
SSR: Enhancing Depth Perception in Vision-Language Models via Rationale-Guided Spatial Reasoning

Yang Liu^{1,†}, Ming Ma^{3,†}, Xiaomin Yu^{4,†}, Pengxiang Ding^{1,2,†,§},
Han Zhao^{1,2}, Mingyang Sun^{1,2,5}, Siteng Huang², Donglin Wang^{1*}
¹Westlake University, ²Zhejiang University, ³Harbin Institute of Technology,
⁴The Hong Kong University of Science and Technology (Guangzhou),
⁵Shanghai Innovation Institute
{liuyang67, wangdonglin}@westlake.edu.cn

Abstract

Despite impressive advancements in Visual-Language Models (VLMs) for multi-modal tasks, their reliance on RGB inputs limits precise spatial understanding. Existing methods for integrating spatial cues, such as point clouds or depth, either require specialized sensors or fail to effectively exploit depth information for higher-order reasoning. To this end, we propose a novel Spatial Sense and Reasoning method, dubbed SSR, a novel framework that transforms raw depth data into structured, interpretable textual rationales. These textual rationales serve as meaningful intermediate representations to significantly enhance spatial reasoning capabilities. Additionally, we leverage knowledge distillation to compress the generated rationales into compact latent embeddings, which facilitate resource-efficient and plug-and-play integration into existing VLMs without retraining. To enable comprehensive evaluation, we introduce a new dataset named SSR-CoT, a million-scale visual-language reasoning dataset enriched with intermediate spatial reasoning annotations, and present SSRBENCH, a comprehensive multi-task benchmark. Extensive experiments on multiple benchmarks demonstrate SSR substantially improves depth utilization and enhances spatial reasoning, thereby advancing VLMs toward more human-like multi-modal understanding. Project page: <https://yliu-cs.github.io/SSR>.

1 Introduction

VLMs represent a pivotal advancement in bridging the gap between image and natural language, demonstrating astounding capabilities across myriad multi-modal tasks [1–7]. Nevertheless, relying solely on RGB is inadequate for accurately capturing spatial information such as relative positions and distances, which presents inherent limitations in capturing precise spatial relationships, thereby constraining the capacity of VLMs to comprehend complex scenes. Consequently, enhancing the ability of VLMs to understand and reason about spatial relationships is essential for critical real-world applications, particularly in robotics.

Recent advancements in VLMs have catalyzed research on explicitly incorporating spatial information to enhance model performance. While some methods leverage point cloud data for improved spatial understanding [8–10], they typically rely on specialized sensors (e.g., LiDAR) that are impractical in scenarios restricted to monocular RGB images. In this context, monocular depth estimation has emerged as a compelling alternative, particularly with the proliferation of generative methods [11, 12]. These methods enable the acquisition of high-quality depth images from standard 2D images through

*Corresponding author. †Equal contribution. §Project lead.

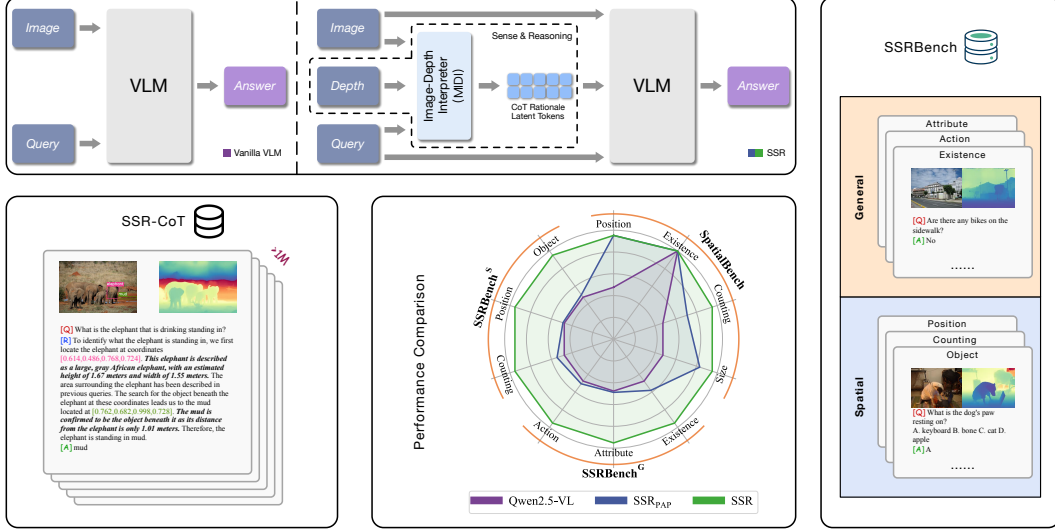


Figure 1: Unlike conventional VLMs, SSR integrates depth perception to enhance spatial reasoning. We introduce a curated dataset SSR-CoT and benchmark SSRBENCH, demonstrating significant improvements in spatial reasoning tasks.

various pre-trained models [13–16], eliminating additional hardware requirements. By leveraging visual encoders pretrained on RGB images, depth features can be efficiently encoded and seamlessly integrated into VLMs, offering a promising pathway for enhancing spatial awareness.

However, a critical limitation of current methods lies in their superficial utilization of depth information [8–10, 17–20]. Unlike humans, who intuitively employ depth as an integral component within broader reasoning processes, existing methods incorporate depth explicitly without capitalizing on its inferential value [17]. Consider a query such as *Are objects A and B far apart?* Human cognition naturally analyzes the spatial relationship between objects and then leverages this understanding to inform subsequent reasoning. This implicit reasoning process underscores the necessity for more sophisticated integration of depth information into VLMs, not merely as supplementary input, but as a fundamental component that facilitates complex spatial reasoning. Developing methodologies that emulate this human-like implicit utilization of depth could substantially enhance VLM capabilities.

