  doorway isn't directly next to the sofa in a way that makes comparison easy.
    *   **Light Switch:** Standard light switch plates are roughly 8 cm wide and 12 cm tall.
    *   **Electrical Outlet:** Similar dimensions to the light switch plate.
    *   **Window:** Windows vary greatly.
    *   **TV:** TV screen sizes are measured diagonally, and widths vary.
    *   **Armchair:** The armchair (looks like an IKEA Poäng or similar) has a typical width of
    ↪  around 68 cm.
    *   **Coffee Table:** The white coffee table looks like a standard small IKEA Lack table,
    ↪  which is 55x55 cm.
    *   **Poster (""Fenêtre sur Cour""):** This is a standard movie poster size, maybe A1 (59.4
    ↪  x 84.1 cm) or similar, but it's hard to tell for sure, and perspective makes it
    ↪  difficult to use accurately.
    *   **Person (briefly seen in doorway):** People vary too much in height/width.
    *   **Ceiling Height:** Standard ceiling heights are often around 240-270 cm, but this is
    ↪  hard to relate directly to the sofa length.

5.  **Select the Best References and Estimate:**
    *   **Floorboards:** Let's count the floorboards along the length of the sofa (around
    ↪  0:18-0:20). It looks like the sofa spans approximately 10-12 full floorboard widths. If
    ↪  we assume an average width of 15 cm per board: 10 * 15 cm = 150 cm; 12 * 15 cm = 180 cm.
    ↪  If we assume a wider board, say 19 cm: 10 * 19 cm = 190 cm; 12 * 19 cm = 228 cm. This
    ↪  gives a range, maybe centering around 180-200 cm.
    *   **Coffee Table:** The coffee table (likely 55 cm wide) is placed on a rug. Comparing the
    ↪  sofa length to the table width is tricky due to perspective, but the sofa looks
    ↪  significantly longer, perhaps 3 to 4 times the table width. 3 * 55 cm = 165 cm; 4 * 55
    ↪  cm = 220 cm. This aligns roughly with the floorboard estimate.
    *   **Armchair:** The armchair (est. 68 cm wide) is next to the coffee table. Again,
    ↪  comparing directly is hard, but the sofa appears roughly 2.5 to 3 times the width of
    ↪  the armchair. 2.5 * 68 cm = 170 cm; 3 * 68 cm = 204 cm. Consistent range.
```

```
    *   **Sofa Type:** It looks like a standard 3-seater sofa, possibly a sofa bed. These
    ↪   typically range from 180 cm to 240 cm in length.

6.  **Refine the Estimate:** The different reference points (floorboards, coffee table,
↪   armchair, typical sofa size) seem to converge on a range roughly between 170 cm and 220 cm.
↪   A common size for this type of sofa would be around 190 cm or 200 cm. The perspective makes
↪   it look quite long. Let's re-examine the floorboards (0:19). Counting more carefully, it
↪   seems closer to 11-13 boards. Using 15cm/board: 11*15 = 165, 13*15 = 195. Using 18cm/board:
↪   11*18 = 198, 13*18 = 234. 200 cm seems like a reasonable central estimate.

7.  **Consider Potential Issues:**
    *   **Lens Distortion:** The video uses a wide-angle (fisheye) lens, which distorts
    ↪   perspective, especially near the edges. This makes visual estimation less accurate.
    ↪   Objects closer to the camera appear larger, and straight lines can appear curved.
    *   **Unknown Dimensions:** The exact dimensions of the reference objects (floorboards,
    ↪   furniture) are unknown.
    *   **Viewing Angle:** The camera moves, changing the perspective constantly.

8.  **Formulate the Answer:** Based on the visual estimation using reference objects like the
↪   floorboards and the likely size of a typical 3-seater sofa/sofa bed, the largest side
↪   (length) appears to be approximately 190-210 cm. I'll state a specific number within this
↪   range as the best estimate, acknowledging it's an approximation. 200 cm seems like a
↪   plausible and common dimension.

9.  **Final Check:** Does 200 cm (2 meters) look reasonable for the sofa shown in the context of
↪   the room? Yes, it fits the visual scale relative to the other furniture and the room size."

