Towards Visuospatial Cognition via Hierarchical Fusion of Visual Experts[*]

Qi Feng[1]

[1]Kyoto University

feng.qi.45n@st.kyoto-u.ac.jp

https://github.com/nkkbr/ViCA

Abstract

While Multimodal Large Language Models (MLLMs) excel at general vision-language tasks, *visuospatial cognition*—reasoning about spatial layouts, relations, and dynamics—remains a significant challenge. Existing models often lack the necessary architectural components and specialized training data for fine-grained spatial understanding. We introduce **ViCA2** (**Vi**suospatial **C**ognitive **A**ssistant 2), a novel MLLM designed to enhance *spatial reasoning*. **ViCA2** features a dual vision encoder architecture integrating `SigLIP` for semantics and `Hiera` for spatial structure, coupled with a token ratio control mechanism for efficiency. We also developed **ViCA-322K**, a new large-scale dataset with over 322,000 spatially grounded question-answer pairs for targeted instruction tuning. On the challenging **VSI-Bench** benchmark, our **ViCA2-7B** model achieves a state-of-the-art average score of 56.8, significantly surpassing larger open-source models (e.g., **LLaVA-NeXT-Video-72B**, 40.9) and leading proprietary models (**Gemini-1.5 Pro**, 45.4). This demonstrates the effectiveness of our approach in achieving strong visuospatial intelligence with a compact model. We release **ViCA2**, its codebase, and the **ViCA-322K** dataset to facilitate further research.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have led to remarkable progress in visual-language tasks such as image captioning, visual question answering (VQA), and video understanding. By aligning vision and language modalities through large-scale pretraining, models such as Flamingo(Alayrac et al., 2022), LLaVA(Liu et al., 2024), and their successors have shown strong generalization capabilities across a wide
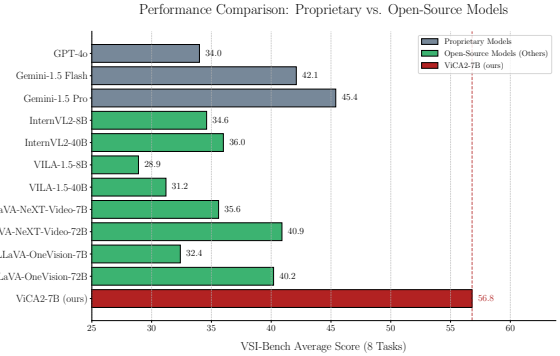

Figure 1: **Average performance comparison on VSI-Bench.** We compare ViCA2-7B with both proprietary API-based models and leading open-source multimodal models on the VSI-Bench benchmark, which evaluates visuospatial reasoning across 8 tasks. Despite having only 7B parameters, ViCA2-7B significantly outperforms all other models, achieving an average score of **56.8**, surpassing the best-performing proprietary model (Gemini-1.5 Pro(Team et al., 2024), 45.4) and the strongest open-source baseline (LLaVA-NeXT-Video-72B, 40.9) by a large margin. This highlights the effectiveness of our dual vision encoder design and hierarchical fusion approach for spatial understanding.

range of benchmarks. More recently, models like LLaVA-OneVision(Li et al., 2024) and LLaVA-NeXT-Video(Zhang et al., 2024) have extended these capabilities to multi-image and video scenarios, enabling unified processing of diverse visual inputs via high-resolution token streams, instruction tuning, and stronger visual backbones such as SigLIP(Zhai et al., 2023).

While these advances are significant, an important aspect of human-level visual intelligence remains largely underexplored: *visuospatial cognition*—the capacity to construct, retain, and reason over spatially coherent internal representations of complex environments. In contrast to general scene understanding, visuospatial reasoning requires fine-grained comprehension of object layouts, spatial relations, temporal order, and geometric attributes.

---

[*]This is a technical report and a draft version. Work in progress.

1

Such capabilities are critical for downstream applications including indoor navigation, assistive robotics, video summarization, and embodied question answering.

Recent work has begun to probe this capacity. Yang et al. (2024) introduced VSI-Bench, a challenging benchmark targeting spatial intelligence through a suite of question types over real indoor videos. Their analysis revealed that even state-of-the-art MLLMs like LLaVA-OneVision-72B and LLaVA-NeXT-Video-72B struggle significantly on spatially grounded tasks. Although proprietary models such as Gemini-1.5 Pro show stronger spatial abilities, these systems remain closed-source and resource-heavy. A key bottleneck identified in open models is their reliance on a single semantic vision encoder, which often fails to capture high-resolution spatial layouts due to low input resolutions and an architecture optimized for global semantic alignment. Moreover, the lack of task-specific training data targeting spatial reasoning further limits their capacity to internalize structured spatial knowledge.

Our Contributions In this paper, we propose **ViCA2** (**Vi**suospatial **C**ognitive **A**ssistant 2), a lightweight yet powerful multimodal framework designed to advance visuospatial understanding in both image and video settings. Our approach builds upon **LLaVA-OneVision** and introduces several key innovations:

1. **Dual Vision Encoder Architecture for Enhanced Spatial Awareness:** We propose a novel dual vision encoder architecture that integrates SigLIP for global semantic abstraction with Hiera—a hierarchical visual encoder from SAM2(Ravi et al., 2024)—for modeling fine-grained spatial structure. The two feature streams are individually projected into the language model's embedding space and fused prior to decoding, enabling joint reasoning over semantics and spatial cues.

2. **Efficient Token Ratio Control Mechanism:** We develop a token ratio control mechanism that adjusts frame sampling density, feature extraction depth, and spatial pooling stride for each encoder. This offers explicit control over the balance between semantic abstraction and spatial detail, particularly crucial under memory constraints for handling high-resolution or long-duration visual inputs.

3. **State-of-the-Art Performance with a Compact Model:** Our experimental results demonstrate that **ViCA2-7B**, despite using a relatively compact 7B language model, achieves state-of-the-art performance on the challenging VSI-Bench benchmark (Figure 1). Specifically, **ViCA2-7B** significantly outperforms much larger open-source models, including LLaVA-OneVision-72B and LLaVA-NeXT-Video-72B, and even surpasses the proprietary Gemini-1.5 Pro in several spatial reasoning categories, highlighting the efficacy of our architectural and data-centric design choices.

To facilitate further research in visuospatial cognition, we open-source our model weights, codebase, and the **ViCA-322K** dataset.

2 Related work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) integrate vision and language understanding within a unified architecture by coupling pretrained language models with visual encoders through lightweight projection modules. Early models such as Flamingo and MiniGPT-4 (Zhu et al., 2023) demonstrated the effectiveness of aligning image features with language inputs. Recent open-source models like LLaVA-OneVision and Video-LLaMA (Zhang et al., 2023a) extend this capability to multi-image and video understanding using variable-length visual sequences. Meanwhile, proprietary systems like Gemini 2.5 Pro (Kavukcuoglu, 2025) and GPT-4V (OpenAI, 2023) exhibit strong performance on vision-language tasks, albeit with closed architectures.

2.2 Visual Encoders for Multimodal Learning

Most MLLMs rely on visual encoders pretrained on image-text contrastive learning, such as CLIP (Radford et al., 2021), EVA-CLIP (Sun et al., 2023), or SigLIP. While effective for global semantics, these models are typically trained on low-resolution inputs, limiting their ability to capture fine-grained spatial details. Recent solutions adopt higher resolutions or tiling strategies (e.g., LLaVA-UHD (Guo et al., 2024), InternLM-XComposer (Zhang et al., 2023b)) to improve spatial coverage, but these approaches do not fundamentally enhance spatial reasoning. In contrast, encoders pretrained on dense prediction tasks—such as Hiera (Ryali et al., 2023), a hierarchical vision backbone used
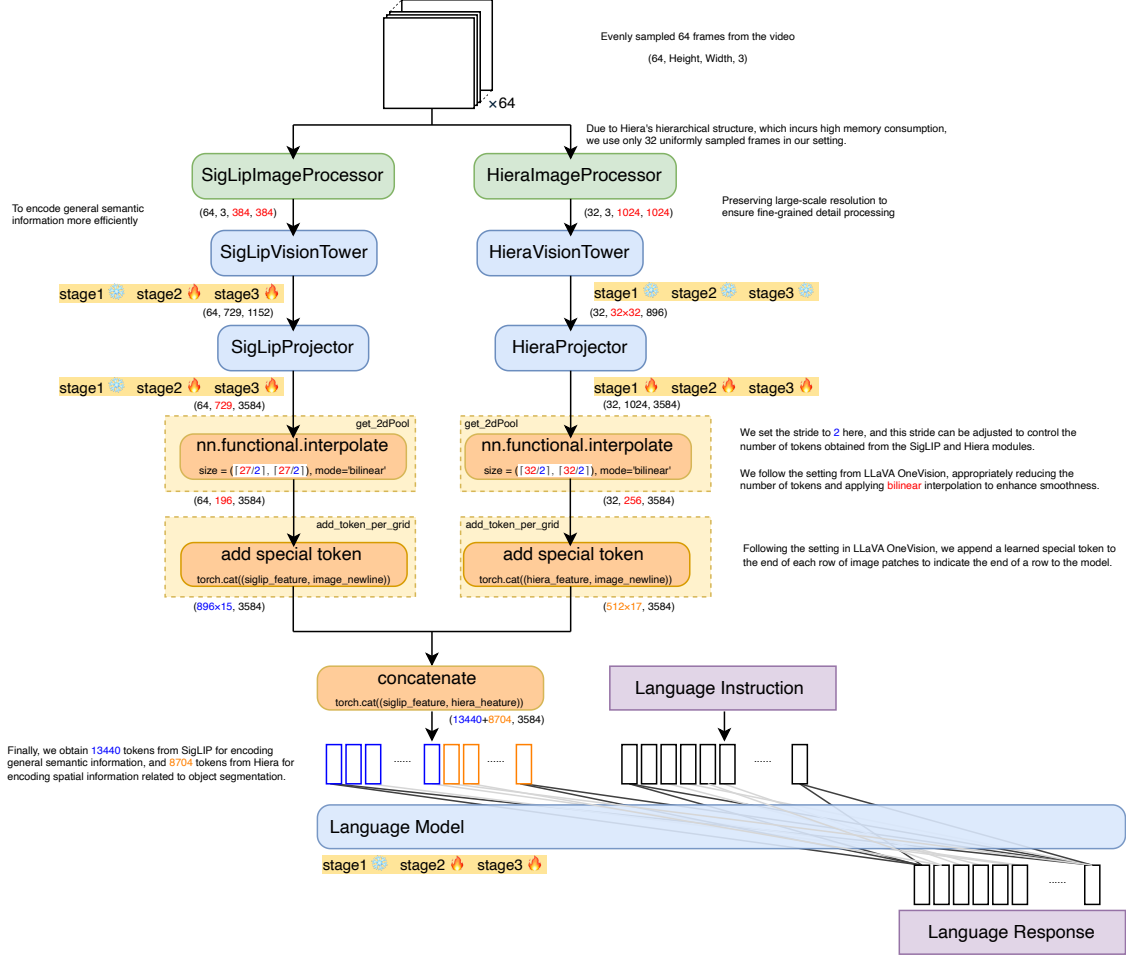
Figure 2: **An overview of the ViCA2 architecture.** The model integrates SigLIP for global semantic understanding and Hiera for spatial structure encoding, with a token ratio control mechanism to balance expressiveness and memory efficiency. A complete version of the architecture diagram is provided in the Appendix(Figure 6).

in SAM2—offer stronger spatial sensitivity. Our ViCA2 model leverages a dual-encoder design combining SigLIP for global semantics and Hiera for spatial structure, with independent projection heads and token ratio control to balance expressiveness and efficiency.

2.3 Visuospatial Reasoning in MLLMs

While MLLMs excel at general vision-language tasks, they remain limited in visuospatial reasoning, which involves understanding object positions, spatial relations, and temporal dynamics. VSI-Bench highlights this gap by evaluating models on spatial questions grounded in real-world indoor videos. Even advanced models such as LLaVA-OneVision-72B and Gemini-1.5 Pro struggle with consistent spatial understanding due to inadequate training supervision and architectural constraints.

3 ViCA2

3.1 The Overview

LLaVA-OneVision is a state-of-the-art open-source large multimodal model that unifies single-image, multi-image, and video understanding within a single architecture. It integrates a strong visual encoder, `SigLIP`, with a powerful language model, `Qwen-2`, and employs a lightweight MLP-based projection layer to bridge vision and language modalities. The model achieves high performance across a wide range of visual-language benchmarks by leveraging a unified representation strategy and a large-scale, high-quality training corpus. Its success demonstrates the effectiveness of multi-stage instruction tuning and the benefits of treating visual input—regardless of type—as sequences of visual tokens.