To this end, we propose SSR, a novel paradigm designed to redefine the integration of depth information within VLMs. Specifically, SSR translates raw depth data into a structured *rationale language*, providing an interpretable intermediate representation that bridges low-level depth perception and higher-level reasoning. This rationale-based language facilitates VLMs in generating outputs that are both more accurate and contextually appropriate, while also enabling the previously underutilized inferential capabilities inherent to spatial depth information. By converting modality-specific depth data into semantically rich and inherently aligned representations, SSR effectively overcomes the interpretability limitations associated with traditional approaches. Consequently, this method significantly enhances the utilization of depth information, laying the groundwork for more robust and human-like spatial reasoning capabilities within contemporary VLMs. To further enhance the efficiency of rationale-language utilization, we transform depth information into a compact latent embedding. Specifically, we apply a knowledge-distillation strategy to compress rationale-language representations into concise latent embeddings [21]. Dissimilar to vanilla Chain-of-Thought (CoT) methods [22, 21, 23, 24] that rely primarily on textual explanations, our distillation strategy significantly reduces computational overhead while preserving the depth and inferential richness inherent to rationale-based representations [25–27]. Importantly, this module can be seamlessly integrated into existing VLMs via a training-free mechanism, highlighting the flexibility and broad applicability of the proposed framework. To achieve SSR, we first curate SSR-CoT, a million-level vision-language spatial reasoning dataset that facilitates depth-aware reasoning and provides a robust foundation for developing sophisticated spatial reasoning models. To validate our approach, we perform extensive experiments across multiple benchmarks. Specifically, we also evaluate on our benchmark SSR-BENCH, which comprises six distinct tasks spanning both general and spatial domains. Extensive

experiments and analysis demonstrate that SSR substantially enhances spatial reasoning capabilities across diverse tasks, highlighting the effectiveness and broad utility of our proposed method.

Overall, our principal contributions in this paper are illustrated in Figure 1 and summarized as follows:

- We propose an efficient VLM, dubbed SSR, capable of simultaneously performing depth perception and spatial reasoning, and generating answers based on implicit reasoning rationales.
- We introduce SSR-CoT, a million-scale visual-language reasoning dataset enriched with intermediate spatial reasoning annotations, and present SSRBENCH, a comprehensive multi-task benchmark.
- Extensive experiments and solid analysis across various benchmarks demonstrate our SSR can efficiently and dramatically enhance the spatial understanding of existing VLMs.

2 Related Work

2.1 Visual-Language Models

LLMs [24, 28–38] have led to major advancements in Natural Language Processing (NLP) tasks, and also have incited interest in developing VLMs. Building a unified LLM with visual inputs for visual language tasks thus remains one of the most important desiderata for VLMs. Over the last few years, VLMs achieved significant performance improvements in multi-modal tasks by integrating a pre-trained visual encoder and projecting the feature into semantic space into LLMs as well as training on large-scale multi-modal question-answering pairs [1, 2, 39–43]. This straightforward method can work well for general tasks, but expecting the model to deduce answers for more complex tasks without deep reasoning can be daunting.

2.2 Multi-Modal Reasoning

LLMs display an emergent capability for step-by-step reasoning through in-context learning, a phenomenon referred to as CoT reasoning. Such reasoning significantly enhances the performance of LLMs on complex reasoning tasks [22, 44, 45]. Concurrently, notable advancements have also occurred in multi-modal CoT research, a paradigm appealing due to its similarity to human problem-solving behaviors [46, 47]. Current research efforts in multi-modal CoT primarily emphasize the construction of intermediate reasoning rationale datasets to train image-text reasoning models. Many existing studies adopt rich textual captions and detailed descriptions as intermediate rationales [48–51]. Beyond text-based rationales, recent approaches have leveraged multi-modal rationales for more comprehensive reasoning [52–56]. However, existing multi-modal CoT methods primarily focus on tasks involving code generation, mathematical problem solving, and general question answering, which require sophisticated reasoning to achieve accurate responses. In contrast, this paper introduces an efficient CoT method that leverages depth images to enhance the performance of VLMs, particularly by improving spatial understanding.

2.3 Spatial Intelligence

Spatial reasoning is an essential capability for VLMs and has therefore been included in several Visual Question Answering (VQA) benchmarks [57–60]. However, the majority of existing VLMs [7, 61–64] are primarily trained on two-dimensional images paired with textual data, a setting that inherently lacks comprehensive spatial information. Consequently, these models exhibit limited performance in spatial reasoning tasks. To overcome this limitation, recent works such as SpatialVLM [10], SpatialRGPT [65] and RoboRefer [66] have sought to improve the spatial reasoning capacity of VLMs by compiling specialized spatially-oriented question-answer datasets and fine-tuning models accordingly. Nevertheless, despite these advancements, prior approaches largely neglect the integration of language-based reasoning capabilities within the spatial reasoning framework. This omission hampers the effectiveness of existing VLMs in addressing more complex tasks, particularly those requiring intricate or multi-step reasoning processes.