---

### Example 5

"Here's my thought process for determining the distance between the table and the shelf:

1.  **Understand the Goal:** The request asks for the distance in meters between the *closest
↪   points* of the dining table and the shelf (specifically the bookshelf visible further down
↪   the room).

2.  **Analyze the Video:**
    *   **Starting Point (0:00-0:04):** The video starts looking through a doorway with frosted
    ↪   glass panels. We see a long room with a dining table in the foreground and kitchen
    ↪   elements further back. The camera moves forward down a few steps into the room.
    *   **Moving into the Room (0:05-0:14):** The camera pans around the dining area. The
    ↪   dining table is wooden, rectangular, and has several chairs around it. There's a large
    ↪   window on the left wall. On the right wall, there's a fireplace/wood stove. A large
    ↪   black pendant light hangs over the table.
    *   **Focusing on the Kitchen Area (0:15-0:22):** The camera moves further down the room,
    ↪   past the dining table. We see a kitchen island/peninsula with stools. Beyond that is
    ↪   the main kitchen counter area with cabinets, a sink, and windows.
    *   **Panning Back (0:23-0:38):** The camera pans back towards the dining area and the
    ↪   fireplace/wood stove on the right wall. It then continues panning left, showing the
    ↪   stairs leading back up to the initial doorway.
    *   **Further down the room (0:46-0:50):** The camera moves towards the back of the kitchen.
    ↪   We see the main kitchen counter with the stove and range hood. Further down, past the
    ↪   kitchen island, there's a doorway leading into a pantry/room lined with shelves filled
    ↪   with items (jars, cans, etc.). This is the ""shelf"" mentioned in the prompt.
    *   **Moving Around Kitchen (0:51-1:16):** The camera moves around the kitchen island and
    ↪   along the main kitchen counter, eventually looking back towards the dining area and the
    ↪   pantry doorway.

3.  **Identify the Objects:**
    *   **Table:** The prominent wooden dining table in the foreground/middle of the long room.
    *   **Shelf:** The shelving unit visible through the doorway at the far end of the kitchen
    ↪   area (the pantry).

4.  **Determine the ""Closest Points"":**
    *   The closest point of the *table* to the shelf would be the end of the table nearest the
    ↪   kitchen area.
    *   The closest point of the *shelf* to the table would be the front edge of the shelves
    ↪   within the pantry doorway.
```

```
5.  **Estimate the Distance:**
    *   **Visualize the Space:** The room is long and relatively narrow. It appears to be an
    ↪   open-plan kitchen/diner.
    *   **Break Down the Distance:**
        *   From the end of the dining table to the *start* of the kitchen island/peninsula.
        ↪   This looks like maybe 2-3 meters of clear floor space.
        *   The length of the kitchen island/peninsula. This looks roughly 1.5-2 meters long.
        *   The space between the kitchen island and the main kitchen counter along the wall.
        ↪   This is the main walkway, perhaps 1-1.5 meters wide. (Though the distance we care
        ↪   about is *along* the length of the room).
        *   From the *end* of the kitchen island (closest to the pantry) to the pantry doorway.
        ↪   This looks like another 2-3 meters.
    *   **Refined Estimation:** Let's trace the path *straight* from the table end to the
    ↪   pantry door.
        *   The table seems to end roughly parallel to the fireplace (0:09).
        *   The kitchen island starts a bit further down (0:15). Let's say 2 meters from table
        ↪   end to island start.
        *   The island itself looks about 1.5 meters deep (front to back, relative to the
        ↪   camera's initial view). Let's estimate its length (along the room) as maybe 2
        ↪   meters (0:21 shows its length better).
        *   From the end of the island to the pantry doorway (0:48-0:50). This distance looks
        ↪   similar to the gap between the table and the island, maybe another 2-3 meters.
    *   **Total Estimation:** 2m (table to island) + 2m (island length, effectively part of the
    ↪   distance to cover) + 3m (island end to pantry door) = approximately 7 meters.