Building on LLaVA-OneVision, our proposed model, ViCA2(Figure 2), extends the multimodal capabilities by introducing a second vision encoder derived from `Hiera`, the hierarchical visual backbone used in SAM2. While LLaVA-OneVision utilizes SigLIP to capture global semantic features, our ViCA2 further incorporates the spatially structured features extracted by `Hiera` to enhance fine-grained understanding of object-centric and region-level information. The outputs from both vision encoders are individually projected into the language model's embedding space via two separate linear projectors, and then fused before being consumed by the decoder. This dual-encoder architecture enables the model to jointly reason over high-level semantics and spatial visual cues, significantly improving its performance on spatially grounded tasks.

3.2 Dual Vision Encoders for Spatially-Grounded Understanding

We employ `SigLIP` as one of the vision encoders in our model to extract high-level semantic representations from visual inputs. SigLIP benefits from contrastive pretraining on large-scale image-text pairs. It produces compact and expressive visual embeddings that align well with natural language, making it highly effective for tasks such as image captioning and general visual question answering. However, models employing a CLIP-like contrastive learning paradigm, including `SigLIP`, are primarily optimized for global semantic alignment rather than fine-grained spatial reasoning. This limitation is particularly evident in phenomena such as "CLIP-blind pairs" (Tong et al., 2024), where visually distinct images are mapped to similar embeddings due to semantic proximity. Furthermore, the relatively low-resolution input constraints (e.g., 336×336 or 512×512) make it difficult for these models to preserve detailed local features, which are essential for dense OCR, visual grounding, and region-level understanding.

To address these limitations, we introduce a second vision encoder based on `Hiera`, the hierarchical visual backbone from SAM2.1 (specifically, the `sam2.1_hiera_b+` variant), to complement the `SigLIP` encoder. Unlike the ViT-H encoder used in the original SAM, `Hiera` adopts a multi-stage hierarchical structure that extracts multiscale visual features through progressive downsampling, similar in spirit to classical feature pyramids. Pretrained with Masked Autoencoding (MAE), `Hiera`

is particularly well-suited for spatially dense tasks, enabling robust recognition of objects and regions that vary in size and location. This is especially beneficial for video or multi-image scenarios, where spatial consistency and resolution variance are critical. Compared to ViT-H, `Hiera` not only improves segmentation and localization accuracy but also reduces computation time significantly, making it a more efficient and spatially aware encoder for large-scale multimodal models.

In our architecture, the outputs from both `SigLIP` and `Hiera` are processed through separate linear projectors to align their feature spaces with the language model's embedding space. The projected features are then concatenated along the sequence dimension before being fed into the language decoder. This dual-encoder setup allows our model to jointly reason over global semantic content and spatially grounded visual cues, thereby enhancing its spatial understanding capabilities. By leveraging `SigLIP`'s strength in semantic abstraction and `Hiera`'s ability to model fine-grained spatial structures, the model is better equipped to tackle challenging tasks that require object-level reasoning, region grounding, and multi-object spatial relation understanding across both images and videos.

3.3 Video Feature Extraction and Dual Encoder Token Balancing

To effectively process video inputs, we uniformly sample $N_{\text{total}} = 64$ frames from each video. These frames are fed into two distinct visual encoders: `SigLIP` and `Hiera`. While `SigLIP` processes all 64 frames to extract high-level semantic features, the computational cost of `Hiera`—particularly its hierarchical structure and high spatial resolution in early stages—necessitates a more selective approach. To strike a balance between expressiveness and memory usage, we feed only $N_{\text{hiera}} = 32$ uniformly sampled frames into `Hiera`.

Each frame destined for `Hiera` is first preprocessed using the SAM2.1 `HieraImageProcessor` and then passed through the `Hiera` backbone, which comprises four hierarchical stages. As the stage depth increases, spatial resolution is progressively reduced while channel capacity increases. Early stages preserve fine-grained visual details such as edges and textures, whereas later stages encode global semantic structures such as object identities and scene layouts. In our configuration, we extract features from Stage 4, the deepest layer,

which captures the most abstract and semantically rich information. These features are then passed through a 2D spatial pooling module with stride $S_{\text{pool}} = 2$, implemented using bilinear interpolation, following the token downsampling strategy used in LLaVA-OneVision. This reduces the token count per frame while preserving the spatial continuity of features, enabling a higher number of frames to be processed within the same memory budget.

To further enhance spatial layout awareness, we append a learned row-wise special token at the end of each row in the resulting 2D token grid, providing the model with positional cues about horizontal boundaries. After this transformation, we obtain $T_{\text{siglip}} = 13,440$ tokens from SigLIP and $T_{\text{hiera}} = 8,704$ tokens from Hiera. These two token sequences are concatenated along the sequence dimension and fed into the language decoder. The token ratio $T_{\text{siglip}} : T_{\text{hiera}} = 1.54$ reflects a deliberate design choice to balance global semantics and spatial precision.

We introduce a token ratio control strategy based on a configurable triplet $(N_{\text{hiera}}, S_{\text{stage}}, S_{\text{pool}})$, which respectively denote: the number of frames passed to Hiera, the selected stage for feature extraction, and the spatial pooling stride. This mechanism enables fine-grained control over the balance between semantic and spatial representations. A high proportion of SigLIP tokens tends to bias the model toward global semantic alignment, while an excess of Hiera tokens increases spatial sensitivity but may dilute language grounding. Our chosen configuration $(32, 4, 2)$ offers a favorable trade-off, enhancing the model's ability to perform spatially grounded reasoning tasks without exceeding hardware limitations.

4 Training

4.1 Initialization of Model Parameters

Our model is initialized based on the publicly released checkpoint `lmms-lab/LLaVA-Video-7B-Qwen2` on Hugging Face. Specifically, the language model is initialized from `Qwen2-7B`, and the primary vision encoder, SigLIP, is initialized from the pretrained weights of `google/siglip-so400m-patch14-384`. To enhance spatial reasoning capabilities, we incorporate a second vision encoder, Hiera, which is initialized from the Hiera module in `facebook/sam2.1-hiera-base-plus`. The

associated linear projector that maps Hiera features into the language model's embedding space is randomly initialized using Xavier initialization. The projected features from both SigLIP and Hiera are directly concatenated along the sequence dimension without introducing any additional parameterized fusion module.

4.2 Training Strategy and Data

We adopt a three-stage training strategy to progressively adapt our model—particularly the newly introduced Hiera encoder and its associated projection module—for spatially grounded multimodal understanding. Each stage targets specific alignment challenges and uses carefully selected datasets with different scales and characteristics.

In the first stage, we focus on calibrating the newly added Hiera projector, a two-layer MLP that is randomly initialized and initially incompatible with the language model's input space. To facilitate coarse visual-language alignment, we employ the `liuhaotian/LLaVA-CC3M-Pretrain-595K` dataset, which contains large-scale image-text pairs with relatively simple and descriptive captions. This dataset is well-suited for warming up the Hiera pathway without introducing overly complex semantics. During this stage, we freeze all other model components and only update the Hiera projector using a learning rate of 1e-3. This step allows the model to establish a basic alignment between Hiera-derived features and the language model input, laying the foundation for downstream integration.

The second stage aims to restore the full multimodal capability of the base model, `LLaVA-Video-7B-Qwen2`, which has already been fine-tuned on the full `lmms-lab/LLaVA-OneVision-Data` corpus. Since the introduction of Hiera and its randomly initialized projector may disrupt previously learned representations, we propose to refine this pathway using only a 10% subset of the original dataset. In our implementation, this corresponds to 279,353 training samples. This lightweight stage serves to re-align the newly added features from Hiera with those of SigLIP and the language model, without the need for full-scale retraining. The learning rate is set to 1e-5, and during this stage we update SigLIP, both vision projectors, and the language model, while keeping the Hiera encoder frozen due to memory constraints.

In the third stage, we conduct targeted fine-

	Stage 1	Stage 2	Stage 3	Thinking
Dataset	LLaVA-CC3M-Pretrain-595K	LLaVA-OneVision-Data	ViCA-322K	ViCA-Thinking-2.68K
Training Data Samples	595,375	279,353	322,003	2,680
Trainable Module	Hiera Projection Layer	Full Model (excluding Hiera Module)		
Trainable Parameters	16M	8.04B	8.04B	8.04B
Learning Rate	1e-3	1e-5	1e-5	1e-5
Epochs	1	1	1	1
DeepSpeed Stage	ZeRO-2	ZeRO-3	ZeRO-3	ZeRO-3

Table 1: **Training configuration across all four stages of our hierarchical vision-language model.** Stage 1 pretrains only the Hiera projection layer, while stages 2-4 fine-tune the full model excluding the Hiera module.

tuning on the ViCA-322K dataset, which we curated and publicly released on Hugging Face. This dataset comprises over 300K spatially grounded question-answer pairs derived from indoor videos, covering a wide spectrum of spatial reasoning tasks, ranging from direct spatial knowledge questions—such as room size estimation, relative distances between objects, and temporal ordering of object appearances—to more complex spatial inference and relational understanding. This stage is designed to enhance the model's spatial perception and reasoning abilities, which are critical for indoor visual-language applications. The training configuration remains the same as in the second stage, with a learning rate of 1e-5, updating all major trainable components except the frozen `Hiera` backbone.

To manage computational costs, we use DeepSpeed ZeRO-2 with offloading in the first stage, which involves relatively few trainable parameters. In the second and third stages, we adopt DeepSpeed ZeRO-3 with offloading to support full fine-tuning under our hardware constraints. An overview of the training schedule, data composition, and optimization setup is summarized in Table 1.

4.3 Stage-wise Results and Observations

To better understand the effectiveness of our three-stage training strategy, we analyze both the *training dynamics* and the *output performance* of the model across four checkpoints: (1) the randomly initialized projector checkpoint, and (2–4) the model after each of the three fine-tuning stages.

We first examine the training loss curves. In both the first and third stages, the loss decreases sharply at the beginning of training and stabilizes thereafter—indicating successful learning and convergence. In contrast, the second stage exhibits only a mild reduction in training loss. We attribute this to the fact that the base model,
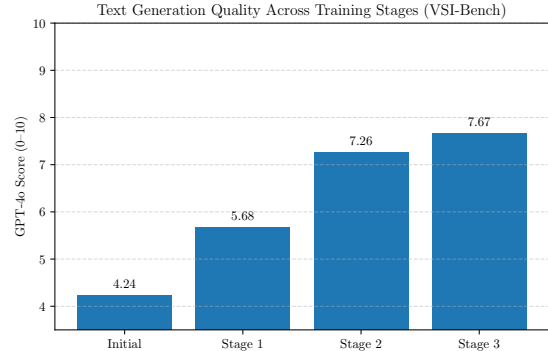

Figure 3: **Average gpt-4.1-mini scores (0–10) for in-detail descriptions generated by each of the four checkpoints on VSI-Bench.**

LLaVA-Video-7B-Qwen2, had already been thoroughly fine-tuned on the same data used in this stage, leaving limited room for further optimization. These trends are visualized in Appendix Figure 5.

To assess the impact of each stage on text generation quality, we conduct a controlled evaluation using the **VSI-Bench** dataset, which consists of 288 indoor video clips. For each video, we prompt the model to generate a *detailed description*. We compare the outputs from the four checkpoints using **gpt-4.1-mini**(OpenAI, 2025) as a judge. Gpt-4.1-mini assigns a quality score in the range $[0, 10]$ for each output based on informativeness, fluency, and spatial relevance. Since some undertrained models may fail to generate the end-of-sequence (EOS) token properly, we set a maximum generation length of 4096 tokens to ensure termination.

As shown in Figure 3, we observe a *progressive and significant improvement* in average generation score across the four checkpoints. Notably, the model fine-tuned with the **ViCA-322K** dataset in the final stage achieves the highest score, confirming that instruction tuning on a spatially grounded, mixed-format dataset not only enhances *spatial reasoning*, but also boosts the overall *language*

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
ViCA2-7B (**ours**)	**56.8(+11.4)**	**65.7(+9.5)**	**51.0(+20.1)**	**75.5(+11.4)**	**71.4(+17.0)**	**51.6(+0.3)**	34.6	**38.1(+2.1)**	**66.5(+17.9)**

Table 2: **Comparison of different models on VSI-Bench.** Our ViCA2-7B 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., +20.1) represent the improvement margins over the best-performing model excluding ViCA2-7B.

generation capability of the model.[1]

5 Experiements Results

5.1 Evaluation Setup

We evaluate our model on the VSI-Bench benchmark, which comprises 5,130 question-answer pairs derived from 288 egocentric indoor videos. The benchmark encompasses eight tasks categorized into configurational, measurement estimation, and spatiotemporal reasoning.

For consistency and reproducibility, we adhere strictly to the evaluation protocols outlined in the VSI-Bench paper. This includes utilizing the provided prompts and employing greedy decoding with parameters set to `do_sample=False` and `temperature=0`. Our evaluation scripts are publicly available to facilitate replication and further research.

5.2 Overall Performance on VSI-Bench

ViCA2-7B achieves state-of-the-art results on VSI-Bench, outperforming both proprietary and open-source models by a large margin(Table 2). It attains an average score of 56.8, exceeding the best proprietary model, Gemini-1.5 Pro (45.4), by 11.4 points, despite having significantly fewer parameters (7B vs. 72B or more).