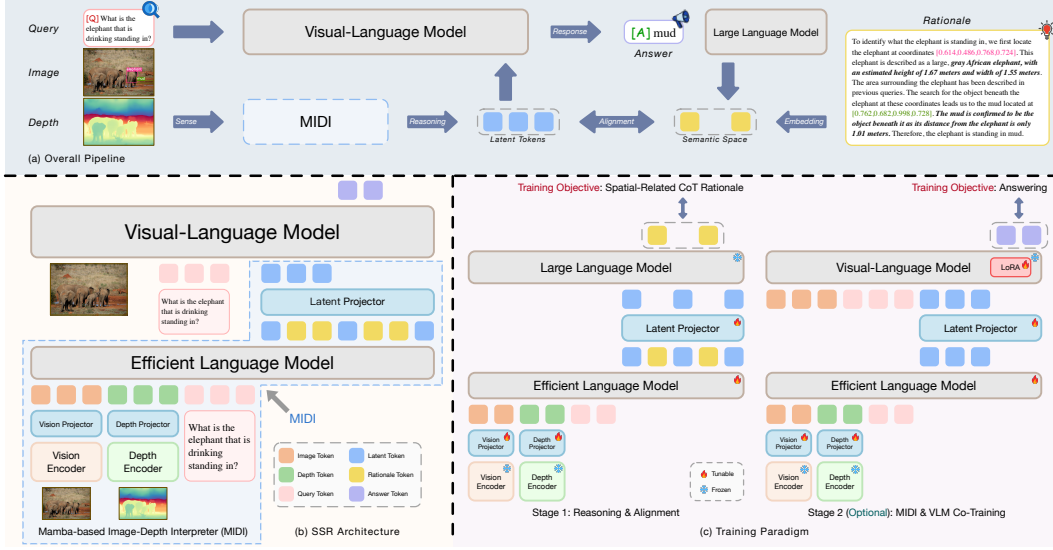


Figure 2: Schematic of SSR framework. (a) Overall pipeline. (b) Full architecture of SSR, comprising the MIDI module followed by the VLM. (c) Two training stages of the SSR. In the stage 1, the LLM provides alignment supervision for the MIDI module, whereas the stage 2 is optional.

3 Methodology

3.1 Architecture

The primary goal of our proposed SSR is to effectively leverage the reasoning capability of efficient language models effectively enhance the depth understanding and spatial reasoning capability for existing VLMs. The overall framework is illustrated in Figure 2.

3.1.1 Image-Depth Interpreter

To achieve comprehensive spatial understanding via depth interpretation, we propose a simple yet effective plug-and-play module entitled Mamba-based Image-Depth Interpreter (MIDI). MIDI generates enriched depth-aware latent token representations, providing essential spatial reasoning information before feeding these tokens into the VLM.

Given an input image $X_V \in \mathbb{R}^{H \times W \times 3}$ and a corresponding textual query X_T , we first utilize a pretrained monocular depth estimation model, Depth Pro [15], to produce a depth $X_D \in \mathbb{R}^{H \times W \times 1}$ in the image. Subsequently, image features Z_V and depth features Z_D are extracted from X_V and X_D , respectively. Specifically, we employ pre-trained CLIP ViT-L/14 [67, 68] as the visual encoder $\mathcal{E}_V$, and SigLIP [69] as the depth encoder $\mathcal{E}_D$: $H_\alpha = \mathcal{E}_\alpha(X_\alpha), \alpha \in \{V, D\}$. Then we apply Multi-Layer Perceptron (MLP) modules, comprising two fully connected layers with GELU [70] activation, as projectors ϕ_V and ϕ_D , transforming these visual features into the semantic embedding space compatible with the subsequent efficient language model: $Z_\alpha = \phi_\alpha(H_\alpha), \alpha \in \{V, D\}$. To jointly encode visual and depth information conditioned on the textual query, we introduce an intermediate reasoning module implemented as the Mamba-based language model [71], denoted as f_{LM} . This module produces latent tokens representing intermediate spatial rationales: $H_R = f_{LM}(Z_V, Z_D, X_T)$. Specifically, we uniformly insert several additional special tokens into the rationales to facilitate the knowledge distillation process and encode textual tokens into latent representations [21]. Finally, similar to previous steps, we apply an additional latent projection module ϕ_R to map these latent rationale tokens into another semantic embedding space, matching the dimensionality of the word embeddings used in the subsequent VLM: $Z_R = \phi_R(H_R)$.

Hence, our proposed MIDI module generates a sequence of spatial-aware latent tokens Z_R . These tokens can easily be plugged into the query sequence for existing VLMs, effectively injecting depth-based spatial reasoning information and significantly enhancing the spatial understanding capabilities.

Table 1: The mixture detail of SSR-CoT dataset. SSR-CoT consist over 1 million image-depth-question-rationale-answer pairs, where the rationale containing rich spatial-related knowledge the enhance Visual-Language Models (VLMs).

Dataset	Source	Size	Dataset	Source	Size
LLaVA-CoT [51]	ShareGPT4V [72]	31.3k	Visual-CoT [52]	Flickr30k [73]	136k
	ChartQA [74]	17.2k		GQA [57]	88k
	A-OKVQA [75]	16.1k		Visual7W [76]	43k
	AI2D [77]	11.4k		OpenImages [78]	43k
	GeoQA+ [79]	11.4k		Birds-200-2021 [80]	10k
	ScienceQA [81]	5.6k		VSR [82]	3k
	DocVQA [83]	4.0k	SCREENED TOTAL		289k
	PISC [84]	1.0k	VoCoT [53]	GQA [57]	72k
	CLEVR [85]	0.5k		LLaVA-Instruct [86]	6k
	CLEVR-Math [87]	0.5k		LVIS [88]	2k
TOTAL		98k	SCREENED ONE-TURN TOTAL		317k
SpatialQA [17]	Bunny [89]	695k	OXE [90]		7.5k
				SCREENED TOTAL	501k

3.1.2 Spatial Sense and Reasoning

Our proposed MIDI module fully leverages spatial information derived from the depth images and generates a latent representation Z_R , which encodes intermediate reasoning rationales essential for producing the response. Subsequently, we input these latent tokens Z_R , alongside the original image X_V and textual question X_T , into an existing VLM f_{VLM} to generate the answer Y_A in an auto-regressive manner: $Y_A = f_{VLM}(X_V, Z_R, X_T)$.

3.2 Training Paradigm

For training the proposed SSR, we adopt a two-stage procedure, as illustrated on the bottom-right side of Figure 2. In Stage 1, we train the underlying MIDI module to generate rationale latent tokens and project them into the language semantic space. In Stage 2, we conduct joint training of the MIDI module and existing Vision-Language Models (VLMs) to further enhance performance. Notably, Stage 2 is optional due to the modular and plug-and-play nature of our MIDI module, enabling straightforward integration into existing VLM frameworks.