6.  **Consider Scale Clues:**
    *   **Floorboards:** Standard floorboards are often around 10-15 cm wide. Counting boards
    ↪   is difficult due to perspective and movement, but it suggests a significant length.
    *   **Doors:** Standard door height is ~2m. The pantry doorway looks standard.
    *   **Counter Depth:** Standard kitchen counters are ~60cm deep. The island looks deeper,
    ↪   maybe 90cm-1m.
    *   **Table Length:** A table seating 6-8 people is typically 1.8m to 2.4m long.
    *   **Room Length:** Combining these, the dining area looks maybe 4m long, the kitchen area
    ↪   (island + walkway + counter space before pantry) looks another 4-5m long. Totaling 8-9m
    ↪   from the window end to the pantry door. The table occupies a portion of the dining area.

7.  **Refine based on Scale:** If the whole room (dining + kitchen) is ~8-9m long, and the
↪   table takes up maybe 2.5m of that length, starting maybe 1m from the window wall, then the
↪   end of the table is roughly 3.5m into the room. The pantry door is at the far end (8-9m).
↪   The distance would be 8.5m (total est.) - 3.5m (table end) = 5 meters.

8.  **Reconcile Estimates:** My first estimate was ~7m, my second based on rough room
↪   dimensions is ~5m. Let's re-examine the middle ground.
    *   Look at 0:10. The table ends.
    *   Look at 0:15. The camera has moved forward. The kitchen area starts properly.
    *   Look at 0:48. We are near the pantry door.
    *   The distance from the end of the table (around 0:10-0:11) to the pantry doorway
    ↪   (visible clearly around 0:48) seems substantial. The kitchen island occupies a
    ↪   significant chunk of this space. Let's use the island as a reference. It looks around
    ↪   2m long. The gap *before* the island (from the table) looks similar in length to the
    ↪   gap *after* the island (to the pantry door). Let's say Gap1 = 2m, Island = 2m, Gap2 =
    ↪   2.5m. Total distance = Gap1 + Island + Gap2 = 2 + 2 + 2.5 = 6.5 meters. This feels more
    ↪   plausible than 5m, given the visual length.

9.  **Final Check:** View the transition from 0:11 (near table end) to 0:48 (near pantry). It's
↪   a considerable walk/pan. 6-7 meters seems reasonable for the space shown, accounting for
↪   the kitchen island and walkways.

10. **Format the Answer:** State the estimated distance clearly in meters, acknowledging it's
↪   an estimate based on visual cues."

---

## Guide for Writing “Thoughts” (Step-by-Step Reasoning for Video QA)

**Purpose:**
These “thoughts” are written to show a clear, step-by-step reasoning process for answering
↪   questions based on video content. They help make the reasoning *transparent*, *logical*,
↪   and *explainable*.
```

```
---

### Structure of a Good "Thoughts"

1. **Clarify the Question**
   - Start by rephrasing or clearly identifying what the question is asking.
   - Example: *“The goal is to determine which object appears first in the video.”*

2. **Observe the Video and Gather Evidence**
   - Mention specific timestamps (e.g., *“At 00:15, the sink becomes visible”*).
   - Describe what is seen, including spatial positions, sizes, and context.
   - Use clear bullet points or numbered lists if multiple observations are needed.

3. **Use Common Knowledge and Estimation**
   - Use real-world dimensions or typical object sizes for scale comparisons.
   - Example: *“Standard kitchen counters are about 60 cm deep, so the sink appears slightly shorter than that.”*

4. **Compare and Reason**
   - Compare objects, locations, appearances, or distances based on what is seen.
   - Discuss alternatives and edge cases.
   - Example: *“The stool is at the foot of the bed, so it's farther from the toilet than the bed's head, which is near the door.”*

5. **Rule Out Incorrect Options (if applicable)**
   - If a multiple-choice question is given, explain why incorrect choices can be eliminated.
   - Example: *“Option A lists the oven before the table, but the oven appears much later.”*

6. **State a Final Answer**
   - Conclude clearly and confidently.
   - Example: *“Therefore, the correct order is: sofa, table, TV monitor, oven.”*

---

### Style and Tone

- Write in **clear, logical English**.
- Keep the tone **neutral and analytical**.
- Use **present tense** when describing the video content.
- Be detailed but not overly wordy—aim for **clarity over length**.