Performance gains are consistent across most tasks. Notably, ViCA2-7B surpasses all baselines by +20.1 in Absolute Distance and +17.0 in Room Size. The only category where it underperforms is Relative Direction. This is likely due to limited

task-specific supervision: our ViCA-322K dataset focuses on object count, absolute and relative distances, object size, room size, and appearance order, but does not explicitly cover relative direction or route planning. Nevertheless, ViCA2-7B still achieves the highest score on Route Planning (+2.1), suggesting strong cross-task generalization from learned spatial reasoning patterns.

5.3 Impact of Training Data Size

To investigate the effect of data scale on model performance, we conduct a controlled experiment by training ViCA2-7B on progressively larger subsets of the ViCA-322K dataset. As shown in Figure 4, model accuracy consistently improves with increasing data percentage, reaching a peak average score of 56.81 at 100% of the training data. Notably, performance continues to improve even in the final stages (from 95% to 100%), in contrast to the saturation trends observed in our prior work with ViCA, where accuracy plateaued earlier under similar conditions.

We attribute this continued improvement to the increased architectural complexity of ViCA2—particularly the dual-encoder design and token balancing mechanism—which requires greater training data to reach full capacity. While ViCA2-7B remains relatively compact in parameter size, its stronger representational capacity enables it to benefit from additional supervision, especially in spatially grounded reasoning tasks.

These findings suggest that the current training data has not yet saturated the model's learning capacity. ViCA2's architecture retains further performance potential within the 7B-8B model scale,

---

[1] See Appendix Figure 7 for qualitative comparisons of the four checkpoints on the same video sample.
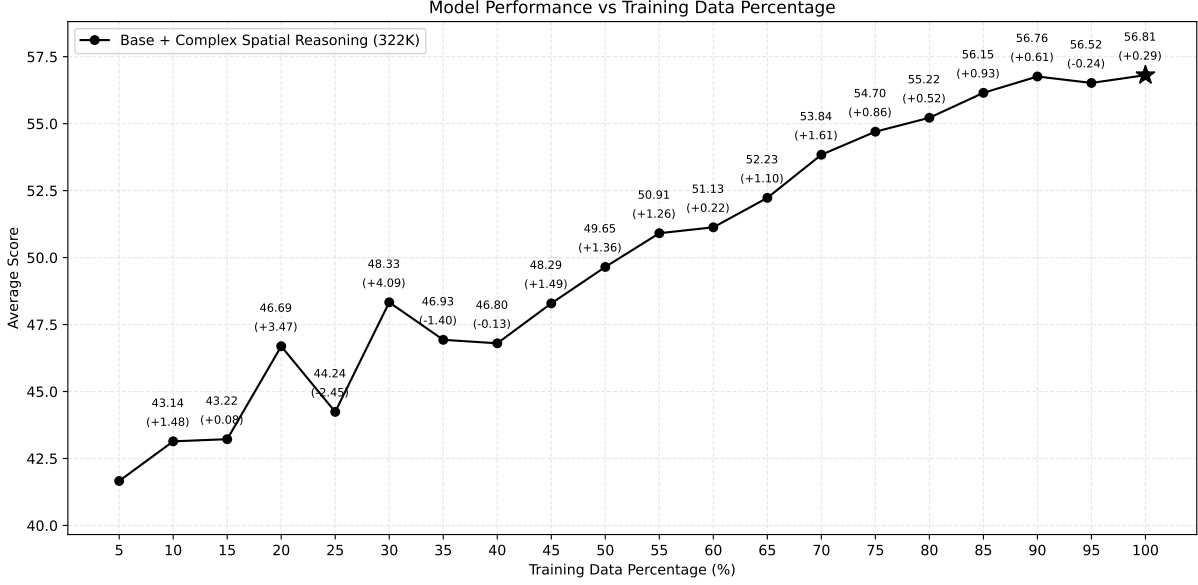
Figure 4: **Impact of training data size on ViCA2-7B performance on VSI-Bench.** The average score consistently improves as the percentage of ViCA-322K training data increases, demonstrating that **ViCA2-7B** benefits from larger-scale spatially grounded instruction tuning and has not yet reached saturation with the current 322K samples.

and could benefit from additional high-quality spatial data. This opens avenues for future scaling of spatial instruction datasets to match increasingly capable vision-language architectures.

6 Conclusion

In this paper, we addressed the pressing challenge of visuospatial cognition in Multimodal Large Language Models (MLLMs), a crucial yet relatively underexplored area for human-like visual intelligence. While existing MLLMs excel at general visual-language tasks, their fine-grained spatial understanding often remains limited. To bridge this gap, we introduced **ViCA2**, a novel framework building upon **LLaVA-OneVision**. **ViCA2** incorporates a dual vision encoder architecture (`SigLIP` for semantics, `Hiera` for spatial structure), augmented by a token ratio control mechanism to balance semantic richness and spatial precision under computational constraints. Furthermore, we developed **ViCA-322K**, a new large-scale dataset with over 322,000 spatially grounded question-answer pairs for targeted instruction tuning of spatial reasoning.

Our comprehensive experiments on the demanding **VSI-Bench** benchmark unequivocally demonstrate **ViCA2**'s superior performance. Despite its compact 7B parameter scale, **ViCA2-7B** significantly outperformed larger open-source counterparts like **LLaVA-NeXT-Video-72B** and surpassed

leading proprietary models such as **Gemini-1.5 Pro** in average scores. These outcomes highlight the efficacy of our architectural innovations and targeted training data in fostering robust spatial reasoning without relying on prohibitively large models. Our ablation studies further suggest that **ViCA2**'s architecture can derive continued benefit from larger spatial instruction datasets, indicating its learning potential is not yet saturated.

These findings underscore the potential of modular architectures and specialized data curation in advancing specific MLLM cognitive abilities, offering a path towards more resource-efficient, capable systems. While **ViCA2** performs strongly on most **VSI-Bench** tasks, areas like 'Relative Direction' indicate avenues for refinement, likely through expanded and diversified training data.

Future work will focus on expanding **ViCA-322K** with more diverse spatial tasks, exploring sophisticated dual-encoder fusion mechanisms, and investigating scalability to larger language models. Applying **ViCA2** to embodied AI tasks like navigation and robotic interaction also presents a promising direction. By open-sourcing **ViCA2**, its model weights, codebase, and the **ViCA-322K** dataset, we aim to catalyze further research in visuospatial cognition, contributing to the broader endeavor of building more spatially intelligent MLLMs.

Limitations

Despite its promising results, ViCA2 has certain limitations. Firstly, the ViCA-322K dataset, while extensive, exhibits some coverage gaps; for instance, the model's relatively lower performance on 'Relative Direction' tasks on VSI-Bench can be attributed to insufficient explicit training examples for this specific spatial concept. Secondly, due to computational resource constraints, the Hiera vision encoder backbone was kept frozen during the primary fine-tuning stages (Stages 2 and 3). Fine-tuning Hiera alongside other components might unlock further performance gains by allowing deeper adaptation of its spatial feature extraction capabilities to the language model's needs. Thirdly, the current iteration of ViCA-322K predominantly features indoor scenes, which may limit the model's out-of-the-box generalization to complex outdoor environments or other specialized visual domains without additional domain-specific fine-tuning. Lastly, our current architecture employs a straightforward concatenation for fusing features from the dual encoders; exploring more sophisticated, learnable fusion mechanisms could offer additional improvements in how semantic and spatial information are integrated.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Google Blog, Accessed: 2025-05-16.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed: 2025-05-16.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-05-17.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pages 29441–29454. PMLR.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

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*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

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*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

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

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. 2023b. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A   Training Record

Figure 5 presents the training logs for all four stages of ViCA2. [2] In both Stage 1 (pretraining the Hiera projection module) and Stage 3 (spatial reasoning fine-tuning), the training loss exhibits a sharp decline during the initial steps, followed by a steady, gradual decrease. This indicates that the model was able to stably learn meaningful patterns from the training data.

In contrast, Stage 2 does not show such a steep drop in loss at the beginning. This is expected, as the base model—initialized from `lmms-lab/LLaVA-Video-7B-Qwen2`—had already undergone extensive fine-tuning on the same dataset (`lmms-lab/LLaVA-OneVision-Data`). Nevertheless, due to the introduction of a randomly initialized Hiera projection module, we conducted partial re-tuning on a 10% subset of the dataset. The goal was to reconcile the representational gap between the Hiera and SigLIP projections and restore the model's overall capability in understanding and describing general videos.

During the final training stage, **ViCA2-Thinking** (see Appendix C), the training loss again decreases sharply at the beginning from its initial high value for this complex task. It eventually plateaus around 0.9. This final loss is notably

higher than the baseline loss of approximately 0.3 observed for the model (as ViCA2-7B) on simpler, preceding tasks before transitioning to explicit multi-step reasoning. This is likely due to the limited size of the **ViCA-thinking-2.68K** dataset. Transitioning the model from answering numerical or multiple-choice questions to generating explicit multi-step reasoning processes presents a substantially higher level of complexity.

We believe that the current dataset, while valuable, is insufficient for fully unlocking the model's potential in spatial reasoning with natural language explanations. In future work, we plan to construct a significantly larger and more diverse dataset specifically designed for spatially grounded reasoning, to further advance the visuospatial cognition abilities of multimodal large language models.

B   Complete Model Architecture

Due to space limitations in the main text, a simplified version of the model architecture was provided(Figrue 2). In this section, we present the full schematic diagram of the complete model architecture(Figrue 6).

---

[2]For a more detailed and interactive view of these logs, please refer to our public Weights & Biases report.
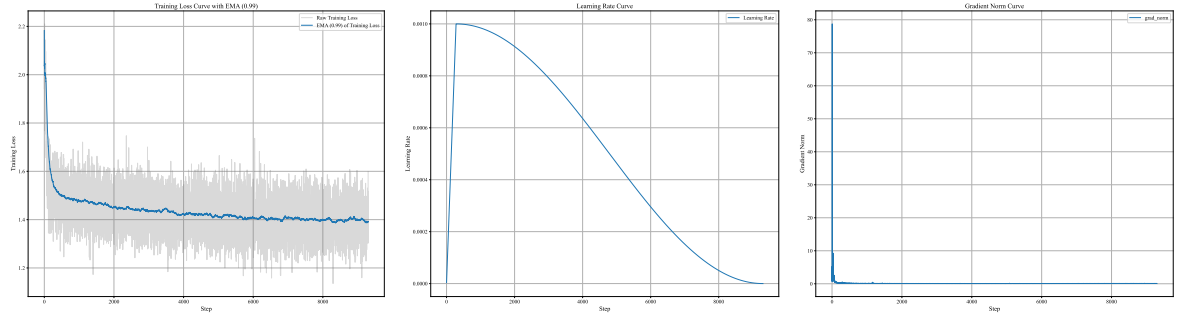
Stage 1: Training Log
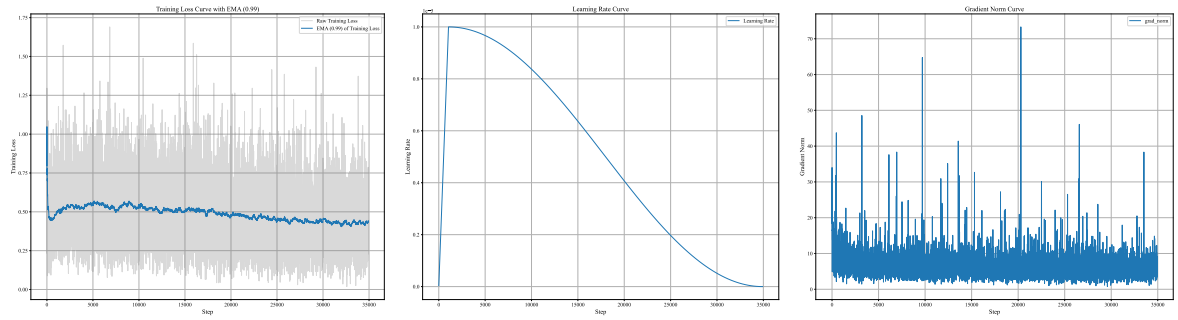
Stage 2: Training Log
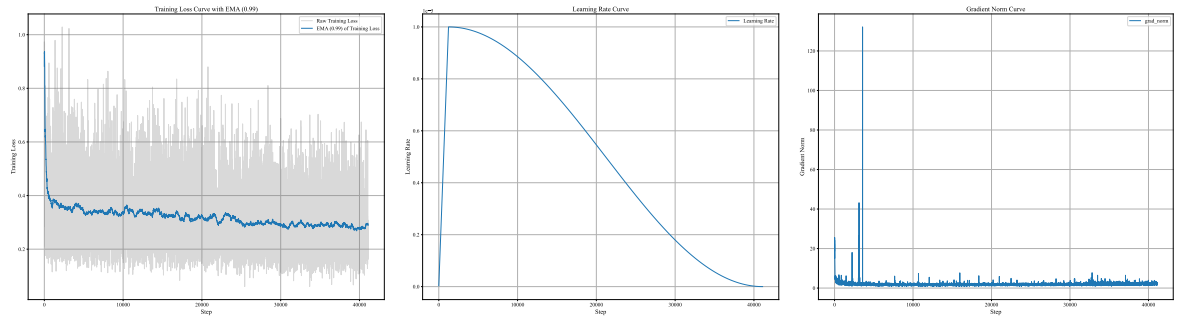
Stage 3: Training Log

ViCA2-Thinking: Training Log

Figure 5: **Training loss curves for the four stages of ViCA2 development.**

Figure 6: **An overview of the ViCA2 architecture.** It integrates SigLIP for global semantics and Hiera for spatial structure, with token ratio control to balance expressiveness and memory efficiency.