3.2.1 Stage 1: Reasoning and Alignment

At the initial stage, we aim to train an efficient language model within the MIDI module, enabling it to generate and encode coherent thought processes represented by a sequence of features consistently aligned with the natural language semantic space. To this end, each training sample at this phase includes a detailed and accurate rationale Y_R as the ground truth. After feeding latent tokens produced by the MIDI module into the subsequent LLM, we require the LLM to reconstruct the original textual rationale solely from these latent representations. Training the LLM for precise rationale recovery depends not only on accurate reasoning capabilities of the MIDI module itself, but also on successfully projecting latent tokens into a semantic space consistent with the frozen-state LLM.

The learning objective for Stage 1 is defined by the standard causal modeling loss, given by:

$$\mathcal{L}_1(\theta) = -\mathbb{E}_{(X_V, X_D, X_T, Z_R, Y_R) \sim D} \left[\frac{1}{|Y_R|} \sum_{i=1}^{|Y_R|} \log P_{\theta}(Y_{R,i} \mid X_V, X_D, X_T, Z_R, Y_{R,<i}) \right]. \quad (1)$$

Following this training stage, latent tokens Z_R generated by the MIDI module can be readily integrated into existing VLM image-text sequences, thereby enhancing their spatial understanding capabilities.

3.2.2 Stage 2: Co-Training

To further enhance the performance of SSR, we jointly train the MIDI module along with existing VLMs. Similar to instruction-tuning, we discard intermediate rationales in the second training stage and allow the VLM to directly generate the final answer. In this setting, accurate answer generation by the VLM requires not only effective reasoning from the MIDI module but also the capacity of

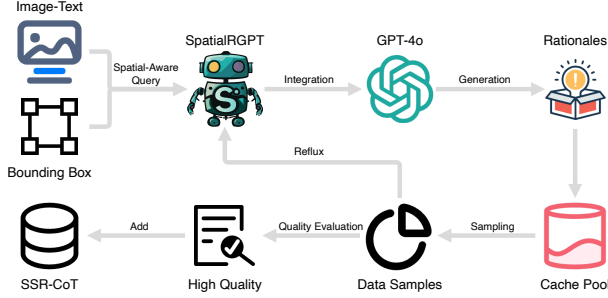


Figure 3: Schematic of SSR-CoT annotation pipeline.

Table 2: Quality evaluation for SSR-CoT dataset. We conduct the evaluation based on the powerful visual-language model Qwen2.5-VL-7B-Instruct [91].

Rationale	Accuracy	Score
✗	67.80	3.6721
✓	79.42	4.1289
	(↑11.62)	(↑0.4568)

Image & Depth				
Question	What is pushed by the person near the person?		What is the elephant that is drinking standing in?	
Rationale	The person is located at coordinates [0.244,0.449,0.485,0.832]. The object being pushed by the person at this location is a motorcycle, positioned at [0.335,0.523,0.501,0.795]. This motorcycle is nearby as indicated by the proximity of the person to it. The sequence of measurements confirms that the motorcycle is in the immediate area where the person is located, indicating that the person is interacting with the motorcycle. Therefore, the motorcycle is the object that the person is pushing.		To identify what the elephant is standing in, we first locate the elephant at coordinates [0.614,0.486,0.768,0.724]. This elephant is described as a large, gray African elephant, with an estimated height of 1.67 meters and width of 1.53 meters. The area surrounding the elephant has been described in previous queries. The search for the object beneath the elephant at these coordinates leads us to the mud located at [0.762,0.682,0.998,0.728]. The mud is confirmed to be the object beneath it as its distance from the elephant is only 1.01 meters. Therefore, the elephant is standing in mud.	
Answer	motorcycle		mud	

Figure 4: Illustrative samples of SSR-CoT dataset.

VLM to comprehend and utilize the reasoning information. Specifically, the second-stage learning objective is formulated as a standard cross-entropy loss for auto-regressive generation of the final answer Y_A :

$$\mathcal{L}_2(\theta) = -\mathbb{E}_{(X_V, X_D, X_T, Y_A) \sim D} \left[\frac{1}{|Y_A|} \sum_{j=1}^{|Y_A|} \log P_{\theta}(Y_{A,j} \mid X_V, X_D, X_T, Y_{A,<j}) \right]. \quad (2)$$

Since the rationale serving as the ground truth for supervised learning is omitted during Stage 2 training, we can incorporate additional VQA pairs to expand the dataset, thereby enhancing the generalization capability of the model. Furthermore, training during this stage is optional due to the modular plug-and-play nature of the MIDI module.

4 Experimentation

4.1 SSR-CoT Collection

There is a scarcity of visual-language CoT datasets with detailed reasoning processes annotations to train the SSR model for depth perception and spatial understanding. Therefore, we curate a new dataset from existing VQA datasets, resulting in over a total of 1 million image-depth-question-rationale-answer pairs. There are four dataset sources we integrated: (1) **LLaVA-CoT** [51]: Systematic and structured reasoning visual-language CoT dataset, including general and science-targeted VQA data source. (2) **Visual-CoT** [52]: Multimodal CoT dataset that takes the bounding box as an intermediate thinking step, including general, relation reasoning and fine-grained science-targeted VQA data source. (3) **VoCoT** [53]: Fine-grained image-text CoT dataset that rationale provides detailed relationships between various objects with bounding box, including general and relation reasoning VQA data source. (4) **SpatialQA** [17]: Spatial QA dataset for sufficient utilization, including depth-related and robotic-related VQA data sources.