---

### Bonus Tips

- If uncertain, show multiple interpretations and explain which is more likely.
- Use indentation or line breaks to separate steps and make the reasoning easier to follow.
- Imagine you are teaching someone else *how* to think through the video—be explicit.

---

## Question

**What is the total number of tables shown in this scene?**


Please answer the question above by following the process and format below:

---

## Thoughts

Your reasoning process

---

## Response

Your answer based on the reasoning process
```

```
---

## Final Answer

\[
\boxed{Your final answer}
\]

---
```

In the prompt template, there is a placeholder line:

What is the total number of tables shown in this scene?

This placeholder question should be replaced with the actual question corresponding to each video in the dataset.

E Supplementary Material for Section 4.5 (Additional Probing Experiments)

The figures and tables supplementing Section 4.5 (Additional Probing Experiments) are provided in this chapter/section.

	ViCA-Base-7B w/o time_inst.	**ViCA-7B** w/o time_inst.	**ViCA-7B** w/ time_inst.
Numerical Answer			
Obj. Count	65.6	**68.8**	68.4 (-0.4)
Abs. Dist.	51.3	**57.0**	56.9 (-0.1)
Obj. Size	74.9	79.2	**79.3**
Room Size	67.1	75.1	**77.6**
Multiple-Choice Answer			
Rel. Dist.	52.0	**58.5**	58.0 (-0.5)
Rel. Dir.	32.6	**42.6**	40.8 (-1.8)
Route Plan	28.4	**34.5**	28.9 (-5.6)
Appr. Order	**70.9**	68.8	69.1 (-1.8)
Average	55.4	**60.6**	59.9 (-0.7)

Table 5: Comparison of ViCA variants with and without time instruction. The third column (ViCA-7B w/ time_instruction) is compared against the better of the two baselines (ViCA-Base-7B and ViCA-7B w/o time_instruction). The values in parentheses indicate the performance difference from the best baseline for each metric. Bold highlights denote the best overall score among the three models.

	ViCA-Base-7B 64 frames	**ViCA-7B** 64 frames	**ViCA-7B** 128 frames
Numerical Answer			
Obj. Count	65.6	**68.8**	68.1 (-0.7)
Abs. Dist.	51.3	**57.0**	54.2 (-2.8)
Obj. Size	74.9	79.2	**79.3**
Room Size	67.1	**75.1**	74.5 (-0.7)
Multiple-Choice Answer			
Rel. Dist.	52.0	**58.5**	58.3 (-0.1)
Rel. Dir.	32.6	**42.6**	40.4 (-2.2)
Route Plan	28.4	**34.5**	30.4 (-4.1)
Appr. Order	**70.9**	68.8	63.1 (-7.8)
Average	55.4	**60.6**	58.5 (-2.0)

Table 6: Comparison of ViCA variants across different frame settings. The third column (ViCA-7B 128 frames) is compared against the better of the two baselines (ViCA-Base-7B and ViCA-7B with 64 frames). The values in parentheses indicate the performance difference from the best baseline for each metric. Bold highlights denote the best overall score among the three models.

F *Thought-Response-Final Answer* Examples generated by ViCA-7B-Thinking

In this section, we present several examples illustrating the *Thought-Response-Final Answer* structure produced by our ViCA-7B-Thinking model.

A representative output produced by ViCA-7B-Thinking

think

1. **Understand the Goal:** The user wants to know the total number of chairs visible in the room shown in the video.
2. **Define "Chair":** I need to identify objects designed for seating one person, typically with a back and legs or a base. This includes dining chairs, office chairs, armchairs, etc.
3. **Scan the Video for Chairs:**
   * **0:00 - 0:05:** The camera shows a floor, a poster, and then pans up towards a doorway leading into a kitchen area. No chairs are visible yet.
   * **0:06 - 0:10:** The camera enters the kitchen area. A stainless steel trash can is visible. No chairs.
   * **0:11 - 0:14:** The camera pans across the kitchen counter, sink, and appliances. No chairs.
   * **0:15 - 0:18:** The camera pans right along the kitchen counter. A built-in oven is visible. Next to the oven, there is a tall wooden cabinet unit. At the base of this unit, on the left side (from the viewer's perspective), a single brown chair with a metal frame is clearly visible. Let's call this **Chair 1**.
   * **0:19 - 0:22:** The camera continues panning right along the kitchen counter. Next to the oven unit, there is a small round white table with two matching chairs. One chair is visible. Let's call this **Chair 2**.
   * **0:23 - 0:27:** The camera pans further right, past the table, showing a TV mounted on the wall, a desk, and a fan. No additional chairs here.
   * **0:28 - 0:31:** The camera pans down from the desk area, revealing a brown leather sofa. On the sofa, there are two patterned cushions that could potentially be used as armrests or small seats, but they are not typical chairs. I will not count these.
   * **0:32 - 0:35:** The camera pans across the sofa and a coffee table. No new chairs.