Model: ViCA2-init **Video:** arkitscenes/41069043

The video begins with a close-up of a small poster on a beige carpeted floor, featuring a person in a blue outfit with the text 'CHARLIE CHAPLIN MODERN CLASSIC' and 'www.charliechaplin.com'. The camera then pans up to reveal a beige wall with a white electrical outlet and a light switch. The scene transitions to a view of a bed with white bedding and a green thro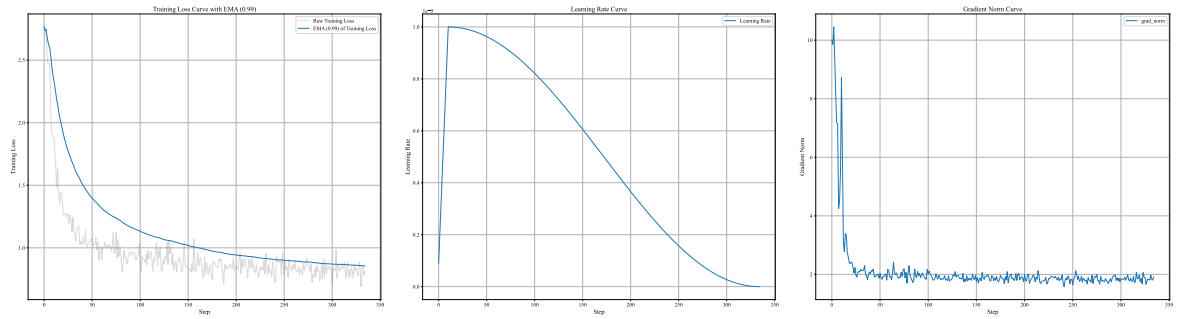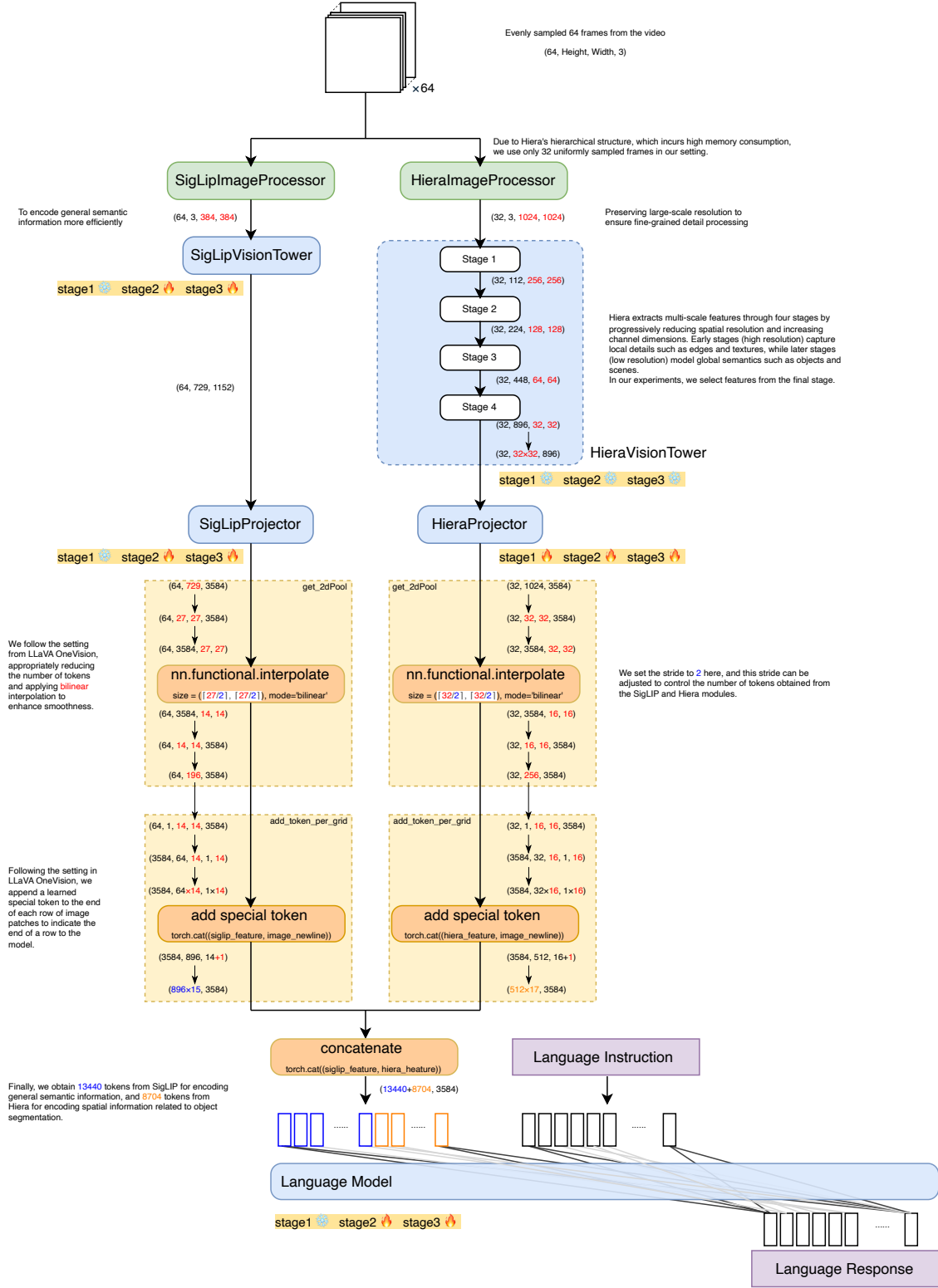w blanket, flanked by two dark brown nightstands, each with a blue lamp. A framed picture hangs above the bed. The camera then pans to the right, showing a window with white curtains and a radiator beneath it. The view shifts to a hallway with a wooden floor and a white wall, leading to a door with a silver handle. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, with a remote control placed on the shelf. The camera then pans to the right, showing a dark brown wardrobe with a silver handle. The view shifts to a window with white curtains and a radiator beneath it. The camera then moves back to the bed, focusing on the nightstand with the blue lamp. The scene transitions to a view of a television mounted on the wall above a white shelf, ...

Model: ViCA2-stage1-align **Video:** arkitscenes/41069043

The video begins with a close-up of a small framed picture on the floor, featuring a person in a blue outfit with the text 'CHARLTON CHAPLIN MODERN CLASSIC' prominently displayed. The camera then pans up to reveal a beige carpeted floor and a white wall with a light switch. The scene transitions to a view of a bed with white bedding and a green throw blanket, flanked by two dark brown nightstands, each with a blue lamp. A framed picture hangs above the bed, and a window with white curtains is visible in the background. The camera then shows a dark brown wardrobe with a flat-screen TV mounted on the wall above it, and a white shelf with a remote control on it. The video continues with a view of the bed and nightstands, then pans to show the window and the dark brown wardrobe. The camera then moves to the door, which is slightly ajar, revealing a hallway with wooden flooring and a light switch on the wall. The video concludes with a view of the hallway, showing a framed picture on the floor near the door, and then pans to show the bed and nightstands again.

Model: ViCA2-stage2-onevision-ft **Video:** arkitscenes/41069043

The video showcases a well-organized bedroom with a modern and minimalist design. The room features a large bed with a dark headboard, covered in white linens and a green blanket. The bed is positioned against a wall with a window that has white curtains, allowing natural light to enter the room. Adjacent to the bed, there is a small, dark-colored dresser with a flat-screen TV mounted on the wall above it. A blue lamp with a cylindrical shade sits on top of the dresser, providing additional lighting.

The floor is carpeted in a light beige color, and the walls are painted in a light yellow hue, creating a warm and inviting atmosphere. A door is visible in the background, leading to another room or hallway. The room also includes a small table with a remote control on it, suggesting that the space is used for relaxation and entertainment.

The video captures the room from various angles, showcasing the layout and design elements in detail. The overall impression is of a clean, modern, and comfortable living space.

Model: ViCA2 **Video:** arkitscenes/41069043

I am currently in a cozy bedroom space. To my left, there is a neatly made bed with a dark headboard, and a soft, light green throw blanket draped over it. The bed is positioned against the wall, and there are several pillows arranged at the head. Directly in front of me, there is a small bedside table with a blue lamp that casts a warm glow, and a power outlet is located nearby. To my right, I can see a tall wardrobe with a mirrored door, which reflects the light from the window. The window is adorned with light, sheer curtains that allow soft natural light to filter in, creating a bright and airy atmosphere. Below the window, there is a radiator, providing warmth to the room. The walls are painted in a soft beige, contributing to the overall calm ambiance. Behind me, there is a door leading to another room, and the floor is covered in a light carpet that adds to the comfort of the space.

Figure 7: Qualitative comparison of the generated texts from four different checkpoints on a shared video sample (ARKitScenes/41069043), highlighting the evolution of generation quality during training.

C ViCA2-Thinking Fine-Tuning and Observations

Following the setup described in our previous ViCA paper, we fine-tuned the ViCA2-7B model using our curated **ViCA-Thinking-2.68K** dataset. The fine-tuning was conducted on four H200 GPUs (141GB each), and completed within 1 hour and 59 minutes. The resulting model, which we denote as **ViCA2-7B-Thinking**, is capable of producing an intermediate *Thoughts* segment prior to generating the final response. This reasoning step aims to enhance the model's spatial cognition capabilities. We provide representative output examples from the model in Figure 8, 9, 10.

In terms of visuospatial performance, ViCA2-7B-Thinking demonstrates a notable drop compared to its predecessor,**ViCA2-7B**, mirroring the degradation previously observed in **ViCA-7B-Thinking** (see Table 3). Its performance converges to the level of baseline 7B/8B open-source models that have not undergone any spatial reasoning-specific fine-tuning.

We hypothesize that this degradation is linked to the model's ability to perform high-quality text generation. The *Thoughts* component demands significantly more from the model's generative capacity compared to structured QA formats. As we showed in Section 4.3, the model's output fluency and informativeness improve progressively with additional training. However, upon examining the outputs from the checkpoint at the end of Stage 3, we observed that almost all video descriptions begin with highly templated and egocentric phrases such as:

> *"I find myself in a compact kitchen space..."*
>
> *"I am seated on a comfortable gray couch..."*
>
> *"I am currently in a small, functional restroom space..."*
>
> *"In this kitchen space, I can see a variety of essential appliances..."*

This observation suggests that the model, during Stage 3 training, acquired a strong tendency to generate egocentric spatial descriptions—likely learned from the spatial caption subset of ViCA-322K. While such descriptions capture local spatial relationships from a first-person perspective, they do not generalize to broader forms of spatial reasoning or flexible narrative generation. Furthermore, the volume of this spatial caption data within ViCA-322K is relatively limited compared to other data types, which restricts the diversity of egocentric descriptive patterns the model can internalize.

We identify three levels of capability relevant to this discussion:

1. Describing object-level spatial relations from an egocentric (first-person) point of view.

2. Providing general descriptions of the overall spatial layout or environment.

3. Performing free-form, coherent, and abstract natural language generation.

These abilities increase in difficulty from (1) to (3). Our findings suggest that ViCA2-7B, after Stage 3, primarily exhibits improvements in the first capability. This may partially explain why the overall performance of ViCA2-7B-Thinking degrades after fine-tuning with ViCA-Thinking-2.68K, despite its high relevance to spatial reasoning.

We further argue that achieving a balanced improvement in both spatial cognition and general language generation requires a revised training strategy. Due to computational constraints, we initialized our model from `lmms-lab/LLaVA-Video-7B-Qwen2` and fine-tuned it sequentially on a subset of `LLaVA-OneVision-Data` and then on ViCA-322K. This strategy, while effective for spatial QA tasks with structured outputs (e.g., numeric or multiple-choice answers), may have inadvertently weakened the model's open-ended generation capabilities—an ability that is essential for high-quality reasoning in ViCA2-7B-Thinking.

Nevertheless, our overall training approach did yield strong gains on VSI-Bench, confirming the effectiveness of both our data curation and architectural innovations. We hypothesize that with access to greater compute, a joint training regime combining `LLaVA-OneVision-Data` and ViCA-322K from the outset—rather than in staged phases—could simultaneously preserve text generation quality and enhance spatial reasoning. Once such a robust base model is obtained, fine-tuning with ViCA-Thinking-2.68K could lead to even greater gains, leveraging the explicit reasoning format introduced by the *Thought* prompts.