To generate visual-language reasoning data enriched with spatial information, we follow a multi-step process, as shown in Figure 3. First, we extract depth estimations from raw images using Depth Pro [15]. For the LLaVA-CoT [51] source, this is the only preprocessing step performed. Second, for

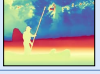
General (432)				Spatial (357)			
Existence (107)			[Q] Are there any bikes on the side walk? [A] No	Counting (41)			[Q] How many birds are parked on the top of the boat? A. six B. two C. one D. three [A] B
			[Q] Are there any clocks in the scene? [A] The answer is yes.				[Q] How many pancakes can you see in the image? [A] I can see four pancakes in the image.
Action (157)			[Q] Why is the man in the black shorts jumping in the air? A. tackling player B. catch frisbee C. dodging ball D. to exercise [A] B	Position (152)			[Q] Is the dog to the right or to the left of the person on the stick? [A] The answer is right.
			[Q] What is the boy doing? A. running B. descending C. ascending D. stretching [A] B				[Q] Is the cab to the left of a girl? [A] No
Attribute (168)			[Q] What is the color and material of the cabinets? [A] The cabinets in the kitchen are made of wood and are brown in color.	Object (154)			[Q] What utensil is in the bowl? [A] Spoon
			[Q] What corporation made the shirt the woman is wearing? A. hanes B. everlast C. adidas D. new balance [A] The cabinets in the kitchen are made of wood and are brown in color.				[Q] What is the dog's paw resting on? A. keyboard B. bone C. cat D. apple [A] A

Figure 5: Examples for each task within the benchmark SSRBENCH.

datasets such as VoCoT [53] and SpatialQA [17], we refine long-form conversations by extracting concise, one-turn question-answer pairs. Third, we leverage SpatialRGPT [65] to comprehensively mine precise spatial attributes within images, such as object size, distance, and relative positioning, based on intermediate reasoning steps, including bounding box annotations from Visual-CoT [52] and VoCoT [53]. Finally, we employ GPT-4o [92] to integrate all extracted information, generating detailed reasoning processes that enhance spatial understanding. Notably, we also incorporate cache pools and perform sampling quality checks within iterative loops to ensure the high quality of the generated data. Specifically, similar to the quality-assessment protocol shown in Table 1, we randomly draw 10% of the cached samples and evaluate VQA accuracy both with and without their generated rationales. Rationales that improve accuracy are retained and incorporated into the final SSR-CoT dataset; those that degrade performance are discarded, and the all samples in cache are re-submitted for re-annotation. Overall, we compile approximately 1.2 million preprocessed data samples into SSR-CoT dataset. Figure 4 illustrates several data samples from the SSR-CoT dataset. Each data instance within SSR-CoT comprises the original image, an associated question-answer pair, the corresponding estimated depth information, and a rationale. The rationale incorporates fundamental reasoning steps used in question-answering tasks and provides detailed spatial reasoning to support accurate answer generation.

To evaluate the quality of SSR-CoT, we conducted an assessment based on the performance of the Qwen2.5-VL-7B-Instruct [91] on the VQA task. This evaluation was carried out on a randomly selected subset comprising approximately 1% of the full dataset, corresponding roughly to 10k samples. Performance metrics include accuracy as well as a quantitative score ranging from 0 to 5, both are produced using the LLM-Assistant powered by the Qwen2.5-14B-Instruct-1M [38, 93]. Further methodological details regarding the evaluation process are described in Appendix D. As presented in Table 2, responses generated with intermediate rationales demonstrate an accuracy improvement of more than 10% compared to direct question-answering. This finding indicates that the intermediate reasoning rationales annotated in our dataset are of high quality and effectively enhance the question-answering performance of VLMs.

4.2 SSRBENCH Construction

Currently, there are no established benchmarks specifically designed for evaluating spatial understanding and reasoning capabilities on image-text pairs. To address this gap, we propose SSRBENCH, a novel evaluation benchmark created from the SSR-CoT dataset. Importantly, the data incorporated into SSRBENCH will be fully removed from SSR-CoT to prevent overlap. SSRBENCH consists of two primary categories, general understanding and spatial understanding, allowing simultaneous evaluation of VLM performance in both general question answering and spatial reasoning tasks. Each category contains three distinct evaluation tasks, with detailed sample sizes provided in Appendix E.

Table 3: Performance comparison on SpatialBench [17] and our SSRBENCH. SSRBENCH^G and SSRBENCH^S denote general and spatial tasks, respectively.

Method	Size		SpatialBench [17]				SSRBENCH ^G			SSRBENCH ^S		
	Mamba	VLM	Position	Existence	Counting	Size	Existence	Attribute	Action	Counting	Position	Object
PROPRIETARY												
GPT-4o-mini [92]	N/A	N/A	47.1	75.0	70.5	21.7	72.0	48.2	63.1	53.7	44.4	46.1
Claude-3.5-Haiku [99]	N/A	N/A	55.9	65.0	72.2	26.7	51.4	52.2	43.7	42.0	34.1	38.3
OPEN-SOURCE												
SpatialVLM [10, 100]	N/A	3B	52.9	80.0	77.1	28.3	31.7	58.3	63.7	31.7	55.8	65.4
LLaVA-1.5 [2]	N/A	7B	44.1	45.0	82.8	30.0	81.3	64.3	66.9	43.9	63.6	63.6
LLaVA-NeXT [86]	N/A	7B	47.1	75.0	84.0	20.0	83.2	66.7	69.4	51.2	69.8	64.9
LLaVA-NeXT [86]	N/A	13B	47.1	75.0	82.9	20.0	86.9	69.6	71.3	41.5	69.8	53.2
SpatialBot [17]	N/A	3B	50.0	80.0	86.7	25.0	75.7	61.3	67.5	39.0	74.1	61.7
Emu3 [101]	N/A	8B	47.1	20.0	10.0	25.0	58.9	35.7	37.6	19.5	51.9	37.0
Qwen2.5-VL [91]	N/A	3B	55.9	80.0	76.4	25.0	66.4	58.9	63.1	34.1	60.5	51.9
Qwen2.5-VL [91]	N/A	7B	61.8	80.0	87.1	30.0	75.7	62.5	70.1	43.9	61.7	55.2
SSR (Ours)	130M	3B	64.7	80.0	82.9	31.7	83.2	82.1	72.6	51.2	83.3	74.7
SSR (Ours)	130M	7B	64.7	85.0	90.2	28.3	90.7	79.2	76.4	65.9	84.6	77.9

Table 4: Performance improvement of SSR compared to the backbone model. SpatialBench [17], SSRBENCH, and CV-Bench [102] report average.