   * **0:36 - 0:40:** The camera pans back towards the entrance area, showing the trash can again. No chairs.
   * **0:41 - End:** The camera moves around the room, revisiting areas already scanned. No new chairs are revealed.

4. **Consolidate Findings:** I identified two distinct chairs:
   * Chair 1: A brown chair under the tall wooden cabinet next to the oven.
   * Chair 2: A brown chair at the small round table near the TV/desk.
5. **Final Count:** There are 2 chairs visible in the room.

response

Based on the video walkthrough of the room:

1. At approximately 0:16, a brown chair is visible tucked under a tall wooden cabinet next to the oven.
2. At approximately 0:20, a second brown chair is visible at a small round table near the TV and desk.

No other chairs are clearly visible in the room. Therefore, there are 2 chairs in the room.

final

\[
\boxed{2}
\]

Figure 7: A representative *Thought-Response-Final Answer* example generated by ViCA-7B-Thinking. The model demonstrates its ability to externalize spatial reasoning in natural language.

A representative output produced by ViCA-7B-Thinking

think

1. **Understand the Goal:** The question asks to identify which of the four listed objects (bed, chair, trash bin, radiator) is closest to the door shown in the video, measuring from the closest points of each object to the door.

2. **Identify the Objects and Door:**
    * **Door:** A white door is visible on the left side of the room, near the start and end of the pan (e.g., 0:03-0:04, 0:28-0:29).
    * **Bed:** A loft bed is positioned along the wall opposite the door (e.g., 0:00-0:02, 0:17-0:19).
    * **Chair:** A black office chair is located at a desk, which is positioned near the window, away from the door (e.g., 0:10-0:12, 0:25-0:26).
    * **Trash Bin:** A small, light-colored trash can is located under the desk, also near the window, away from the door (e.g., 0:10-0:12, 0:25-0:26).
    * **Radiator:** A white baseboard heater (radiator) is located under the window, near the desk and chair (e.g., 0:10-0:12, 0:25-0:26).

3. **Analyze Spatial Relationships:**
    * The video pans across the room, showing the layout.
    * The door is on one wall.
    * The bed is on the opposite wall, further down the room from the door.
    * The desk, chair, and trash bin are grouped together near the window, which is on the same wall as the door, but further down the room than the door.
    * The radiator is located under the window, also on the same wall as the door, but further down the room than the desk/chair area.

4. **Estimate Relative Distances to the Door:**
    * **Bed:** The bed is clearly the furthest object from the door among the options.
    * **Chair/Trash Bin/Radiator:** These three items are clustered together near the window. The distance from the door to the chair, trash bin, or radiator involves traversing the width of the room past the desk area.
    * **Chair:** The chair is positioned at the desk. The closest point of the chair to the door would be its back edge. Visually comparing the position of the chair back (around 0:11) to the position of the door (around 0:03), the chair back appears significantly closer to the door than the bed is. The distance looks like roughly half the width of the room.
    * **Trash Bin:** The trash bin is next to the chair/desk. Its closest point to the door is also likely around half the room's width, similar to the chair.
    * **Radiator:** The radiator is under the window, also near the chair/desk. Its closest point to the door is also likely around half the room's width, similar to the chair and trash bin.

5. **Compare Closest Points:** All three items (chair, trash bin, radiator) are relatively close to the door. However, the chair is positioned at the desk, which is closer to the door than the bed is. The chair's back edge seems to be the closest point among the three options to the door.

6. **Conclusion:** Based on the visual layout and relative positions, the chair is the closest object to the door among the given choices.

response

The video shows a room containing a bed, a desk with a chair, a trash bin, and a radiator. The door is located on one wall. The bed is positioned along the opposite wall from the door. The desk, chair, and trash bin are clustered together near the window, which is on the same wall as the door but further down the room. The radiator is located under the window, also on the same wall but further down the room. Comparing the distances, the bed is clearly the furthest from the door. The chair, trash bin, and radiator are all relatively close, but the chair is positioned at the desk, which is closer to the door than the bed. Therefore, the chair is the closest object to the door.

final