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-7B/8B									
InternVL2-8B	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
VILA-1.5-8B	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
LLaVA-NeXT-Video-7B	35.6	48.5	14.0	47.8	24.2	43.5	**42.4**	34.0	30.6
LLaVA-OneVision-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
ViCA2-7B-thinking									
ViCA2-7B-thinking (**ours**)	35.5	33.9	25.0	42.0	33.1	40.8	34.9	33.1	41.6
Open-source Models-40B									
InternVL2-40B	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
VILA-1.5-40B	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
Open-source Models-72B									
LLaVA-NeXT-Video-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
ViCA2-7B									
ViCA2-7B (**ours**)	**56.8**(+11.4)	**65.7**(+9.5)	**51.0**(+20.1)	**75.5**(+11.4)	**71.4**(+17.0)	**51.6**(+0.3)	34.6	**38.1**(+2.1)	**66.5**(+17.9)

Table 3: **Quantitative comparison of ViCA2-7B-Thinking, ViCA2-7B, and selected baseline models on the VSI-Bench benchmark.** While ViCA2-7B achieves the highest average score, ViCA2-7B-Thinking trades off numerical accuracy for improved interpretability through intermediate reasoning steps.

We leave the empirical validation of this hypothesis for future work.

D   Comparison with ViCA-7B

While our proposed model, **ViCA2-7B**, achieves state-of-the-art performance in visuospatial reasoning compared to existing multimodal large language models (see Table 2), its overall performance is slightly lower than that of our earlier model, **ViCA-7B**, trained in previous work (see Table 4).

We hypothesize that this discrepancy arises from the increased complexity and capacity of ViCA2-7B. Compared to ViCA-7B, ViCA2-7B incorporates a dual vision encoder architecture and significantly more trainable parameters. Consequently, it likely demands a larger and more diverse dataset to reach its full potential. However, the **ViCA-322K** dataset—while already extensive—was originally designed and curated for ViCA-7B. During the construction of this dataset, we conducted performance scaling analysis on ViCA-7B and observed that model performance plateaued and slightly declined beyond a certain data scale, prompting us to halt further data collection at that point (Figure 11).

To better understand how dataset size influences each model, we further analyzed the performance scaling curves of **ViCA-7B** and **ViCA2-7B** across all eight tasks in VSI-Bench. The results are presented in Figure 12, 13, 14, 15, 16, 17, 18, 19.

We observe that for tasks well represented in the ViCA-322K dataset—such as *Object Count*, *Absolute Distance*, and *Room Size*—both models exhibit

	ViCA-7B	ViCA2-7B
Numerical Answer		
Obj. Count	**68.8**	67.7
Abs. Dist.	**57.0**	51.0
Obj. Size	**79.2**	75.5
Room Size	**75.1**	71.4
Multiple-Choice Answer		
Rel. Dist.	**58.5**	51.6
Rel. Dir.	**42.6**	34.6
Route Plan	34.5	**38.1**
Appr. Order	**68.8**	66.5
Average	**60.6**	56.8

Table 4: Comparison of ViCA-7B and ViCA2-7B Performance on VSI-Bench

clear performance gains as training data increases. In contrast, for tasks underrepresented or entirely absent from the dataset, namely *Relative Direction* and *Route Planning*, both models show unstable performance trends, with scores fluctuating without a clear upward trajectory. This suggests that the existing dataset lacks sufficient supervision for these tasks and that performance gains in these categories are unlikely without task-specific data augmentation.

These findings further support our earlier hypothesis: ViCA2-7B, with its greater architectural capacity, remains under-optimized under the current data regime and may benefit from scaling the dataset beyond the limits originally established for ViCA-7B.

A representative output produced by ViCA2-7B-Thinking

think

1. **Understand the Goal:** The question asks for 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, armchairs, office chairs, etc.
3. **Analyze the Video Chronologically:**
   * **0:00 - 0:05:** The video starts showing a poster on the floor, then pans up and moves into a room. A grey, upholstered sofa is visible. This is not a chair.
   * **0:05 - 0:10:** The camera pans right, revealing a dining area. A round glass dining table is visible. Around this table, there are four chairs with grey upholstery and chrome legs. These are clearly chairs. **Count: 4**.
   * **0:10 - 0:15:** The camera continues panning right, showing a radiator and a wall with a colorful artwork. No chairs here.
   * **0:15 - 0:20:** The camera pans further right, showing a large window with a view of the outside. Under the window, there is a radiator. No chairs here.
   * **0:20 - 0:25:** The camera pans right again, showing the sofa again. No chairs here.
   * **0:25 - 0:30:** The camera pans right, showing a TV on a stand. No chairs here.
   * **0:30 - 0:35:** The camera pans right, showing a white kitchen area with cabinets, sink, oven, and stovetop. No chairs here.
   * **0:35 - 0:40:** The camera pans right, showing a microwave and a toaster oven. No chairs here.
   * **0:40 - 0:45:** The camera pans right, showing a doorway. No chairs here.
   * **0:45 - 0:50:** The camera pans right, showing the floor near the doorway. A blue item (possibly a poster or paper) is lying on the floor. No chairs here.
   * **0:50 - 1:00:** The camera pans back across the room, showing the sofa, dining table, chairs, and the initial view. No additional chairs are revealed.
4. **Consolidate and Count:** Throughout the video, only one set of chairs (the four around the dining table) is clearly visible. No other chairs are shown.
5. **Final Conclusion:** There are 4 chairs in the room.

response

The video pans around a room, showing various areas including a dining area and a kitchen area.
- At 0:05-0:10, a round glass dining table is shown. Around this table, four chairs with grey upholstery and chrome legs are visible.
- No other chairs are shown in the rest of the room tour, which includes the sofa, radiator, window, kitchen appliances, and doorway.
Therefore, there are 4 chairs in the room.

final

\[
\boxed{4}
\]

✅

A representative output produced by ViCA-7B-Thinking

think

1. **Understand the Goal:** The question asks for the total number of chairs visible in the room shown in the video.
2. **Define "Chair":** I need to identify objects designed primarily 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 starts by showing a poster on the floor. No chairs are visible yet.
   * **0:05 - 0:10:** The camera pans up and moves into a living/dining area. A dark grey, upholstered **dining chair** is clearly visible next to a glass dining table. Let's call this **Chair 1**.
   * **0:10 - 0:14:** The camera continues panning around the dining area. Another identical dark grey dining chair is visible next to Chair 1. Let's call this **Chair 2**.
   * **0:14 - 0:18:** The camera pans further right, past a radiator and a colorful artwork. No new chairs are seen here.
   * **0:18 - 0:27:** The camera pans towards the windows and balcony door. No chairs are visible in this section.
   * **0:27 - 0:36:** The camera pans back across the sofa and coffee table. No additional chairs are seen here.
   * **0:36 - 0:45:** The camera pans towards a TV stand and then moves into the kitchen area. No chairs are visible in the kitchen.
   * **0:45 - 1:06:** The camera pans back across the living area, showing the sofa, coffee table, TV stand, and kitchen again. No new chairs are identified.
   * **1:06 - 1:16:** The camera pans back towards the entrance area, showing the dining chairs (Chair 1 and Chair 2) again before exiting the room. No other chairs are seen.
4. **Consolidate Findings:** Throughout the video tour of the room, only two distinct chairs were observed: the two matching dining chairs near the glass table.
5. **Final Count:** There are 2 chairs visible in the room.

response

The video shows a tour of a room that appears to be a combined living and dining area. By carefully observing the furniture as the camera pans around the room:
* At timestamp 0:09, a dark grey dining chair is visible next to a glass table.
* At timestamp 0:12, a second identical dark grey dining chair is visible next to the first one.
* The rest of the video shows other furniture like sofas, a coffee table, a TV stand, and kitchen appliances, but no additional chairs are shown.
Therefore, there are 2 chairs in the room.

final

\[
\boxed{2}
\]

❌

Figure 8: Comparison of outputs from **ViCA2-7B-Thinking** and **ViCA-7B-Thinking**.

A representative output produced by **ViCA2**-7B-Thinking

3134
think

1. **Understand the Goal:** The question asks to identify which of the four listed objects (tv, plant, chair, pillow) is closest to the door, measuring from the closest point of each object to the closest point of the door.

2. **Identify the Objects and the Door:**
   * **Door:** The video shows a wooden door with a glass panel, located on the wall opposite the window. It's visible around timestamps 0:15-0:17 and 0:27-0:28.
   * **TV:** A large flat-screen TV is mounted on the wall, above a black media console. It's visible from the start (0:00) and again later (e.g., 0:10, 0:23).
   * **Plant:** A potted plant is sitting on a wooden chest of drawers next to the TV (visible 0:00, 0:10, 0:23).
   * **Chair:** A black leather armchair is positioned near the window, next to a coffee table (visible 0:00, 0:10, 0:23).
   * **Pillow:** A green pillow with white circular patterns is resting on the arm of the sofa (visible 0:00, 0:10, 0:23).

3. **Analyze Spatial Relationships:**
   * The room appears to be a living room or lounge area.
   * The TV, plant, and chair are grouped together on one side of the room, near the window.
   * The sofa is on the opposite side of the room, facing the TV area.
   * The door is located on the wall opposite the window, near the entrance to the room.

4. **Estimate Distances from the Door:**
   * **TV:** The TV is on the wall opposite the window. The distance from the TV to the door involves crossing the width of the room. This appears to be the longest distance among the options.
   * **Plant:** The plant is on a chest of drawers. This chest is positioned between the TV and the sofa. The distance from the plant to the door is likely greater than the distance from the sofa to the door.
   * **Chair:** The chair is near the window. The distance from the chair to the door is likely greater than the distance from the sofa to the door.
   * **Pillow:** The pillow is on the sofa. The sofa is positioned across the room from the door. The distance from the pillow to the door is likely greater than the distance from the sofa to the door.

5. **Compare Distances:**
   * The TV is across the room from the door.
   * The plant, chair, and pillow are all located on the same side of the room, near the window and sofa.
   * Therefore, the distance from the door to the pillow (which is on the sofa) is shorter than the distance from the door to the TV, plant, or chair.

6. **Conclusion:** Based on the visual layout, the pillow is the closest object to the door among the given options.

response

A representative output produced by **ViCA**-7B-Thinking

think

1. **Understand the Goal:** The question asks to identify which of the four listed objects (tv, plant, chair, pillow) is closest to the door shown in the video, measuring from the closest points.

2. **Identify the Door:** The video pans around a living room. A white interior door is clearly visible on the right side of the frame multiple times, for example, at 0:35-0:36, 0:47-0:48, 1:09-1:10, and 1:23-1:24. It appears to lead to another room or hallway.

3. **Locate the Objects Relative to the Door:**
   * **TV:** The TV is located on the left side of the room, near a window (visible at 0:00-0:02, 0:10-0:11). It is far from the door on the right.
   * **Plant:** There are several plants. One is on a trunk near the window (0:00-0:02), another is on a stand near the window (0:00-0:02), and another is on a trunk near the window (0:00-0:02). All these are on the left side of the room, near the window, far from the door on the right.
   * **Chair:** There are two types of chairs:
     * A large, dark grey armchair/loveseat is positioned near the center of the room, facing the TV (visible at 0:35-0:36, 0:47-0:48, 1:09-1:10, 1:23-1:24). This chair is relatively close to the door on the right.
     * A black leather recliner is positioned near the window, next to a keyboard piano (visible at 0:00-0:02, 0:05-0:06, 0:10-0:11, 0:28-0:30, 0:35-0:36, 0:47-0:48, 1:09-1:10, 1:23-1:24). This chair is also relatively close to the door on the right.
   * **Pillow:** Pillows are visible on the sofa (e.g., 0:35-0:36, 0:47-0:48, 1:09-1:10, 1:23-1:24). The sofa is positioned along the wall opposite the TV, near the door. Pillows are therefore located near the door.

4. **Compare Distances to the Door:**
   * The TV, plant, and chairs (both types) are all located on the left side of the room, near the window. The distance from these items to the door on the right side is significant, likely several meters.
   * The pillows are located on the sofa, which is positioned along the wall opposite the TV. The distance from the pillows on the sofa to the door on the right side is also significant, likely several meters.
   * The armchair and recliner are positioned closer to the center of the room, nearer to the door than the furniture on the left side. The distance from these chairs to the door appears to be less than the distance from the TV, plant, or sofa to the door.