Method	Size		SpatialBench [17]	SSRBENCH		CV-Bench [102]	VSR [82]		What's Up [103]
	Mamba	VLM		General	Spatial		Random	Zero-Shot	
Qwen2.5-VL [91]	N/A	3B	59.3	62.8	48.8	67.0	73.0	76.4	85.4
SSR (Ours)	130M	3B	64.8 (↑5.4)	79.3 (↑16.5)	69.7 (↑20.9)	68.9 (↑1.9)	78.6 (↑5.6)	82.9 (↑6.5)	87.9 (↑2.5)
Qwen2.5-VL [91]	N/A	7B	64.7	69.4	53.6	73.0	N/A	N/A	N/A
SSR (Ours)	130M	7B	67.0 (↑2.3)	82.1 (↑12.6)	76.1 (↑22.5)	73.3 (↑0.3)	N/A	N/A	N/A

We illustrate the process to construct SSRBENCH as shown in Appendix E.2. First, we define 6 distinct task categories. Then, we randomly sampled image-text pairs from SSR-CoT, proportionally retaining the distribution of its original data sources. These samples were independently classified into task categories by GPT-4o [92] and Gemini-2.5-Pro [94]. Only instances for which both models agreed on the assigned category were included in SSRBENCH; instances with disagreement were returned to SSR-CoT.

In recent years, LLMs have demonstrated significant advancements in language understanding, reasoning, and text generation, exhibiting strong perceptual and comprehension capabilities through the implicit world knowledge they encapsulate. Therefore, LLMs have increasingly been used as assessors to evaluate generation performance in question-answering tasks [7, 1, 95, 96]. Consistent with our approach in data quality evaluation, we employ the Qwen2.5-14B-Instruct-1M [38, 93], a powerful LLM, to evaluate the performance of VLMs in this benchmark.

4.3 Implementation Details

In this paper, we utilize Mamba [71] as the lower-level efficient language model for reasoning, Qwen2.5 [38] as the LLM for alignment in the first training stage, and Qwen2.5-VL [91] as the VLM supporting multi-modal comprehension in the second training stage. During Stage 1, we exclusively train the MIDI component on our proposed SSR-CoT dataset. In Stage 2, we jointly train the SSR using both the SSR-CoT dataset and the LLaVA-Instruct-150K dataset [1]. Leveraging the efficiency of LoRA [97] and Fully Sharded Data Parallel (FSDP) [98], training SSR requires approximately 19 hours for Stage 1 and 48 hours for Stage 2, using a single Nvidia 8-H800 GPU node equipped with 80GB VRAM. Detailed hyperparameter configurations are provided in Appendix A.

4.4 Main Results

Table 3 presents the comparative performance of the SSR against its backbone and state-of-the-art baselines on SpatialBench [17] and SSRBENCH. As shown by the results, our SSR in 3 billion parameters can achieve comparable or even higher results than large-scale baseline models, including closed-source and backbone models. Our larger variant, comprising 7 billion parameters, yields the best performance on most tasks across the two benchmarks. Compared to the top-performing baselines in each benchmark, SSR exhibited notable improvements in the average question answering accuracy, achieving a maximum enhancement of 13.6 and an average improvement of 6.77. Moreover, we also provide a detailed analysis of the performance improvements compared to the backbone model across additional benchmarks in Section 5.1.

Table 5: Performance comparison on general VQA benchmarks

Method	Size		VQAv2 [104]	TextVQA [105]	POPE [106]	MMBench [107]	GQA [57]
	Mamba	VLM					
Qwen2.5-VL [91]	N/A	3B	72.5	57.0	84.4	75.9	56.2
SSR (Ours)	130M	3B	79.0 ($\uparrow 6.5$)	61.3 ($\uparrow 4.3$)	86.0 ($\uparrow 1.6$)	78.3 ($\uparrow 2.4$)	63.6 ($\uparrow 7.4$)

Table 6: Performance comparison among the backbone model, the SSR with/without the second training stage. _{PAP} indicates that the MIDI module was employed in a plug-and-play manner.

Method	Size		SpatialBench [17]				SSRBENCH ^G			SSRBENCH ^S		
	Mamba	VLM	Position	Existence	Counting	Size	Existence	Attribute	Action	Counting	Position	Object
Qwen2.5-VL [91]	N/A	3B	55.9	80.0	76.4	25.0	66.4	58.9	63.1	34.1	60.5	51.9
SSR_{PAP}	130M	3B	64.7 ($\uparrow 8.8$)	80.0 (0.0)	79.6 ($\uparrow 3.2$)	30.0 ($\uparrow 5.0$)	70.1 ($\uparrow 3.7$)	59.5 ($\uparrow 0.6$)	63.7 ($\uparrow 0.6$)	36.6 ($\uparrow 2.5$)	61.1 ($\uparrow 0.6$)	53.2 ($\uparrow 1.3$)
SSR	130M	3B	64.7 ($\uparrow 8.8$)	80.0 (0.0)	82.9 ($\uparrow 6.5$)	31.7 ($\uparrow 6.7$)	83.2 ($\uparrow 16.8$)	82.1 ($\uparrow 23.2$)	72.6 ($\uparrow 9.5$)	51.2 ($\uparrow 17.1$)	83.3 ($\uparrow 22.8$)	74.7 ($\uparrow 22.8$)