```
\[
\boxed{B. chair}
\]
```

Figure 8: A representative *Thought-Response-Final Answer* example generated by ViCA-7B-Thinking. The model demonstrates its ability to externalize spatial reasoning in natural language.

A representative output produced by ViCA-7B-Thinking

think

1. **Understand the Goal:** The question asks to identify which of the four listed objects (closet, bed, trash bin, mirror) is closest to the TV monitor shown in the video, measuring from the closest points.

2. **Identify the TV Monitor:** The TV monitor is a black flat-screen television located on top of a wooden dresser. It's clearly visible at timestamps like 0:25-0:26 and 1:08-1:10.

3. **Locate the Objects Relative to the TV:**
    * **TV Location:** On top of a wooden dresser.
    * **Bed:** The bed is positioned next to the dresser, along the same wall. The side of the bed frame is very close to the dresser where the TV sits (visible at 0:25, 1:08).
    * **Closet:** A large white closet unit is visible further down the room, near the window (seen around 0:17-0:19). This is significantly further away from the TV than the bed or the dresser it sits on.
    * **Trash Bin:** A green trash bin is located on the floor near the dresser, closer to the foot of the bed (seen at 0:24, 1:07). It is also relatively close to the TV, but perhaps slightly further than the bed frame itself.
    * **Mirror:** A large mirror is mounted on the wall above the dresser (seen at 0:25, 1:08). The bottom edge of the mirror is directly above the TV. The closest point of the mirror is vertically above the TV.

4. **Compare Distances from the TV:**
    * **TV to Bed:** The distance is primarily horizontal, between the side of the bed and the side of the dresser. This appears to be a very small gap, likely less than half a meter.
    * **TV to Trash Bin:** The distance involves both horizontal and vertical components. The bin is on the floor, so the horizontal distance is greater than the bed-to-dresser distance. The vertical distance is also significant, as the bin is below the TV. The total distance seems greater than the bed-to-dresser distance.
    * **TV to Mirror:** The distance is primarily vertical, from the bottom of the TV to the bottom of the mirror. While the horizontal distance between the TV and the dresser is small, the vertical distance from the TV to the mirror might be slightly larger than the horizontal distance to the bed, depending on the exact placement. However, the mirror is directly above the TV, making the closest point of the mirror vertically adjacent to the TV.

5. **Determine the Closest Object:** Comparing the closest points:
    * The bed frame is immediately adjacent to the dresser.
    * The mirror is immediately adjacent to the TV via the dresser.
    * The trash bin is further away horizontally and vertically.
    * The closet is furthest away.

    Between the bed and the mirror, the bed frame appears to be the closest object because it is physically positioned right next to the dresser, which the TV rests upon. The mirror is above the TV, but the bed frame is beside the dresser.

6. **Select the Correct Option:** Based on the visual evidence, the bed is the closest object to the TV among the given choices.

response

The video shows a bedroom containing a bed, a dresser with a TV on top, a closet, and a trash bin. The TV is located on the dresser (visible at 0:25-0:26, 1:08-1:10). We need to determine which object (closet, bed, trash bin, mirror) is closest to the TV.

1. **TV Location:** The TV is on top of a dresser.
2. **Bed Location:** The bed is positioned immediately next to the dresser, along the same wall (visible at 0:25, 1:08). The side of the bed frame is very close to the dresser.
3. **Closet Location:** The closet is located further down the room, near the window (visible at 0:17-0:19), which is significantly farther from the TV than the bed or the dresser.
4. **Trash Bin Location:** The trash bin is on the floor near the dresser, closer to the foot of the bed (visible at 0:24, 1:07). It is also relatively close to the TV, but perhaps slightly further than the bed frame.