5. **Conclusion:** Comparing the locations, the chairs (especially the armchair and recliner) are significantly closer to the door than the TV, plant, or pillows.

response

Based on the video analysis:
1. The **door** is located on the right side of the room (visible at timestamps like 0:35-0:36, 1:09-1:10).
2. The **TV** is located on the left side of the room,

1. The video shows a room containing a TV, a plant, a chair, and a pillow.
2. The **door** is located on the wall opposite the window (visible at 0:15-0:17).
3. The **TV** is mounted on the wall opposite the window, near the window (visible at 0:00, 0:10, 0:23).
4. The **plant** is on a chest of drawers next to the TV (visible at 0:00, 0:10, 0:23).
5. The **chair** is near the window (visible at 0:00, 0:10, 0:23).
6. The **pillow** is on the sofa (visible at 0:00, 0:10, 0:23).
7. The room appears to be roughly rectangular. The TV, plant, and chair are grouped together on one side near the window. The sofa is on the opposite side, facing the TV area.
8. The distance from the door to the TV is across the width of the room.
9. The distance from the door to the plant, chair, or pillow is the distance from the sofa to the door, which is across the width of the room.
10. Comparing the distances, the pillow is on the sofa, which is across the room from the door. The TV, plant, and chair are also across the room from the door, but they are located on the same side of the room as the sofa.
11. Therefore, the pillow is the closest object to the door among the given options.

final

\[
\boxed{D. pillow}
\]

---

 near the window (visible at 0:00-0:02).
3. Several **plants** are located on trunks near the window (visible at 0:00-0:02).
4. Two types of **chairs** (a large armchair/loveseat and a recliner) are located near the center of the room, closer to the door than the TV or plants (visible at 0:35-0:36, 0:47-0:48, 1:09-1:10, 1:23-1:24).
5. **Pillows** are located on the sofa, which is positioned along the wall opposite the TV (visible at 0:35-0:36, 1:09-1:10).

Comparing the distances from the closest point of each object to the door, the chairs (especially the armchair and recliner) are significantly closer than the TV, plant, or pillows.

final

The final answer is $\boxed{C. chair}$

Figure 9: Comparison of outputs from **ViCA2-7B-Thinking** and **ViCA-7B-Thinking**.

A representative output produced by **ViCA2**-7B-Thinking	A representative output produced by **ViCA**-7B-Thinking
think	think

<div style="display:flex">
<div>

ViCA2-7B-Thinking (left column):

think

1. **Understand the Goal:** The user wants to know the sequence of the three actions (turn back, turn left, turn right) performed by a robot starting from the closet, moving towards the window, and finally reaching the black nightstand. I need to watch the video, identify the sequence of these three actions, and select the corresponding option.

2. **Analyze the Video for Key Points and Path:**
   * The video starts by showing a poster on a bed (0:00-0:05).
   * The camera pans right, revealing a window with black curtains (0:05-0:10).
   * The camera pans further right, showing a bed with a purple duvet (0:10-0:15).
   * The camera pans further right, showing a closet area with a clothes drying rack and a mirrored wardrobe (0:15-0:20).
   * The camera pans further right, showing a black nightstand with a box of tissues and a tube of toothpaste (0:20-0:25).
   * The camera pans further right, showing a door leading out of the room (0:25-0:30).
   * The camera pans further right, showing a glimpse of a water cooler and a water bottle on the floor (0:30-0:35).
   * The camera pans back towards the closet area (0:35-0:40).
   * The camera pans back towards the bed (0:40-0:45).
   * The camera pans back towards the window (0:45-0:50).
   * The camera pans back towards the poster on the bed (0:50-0:55).

3. **Identify the Sequence of the Three Actions:**
   * **Identify the Starting Point:** The sequence begins at the start of the video, where the camera is focused on the poster on the bed.
   * **Identify the "Bed":** The bed is the first object clearly identifiable. It's the bed with the purple duvet.
   * **Identify the "Closet Area":** The closet area is the space containing the clothes drying rack and the mirrored wardrobe. This area is visible from around 0:15 onwards.
   * **Identify the "Window":** The window is the large window with black curtains, visible from around 0:05 onwards.
   * **Identify the "Nightstand":** The black nightstand is the one with the box of tissues and toothpaste, visible from around 0:20 onwards.
   * **Identify the "Door":** The door is the white door leading out of the room, visible from around 0:25 onwards.
   * **Identify the "Water Cooler":** The water cooler is the appliance seen briefly around 0:30-0:35.
   * **Identify the "Water Bottle":** The water bottle is the blue bottle seen briefly around 0:30-0:35.

4. **Analyze the Video for the Sequence:**
   * **Initial Focus:** The video starts focused on the poster on the bed (0:00-0:05). This is the starting point.
   * **First Turn:** The camera pans right from the bed towards the window (0:05-0:10). This is the first turn. The

</div>
<div>

ViCA-7B-Thinking (right column):

think

1. **Understand the Goal:** The user wants to know the sequence of turns (turn back, turn left, or turn right) a robot would take starting from the closet area, facing the heater, and moving towards the black nightstand next to the bed.
2. **Analyze the Video for Path and Turn Points:**
   * The video starts by showing a bed with a nightstand on one side (0:00-0:05).
   * The camera pans around the room, showing the bed, window, radiator, and then moves towards the closet area (0:13-0:24).
   * The camera enters the closet space (0:25-0:37).
   * The camera pans back out of the closet towards the bed and nightstand (0:38-0:46).
   * The camera pans across the bed towards the window (0:47-0:53).
   * The camera pans across the window towards the wall where the nightstand is located (0:54-1:01).
   * The camera focuses on the nightstand (1:02-1:05).
3. **Identify Key Locations and Relative Positions:**
   * **Nightstand:** Located on the side of the bed closest to the window.
   * **Heater:** Located under the window.
   * **Closet Entrance:** The point where the camera enters the closet space.
   * **Bed:** The main object in the room.
   * **Window:** The window is on the wall opposite the closet entrance.
4. **Determine the Sequence of Movement:**
   * The robot starts near the closet entrance, facing the heater.
   * It needs to move past the foot of the bed towards the window.
   * It needs to turn to reach the nightstand, which is on the side of the bed closest to the window.
5. **Break Down the Path into Segments:**
   * **Segment 1:** From the closet entrance, moving forward along the length of the bed towards the window. This involves turning left (relative to the initial view) to align with the bed's length.
   * **Segment 2:** Moving along the length of the bed towards the window. This involves going past the foot of the bed.
   * **Segment 3:** Turning right (relative to the initial view) to reach the nightstand, which is on the side of the bed closest to the window.
6. **Sequence the Turns:**
   * The first turn is to the left (Segment 1).
   * The second turn is to the right (Segment 3).
   * The third turn is to the left (Segment 2).
7. **Compare with Options:**
   * A. Turn Right, Turn Left, Turn Left - Incorrect order.
   * B. Turn Right, Turn Right, Turn Left - Incorrect order.
   * C. Turn Right, Turn Right, Turn Right - Incorrect order.
   * D. Turn Back, Turn Right, Turn Left - This matches the sequence derived: Turn (Back/Left), Turn (Right), Turn (Left).

response

The video shows a robot navigating from the closet area

</div>
</div>

bed is the reference point. The camera moves away from the bed towards the window.
   * **Second Turn:** The camera continues panning right, past the window, towards the closet area (0:15-0:20). This is the second turn. The camera moves away from the window towards the closet.
   * **Third Turn:** The camera continues panning right, past the closet area, towards the nightstand (0:20-0:25). This is the third turn. The camera moves away from the closet towards the nightstand.
   * **Final Position:** The camera ends back at the poster on the bed (0:50-0:55).

5. **Determine the Sequence:** The sequence of the three turns is: Bed -> Window -> Nightstand.

6. **Compare with Options:**
   * A. Turn Right, Turn Left, Turn Left
   * B. Turn Right, Turn Right, Turn Left
   * C. Turn Right, Turn Right, Turn Right
   * D. Turn Back, Turn Right, Turn Left

7. **Evaluate Options:** The observed sequence is Bed -> Window -> Nightstand. None of the options match this sequence. The options are:
   * Option A: Right, Left, Left
   * Option B: Right, Right, Left
   * Option C: Right, Right, Right
   * Option D: Back, Right, Left

8. **Re-evaluate and Select Best Fit:** Since none of the options match the observed sequence (Bed -> Window -> Nightstand), there might be an error in the question or the options provided. However, if forced to choose the "best" fit, Option A (Right, Left, Left) is the only one that correctly identifies the first two items in the sequence (Right, Left) and places the third item (Left) last. This is the closest representation of the sequence, despite the incorrect ordering of the last two items.

response

1. The video begins by showing a poster on a bed (0:00-0:05).
2. The camera pans right, moving away from the bed towards the window (0:05-0:10).
3. The camera continues panning right, moving past the window towards the closet area (0:15-0:20).
4. The camera continues panning right, moving past the closet area towards the black nightstand (0:20-0:25).
5. The camera ends back at the poster on the bed (0:50-0:55).

The observed sequence of the three actions is: Bed -> Window -> Nightstand.

Comparing this sequence to the options:
* A. Right, Left, Left
* B. Right, Right, Left
* C. Right, Right, Right
* D. Back, Right, Left

None of the options match the observed sequence. Option A is the only one that correctly identifies the first two

towards the black nightstand next to the bed. To do this, the robot first needs to turn left to align with the bed's length, then move along the bed's length towards the window, and finally turn right to reach the nightstand, which is on the side of the bed closest to the window. Therefore, the sequence of turns is: turn left, turn right, turn left.

final

\[
\boxed{D. Turn Back, Turn Right, Turn Left}
\]
❌

items in the sequence (Right, Left) and places the third item (Left) last. Although the order of the last two items is incorrect, it is the closest option to the observed sequence.

final

\[
\boxed{A. \text{Right, Left, Left}}
\]
✅

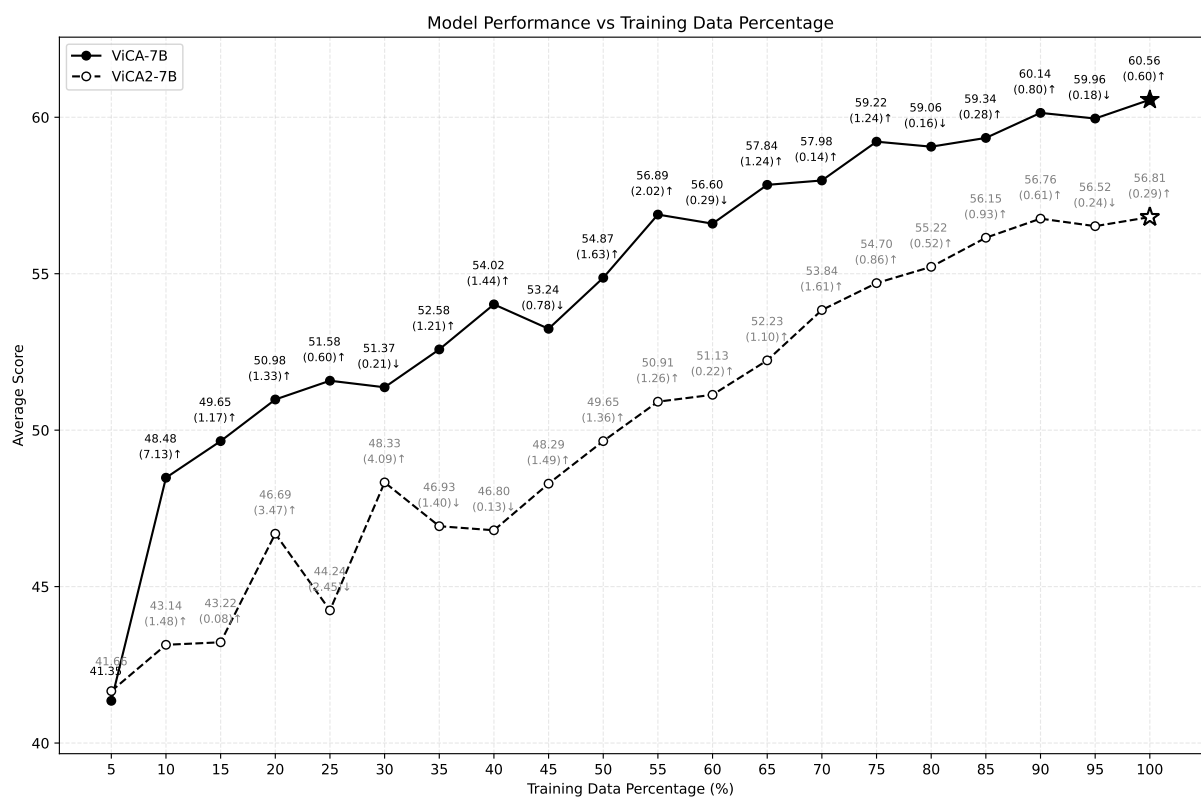Figure 10: Comparison of outputs from **ViCA2-7B-Thinking** and **ViCA-7B-Thinking**.