5 Analysis

5.1 Performance Improvement

We present additional experimental results in Table 4, demonstrating the improved performance of SSR compared to the backbone model across the five benchmarks shown in Table 11 at varying model scales. Specifically, across the three benchmarks reporting average values, SSR models of different sizes demonstrated average improvements of **11.2** and **9.4** compared to the backbone model. The most significant improvements were observed in the space task of the benchmark, where the enhancements reached **20.9** and **22.5**, respectively. This result exceeds the improvements reported in Table 2, indicating that our SSR effectively reasons about information highly relevant to multi-modal VQA tasks without introducing significant additional noise. Furthermore, the training paradigm of SSR enhances performance not only on two evaluation datasets closely related to the training data but also on the out-of-domain CV-Bench [102], VSR [82], What’s Up [103] and multiple general VQA benchmarks [57, 104–107] as shown in Table 5. These findings indicate that our training approach effectively further improves the generality and generalization capability of the SSR in addition to enhanced spatial understanding performance.

5.2 Ablation Studies

As reported in Table 6, these experiments illustrate the performance of the MIDI module when integrated in a plug-and-play manner without second training stage, leading to improved spatial understanding. Specifically, this approach achieves average performance gains of **4.4** and **1.6** on different benchmark datasets. On certain tasks, the plug-and-play approach achieves performance improvements of up to **8.8**, demonstrating the effectiveness of this usage. In addition, after the second stage training, the performance of the complete SSR model will be significantly improved on this basis, achieving average performance gains of **5.7** and **18.7** on different benchmark datasets. Moreover, we provide case studies in Appendix F.

5.3 Efficiency

We evaluate the Qwen2.5-VL [91] after fine-tuning on SSR-CoT dataset, overly protracted and convoluted textual intermediate reasoning chains not only increase the risk of erroneous conclusions but also impose prohibitive computational costs that undermine inference efficiency, which can better reflect the importance of latent reasoning method in CoT application. As illustrated in Table 7, the results demonstrate that, although SSR introduces a modest absolute latency per generated token, its latent-reasoning paradigm dramatically curtails the number of CoT tokens needed to reach a final response. Consequently, under the CoT-based evaluation framework, the overall end-to-end inference speed is substantially improved.

Table 7: Inference efficiency comparison on SpatialBench [17].

Model	Size	SpatialBench [17]	Token Per Sample	Token Per Second	Inference Time Per Sample
Qwen2.5-VL [91] (w/ SFT on SSR-CoT)	3B	51.3	437.28	18.88	23.16s
SSR (Ours)	3B	64.8	2.62	8.18	0.32s

Table 8: LLM-Assistant Evaluation Performance comparison on SSRBENCH, evaluated using different LLMs. (For each result, the left column presents the original Qwen-based evaluation score, while the right column reports the corresponding GPT judgment).

Model	Size	Existence	Attribute	Action	Counting	Position	Object
Qwen2.5-VL [91]	3B	66.4 / 58.9	58.9 / 59.5	63.1 / 67.5	34.1 / 34.1	60.5 / 58.6	51.9 / 52.6
SSR (Ours)	3B	83.2 / 81.3	82.1 / 79.8	72.6 / 73.3	51.2 / 51.2	83.3 / 80.8	74.7 / 74.7

5.4 LLM-Assistant Evaluation

To mitigate potential biases arising from employing models of the same family as judges, we re-assessed the SSRBENCH results with GPT-4o-mini [92], the outcomes are reported in Table 8. High inter-model agreement between the evaluation scores assigned by different LLM judges in the table suggests that simple answer-comparison tasks largely resist bias within model series.

5.5 Rationale Embedding

To analyze whether MIDI effectively captures depth information and conducts spatial reasoning guided by rationale, we visualize the cosine similarity between latent tokens, both with and without rationale. Figure 6 visualizes the cosine similarities between the latent tokens produced in two different paradigms: x-axis: latent tokens inserted inside the rationale, y-axis: latent tokens inserted immediately after the question and used to start the answer generation. Diagonal cells represent these two states of the same sample. High values on the diagonal indicate that the model has learned to map the rationale to the latent representation, confirming that it successfully distills the spatial knowledge embedded in the rationale. Low off-diagonal values indicate that the latent tokens remain sample-specific and do not collapse to a generic representation.

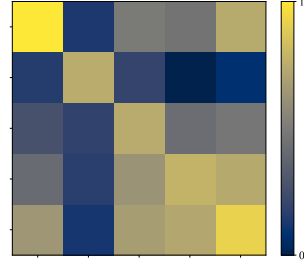


Figure 6: Cosine similarity matrix of reasoning latent tokens with/without rationale.

6 Conclusion

In this paper, we propose a novel VLM SSR with an important module named MIDI to interpret depth for enhancing the depth perception and spatial reasoning capabilities of existing VLMs. MIDI can even be efficiently integrated into existing VLMs in a seamless, plug-and-play manner. To enable effective training and evaluation, we curate a multi-modal CoT dataset SSR-CoT and present a comprehensive benchmark SSRBENCH. Extensive experiments conducted across four distinct benchmarks demonstrate that SSR consistently achieves state-of-the-art performance enhancements over existing approaches, particularly excelling in spatially-oriented visual question answering tasks.

Broader Impacts. Our proposed SSR demonstrates that spatial reasoning capabilities can be incrementally enhanced without adversely affecting its existing VLM functionalities. This provides an innovative avenue for research communities to integrate additional capabilities into VLMs.

Limitation and Future Works. Although SSR shows astounding performance, this study is limited to the Qwen/Qwen-VL series; future work will broaden the VLM scope to test generalizability.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215).