5. **Mirror Location:** The mirror is mounted on the wall above the dresser (visible at 0:25, 1:08). The bottom edge of the mirror is directly above the TV. The closest point of the mirror is vertically adjacent to the TV.

Comparing the distances from the closest point of each object to the closest point of the TV:
* The bed frame is immediately adjacent to the dresser, which the TV rests upon. The horizontal distance is minimal.
* The mirror is immediately adjacent to the TV via the dresser. The vertical distance is minimal.
* The trash bin is further away horizontally and vertically.
* The closet is furthest away.

Therefore, the bed is the closest object to the TV among the given options.

final

\[
\boxed{B. bed}
\]

Figure 9: A representative *Thought-Response-Final Answer* example generated by ViCA-7B-Thinking. The model demonstrates its ability to externalize spatial reasoning in natural language.

A representative output produced by ViCA-7B-Thinking

think

1. **Understand the Goal:** The user wants to determine the chronological order in which four specific categories of objects (refrigerator, closet, door, table) first appear in the video.
2. **Define Categories:**
    * Refrigerator: An appliance for cooling food.
    * Closet: A storage space, often built-in or freestanding, for clothes and other items.
    * Door: A hinged or sliding panel used to enclose or separate spaces.
    * Table: A piece of furniture with a flat top and one or more legs, providing a level surface for eating, writing, or placing things on.
3. **Scan Video for First Appearances:**
    * **0:00 - 0:05:** The video starts showing a corner of a room. A small wooden bookshelf is visible. Next to it, a black sofa is partially visible. On the floor next to the sofa, a pair of red sandals and a pair of blue flip-flops are visible. Further back, a cardboard box labeled "TV" is visible. A whiteboard is also visible on the floor. No refrigerator, closet, door, or table yet.
    * **0:06 - 0:08:** The camera pans right. A wooden bunk bed comes into view. Underneath the bunk bed, a desk is visible. A black office chair is positioned at this desk. A mini-fridge is located under the desk. This is the first appearance of a refrigerator (mini-fridge).
    * **0:09 - 0:10:** The camera continues panning right along the desk area. A wooden wardrobe unit becomes visible against the wall. A towel hangs on the side of this wardrobe. This is the first appearance of a closet.
    * **0:11 - 0:12:** The camera pans further right, past the wardrobe. A wooden door is visible, leading out of the room. This is the first appearance of a door.
    * **0:13 - 0:14:** The camera pans across the room again, showing the previously identified items (wardrobe, door, mini-fridge, desk) from different angles. The desk itself is clearly visible. This is the first appearance of a table (desk).
4. **Establish the Sequence:**
    * Refrigerator (Mini-fridge): ~0:07
    * Closet (Wardrobe): ~0:09
    * Door: ~0:11
    * Table (Desk): ~0:13
5. **Order the Items:** The observed sequence is: Refrigerator, Closet, Door, Table.
6. **Compare with Options:**
    * A. closet, table, door, refrigerator - Incorrect.
    * B. refrigerator, closet, door, table - Correct. Matches the observed sequence.
    * C. refrigerator, table, closet, door - Incorrect.
    * D. door, table, refrigerator, closet - Incorrect.

response

Based on the video analysis:
1. The **refrigerator** (a mini-fridge) first appears around 0:07.
2. The **closet** (a wooden wardrobe) first appears around 0:09.
3. The **door** (the entrance door) first appears around 0:11.
4. The **table** (the desk) first appears around 0:13.

The chronological order of first appearance is: refrigerator, closet, door, table. This corresponds to option B.

final

\[
\boxed{B. refrigerator, closet, door, table}
\]

Figure 10: A representative *Thought-Response-Final Answer* example generated by ViCA-7B-Thinking. The model demonstrates its ability to externalize spatial reasoning in natural language.