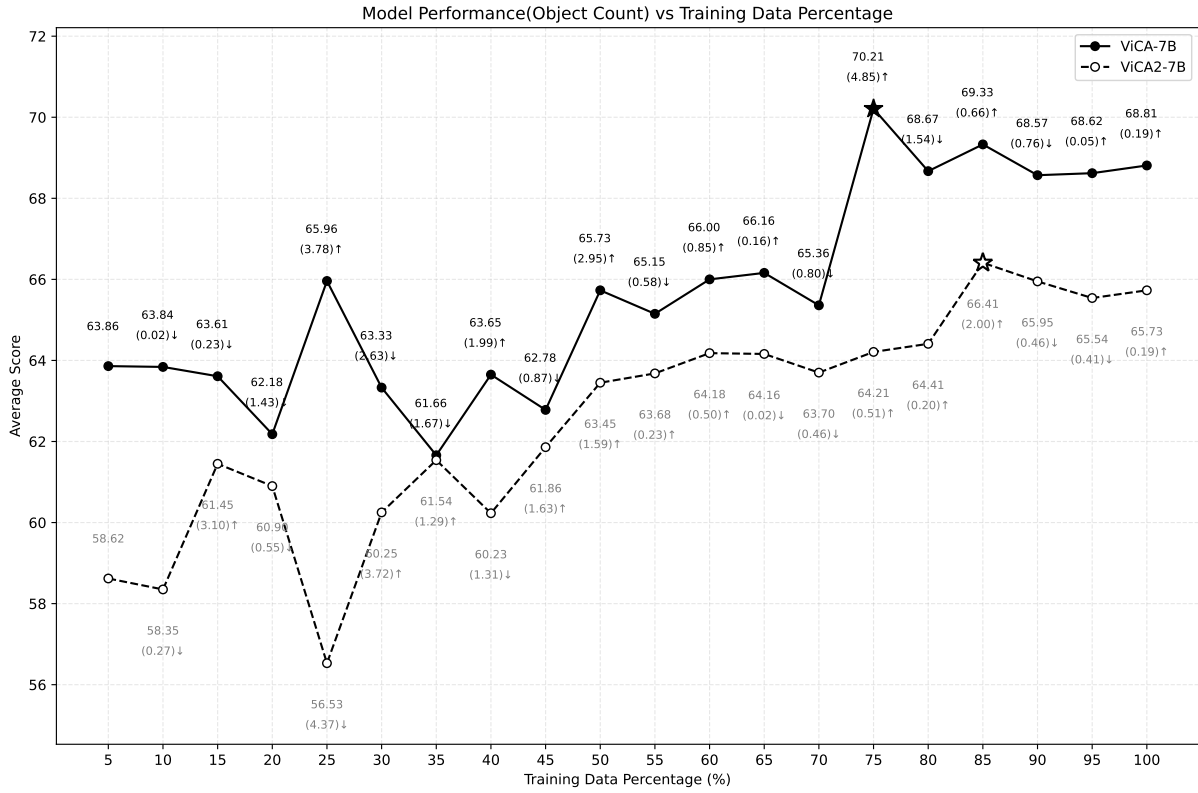Figure 11: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size

Figure 12: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Object Count**.
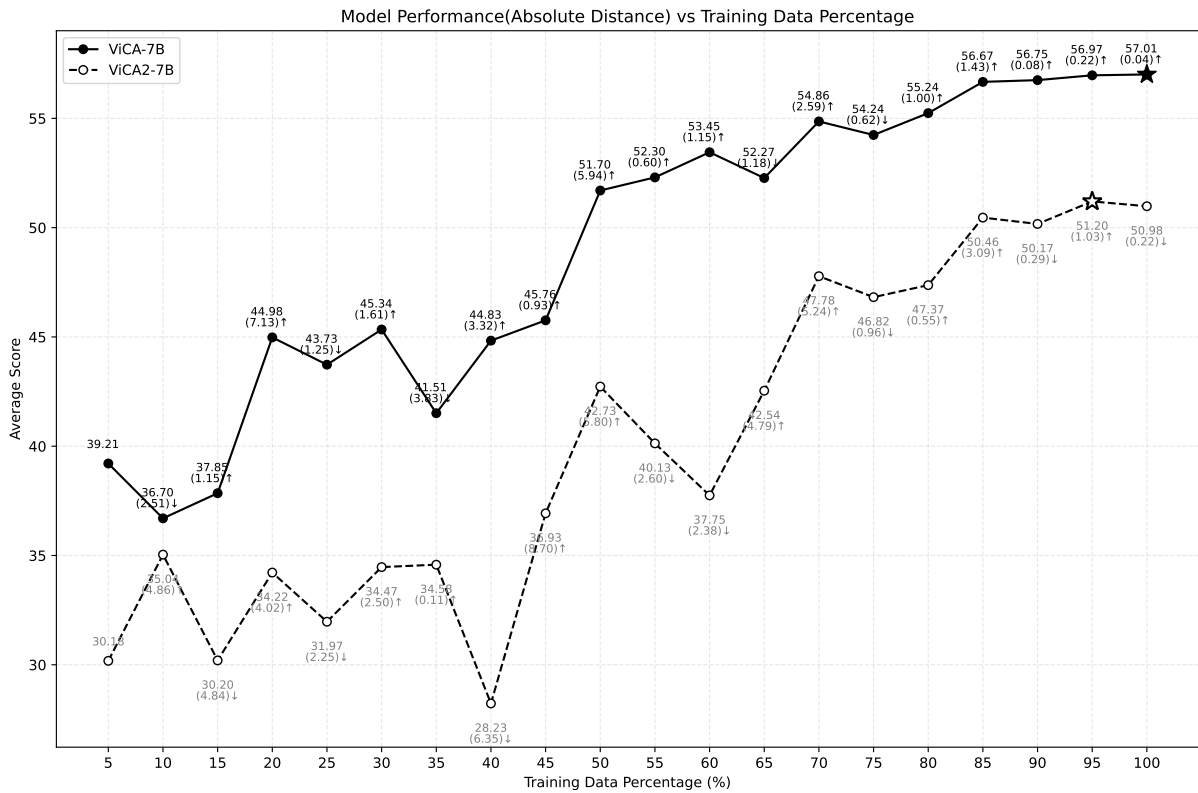

Figure 13: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Absoulte Distance**.
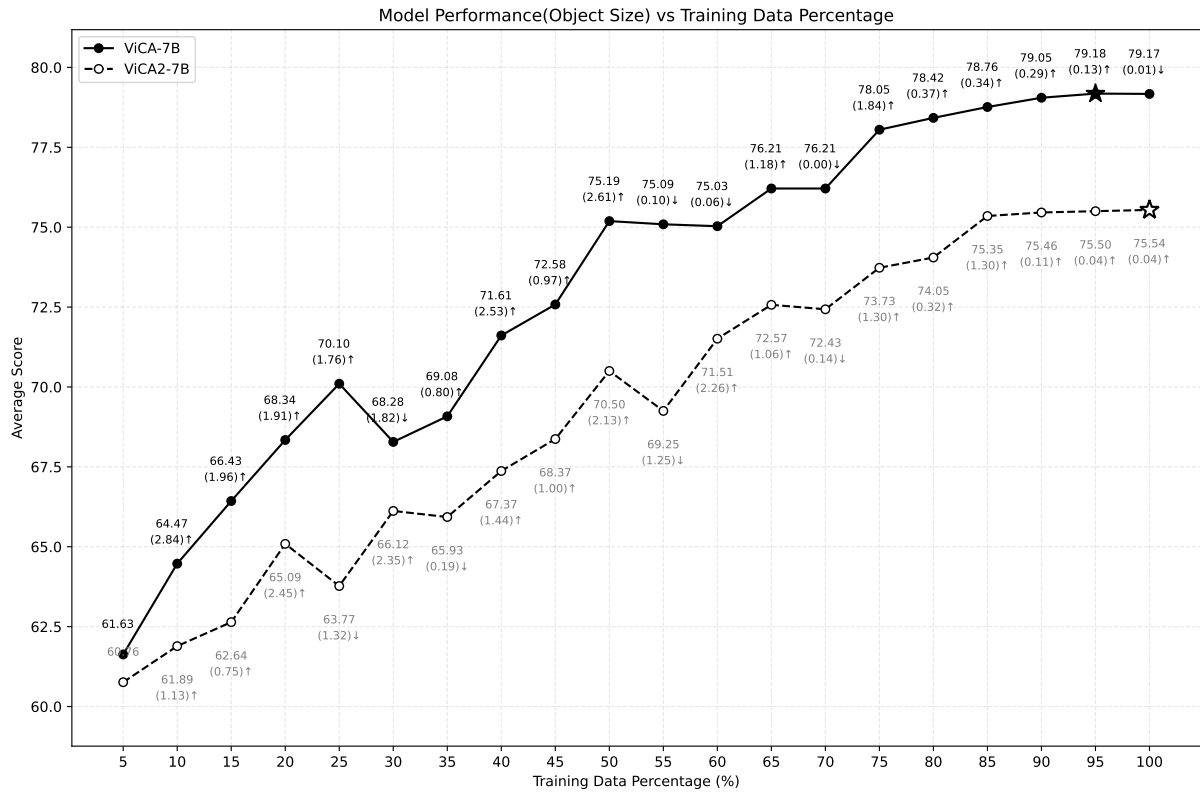
Figure 14: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Object Size**.
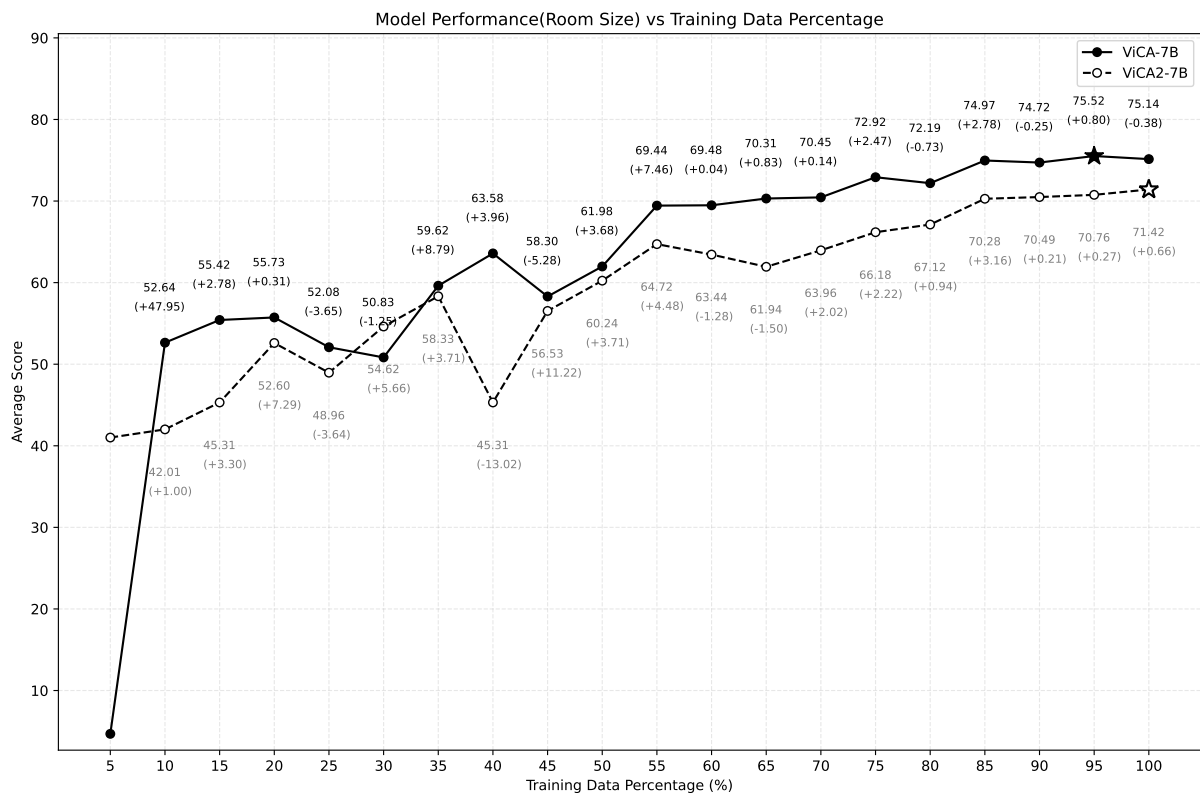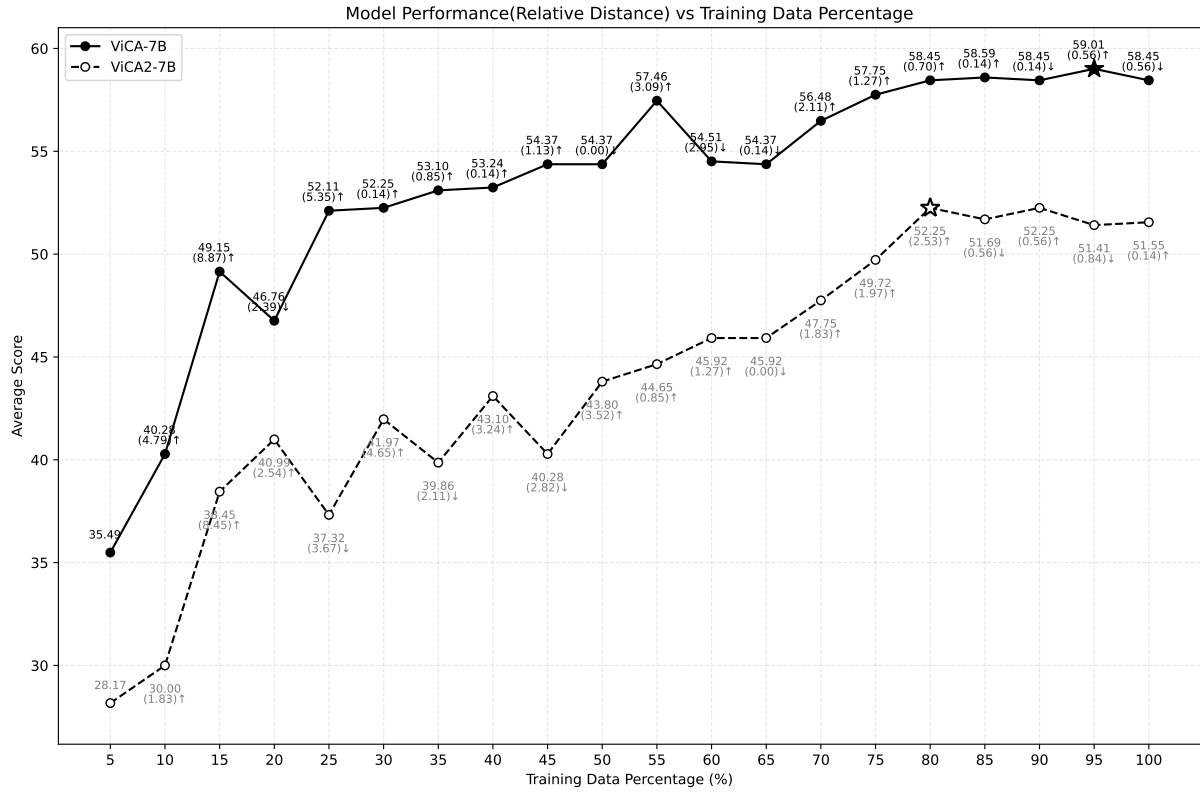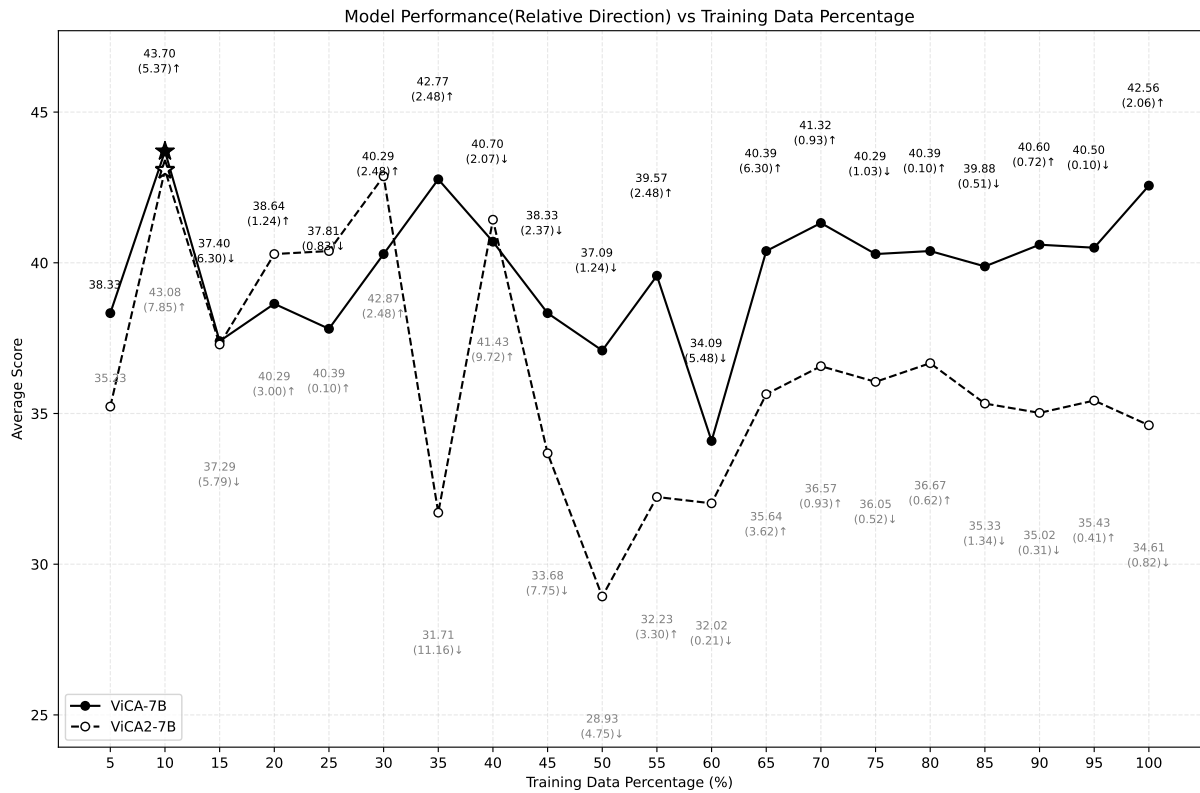

Figure 15: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Room Size**.

Figure 16: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Relative Distance**.

Figure 17: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Relative Direction**.
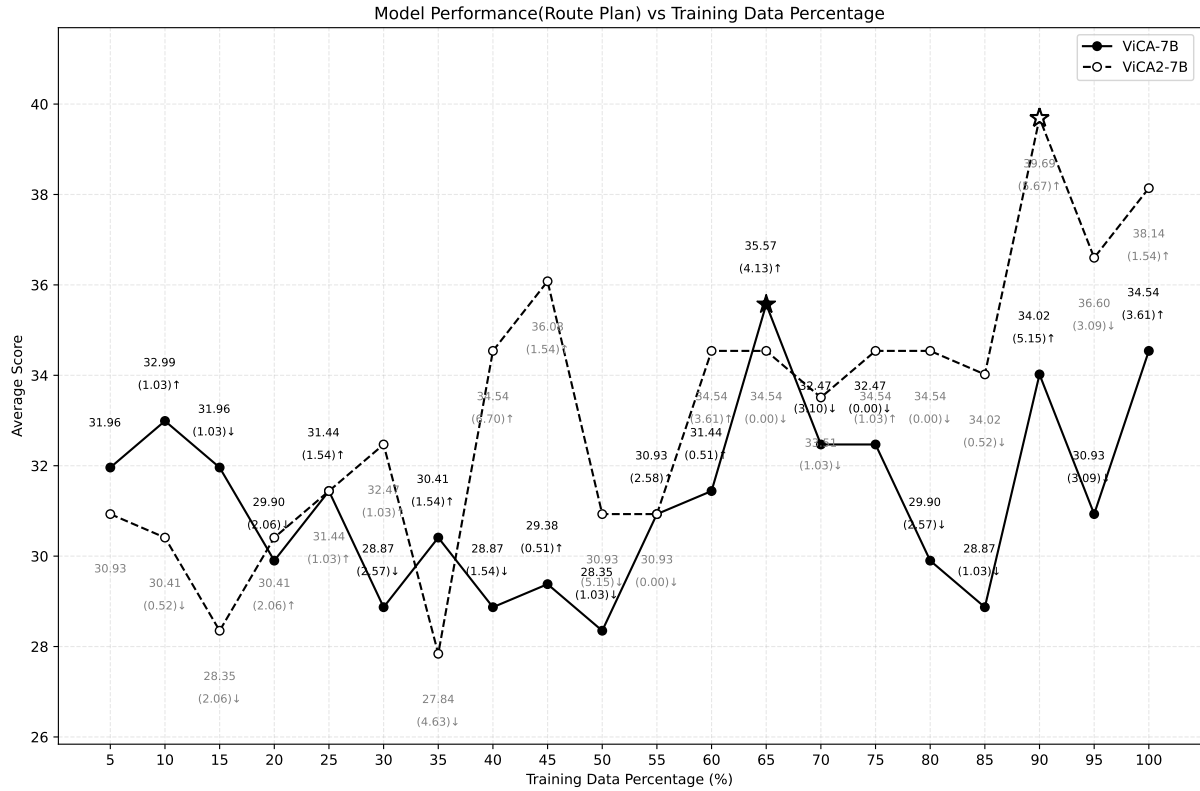
Figure 18: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Route Plan**.
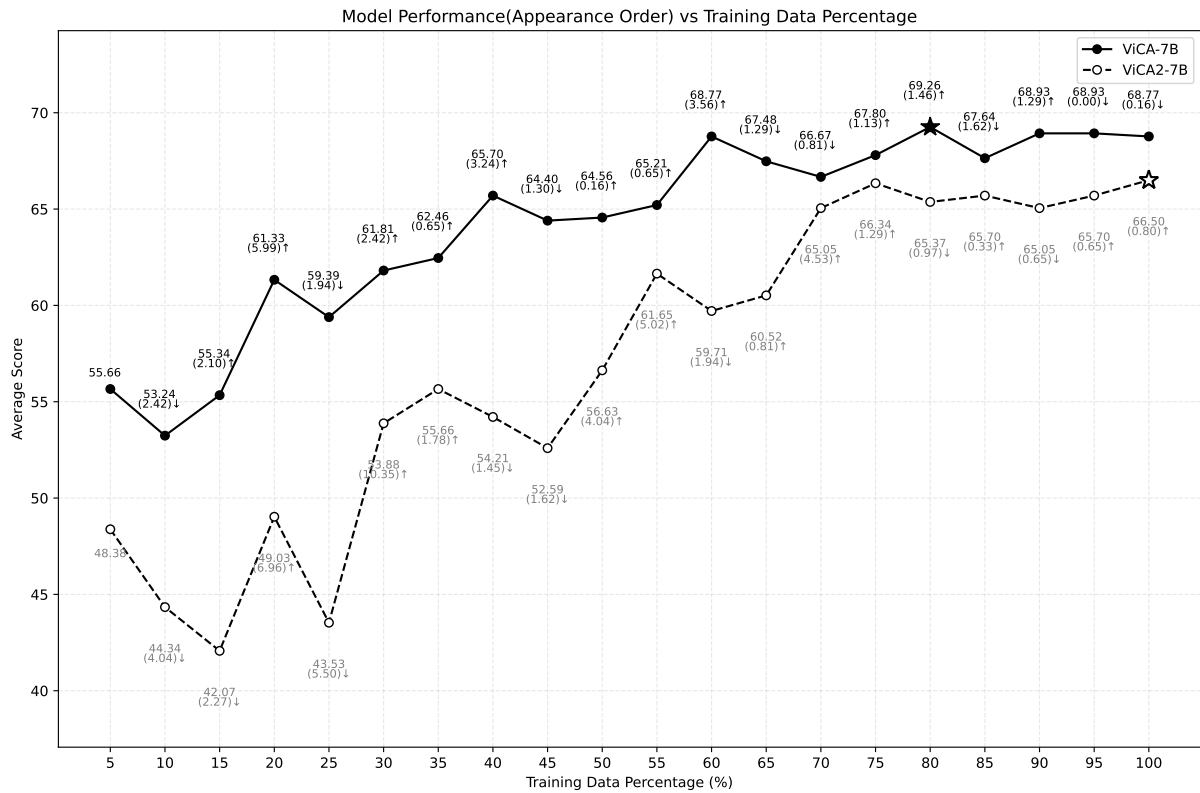

Figure 19: Performance Comparison of ViCA and ViCA2 with Increasing Training Data Size - **Appearance Order**.

E Evaluation Prompts for Checkpoint Text Quality Across Training Stages

The following are detailed descriptions of the same video, generated by four different Vision-Language Large Models (VLLMs). These models will be referred to as Model A, Model B, Model C, and Model D.

Your task is to critically evaluate and score the output from each of these VLLMs. The scoring scale is from 0 (minimum) to 10 (maximum).

Please ensure your evaluation addresses at least the following dimensions for each model's description: 1. Richness of Detail: How comprehensive and specific are the details provided about the video's content? 2. Accuracy: How accurately does the description reflect the presumed events, objects, and context within the video? 3. Organization/Coherence : How logically structured, clear, and easy to follow is the description? Is there a coherent narrative or flow? 4. Language Fluency: How natural, grammatically correct, and well-phrased is the language used? 5. Information Redundancy or Repetition: Does the description contain unnecessary repetition of information or superfluous content?

After providing your detailed textual evaluation for each model, discussing its performance across these dimensions, please conclude by returning only the final scores in the precise JSON format exemplified below.

```json
{
  "A": <score_for_model_A>,
  "B": <score_for_model_B>,
  "C": <score_for_model_C>,
  "D": <score_for_model_D>
}
```

[A]

<Detailed descriptions of the video generated by model A>

[B]

<Detailed descriptions of the video generated by model B>

[C]

<Detailed descriptions of the video generated by model C>

[D]

<Detailed descriptions of the video generated by model D>