References

- [1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” in *Proc. of NeurIPS*, 2023.
- [2] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved Baselines with Visual Instruction Tuning,” in *Proc. of CVPR*, pp. 26286–26296, 2024.
- [3] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu, “Osprey: Pixel Understanding with Visual Instruction Tuning,” in *Proc. of CVPR*, pp. 28202–28211, 2024.
- [4] H. A. Rasheed, M. Maaz, S. S. Mullappilly, A. M. Shaker, S. H. Khan, H. Cholakkal, R. M. Anwer, E. P. Xing, M. Yang, and F. S. Khan, “GLaMM: Pixel Grounding Large Multimodal Model,” in *Proc. of CVPR*, pp. 13009–13018, 2024.
- [5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick, “Segment Anything,” in *Proc. of ICCV*, pp. 3992–4003, 2023.
- [6] Y. Liu, H. Ping, D. Zhang, Q. Sun, S. Li, and G. Zhou, “Comment-aware multi-modal heterogeneous pre-training for humor detection in short-form videos,” in *Proc. of ECAI*, pp. 1568–1575, 2023.
- [7] Y. Liu, P. Ding, S. Huang, M. Zhang, H. Zhao, and D. Wang, “PiTe: Pixel-Temporal Alignment for Large Video-Language Model,” in *Proc. of ECCV*, pp. 160–176, 2024.
- [8] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, “3D-VLA: A 3D Vision-Language-Action Generative World Model,” in *Proc. of ICML*, 2024.
- [9] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, “3D-LLM: Injecting the 3D World into Large Language Models,” in *Proc. of NeurIPS*, 2023.
- [10] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. J. Guibas, and F. Xia, “SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities,” in *Proc. of CVPR*, pp. 14455–14465, 2024.
- [11] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in *Proc. of ICLR*, 2021.
- [12] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Proc. of NeurIPS*, 2020.
- [13] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data,” in *Proc. of CVPR*, pp. 10371–10381, 2024.
- [14] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth Anything V2,” in *Proc. of NeurIPS*, 2024.
- [15] A. Bochkovskiy, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. Richter, and V. Koltun, “Depth Pro: Sharp Monocular Metric Depth in Less Than a Second,” in *Proc. of ICLR*, 2025.
- [16] X. He, D. Guo, H. Li, R. Li, Y. Cui, and C. Zhang, “Distill Any Depth: Distillation Creates a Stronger Monocular Depth Estimator,” *CoRR*, 2025.
- [17] W. Cai, Y. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, “SpatialBot: Precise Spatial Understanding with Vision Language Models,” in *Proc. of ICRA*, 2025.
- [18] S. Chen, X. Chen, C. Zhang, M. Li, G. Yu, H. Fei, H. Zhu, J. Fan, and T. Chen, “LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning,” in *Proc. of CVPR*, pp. 26418–26428, 2024.
- [19] J. Zhang, M. Cai, T. Xie, and Y. J. Lee, “CounterCurate: Enhancing Physical and Semantic Visio-Linguistic Compositional Reasoning via Counterfactual Examples,” in *Proc. of ACL Findings*, pp. 15481–15495, 2024.

- [20] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, “RoboPoint: A Vision-Language Model for Spatial Affordance Prediction for Robotics,” *CoRR*, 2024.
- [21] B. Lee, C. W. Kim, B. Park, and Y. M. Ro, “Meteor: Mamba-based Traversal of Rationale for Large Language and Vision Models,” in *Proc. of NeurIPS*, 2024.
- [22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *Proc. of NeurIPS*, 2022.
- [23] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappeler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O’Connell, I. Osband, I. C. Gilaberte, and I. Akkaya, “OpenAI o1 System Card,” *CoRR*, 2024.
- [24] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, and S. S. Li, “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,” *CoRR*, 2025.
- [25] S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian, “Training Large Language Models to Reason in a Continuous Latent Space,” *CoRR*, 2024.
- [26] J. Zhang, Y. Zhu, M. Sun, Y. Luo, S. Qiao, L. Du, D. Zheng, H. Chen, and N. Zhang, “LightThinker: Thinking Step-by-Step Compression,” *CoRR*, 2025.
- [27] S. Feng, G. Fang, X. Ma, and X. Wang, “Efficient Reasoning Models: A Survey,” *CoRR*, 2025.
- [28] A. Radford and K. Narasimhan, “Improving Language Understanding by Generative Pre-Training,” 2018.
- [29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” 2019.
- [30] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in *Proc. of NeurIPS*, 2020.
- [31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” in *Proc. of NeurIPS*, 2022.

- [32] OpenAI, “ChatGPT: Optimizing Language Models for Dialogue,” 2023.
- [33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and Efficient Foundation Language Models,” *CoRR*, 2023.
- [34] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” *CoRR*, 2023.
- [35] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, “Qwen Technical Report,” *CoRR*, 2023.
- [36] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan, “Qwen2 Technical Report,” *CoRR*, 2024.
- [37] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Srivankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al., “The Llama 3 Herd of Models,” *CoRR*, 2024.
- [38] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, “Qwen2.5 Technical Report,” *CoRR*, 2024.
- [39] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, “Flamingo: a Visual Language Model for Few-Shot Learning,” in *Proc. of NeurIPS*, 2022.
- [40] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Y. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, “OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models,” *CoRR*, 2023.

- [41] J. Xu, X. Zhou, S. Yan, X. Gu, A. Arnab, C. Sun, X. Wang, and C. Schmid, “Pixel Aligned Language Models,” in *Proc. of CVPR*, pp. 13030–13039, 2024.
- [42] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models,” in *Proc. of ICLR*, 2024.
- [43] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, “Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference,” in *Proc. of AAAI*, pp. 10421–10429, 2025.
- [44] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, “Training Verifiers to Solve Math Word Problems,” *CoRR*, 2021.
- [45] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, “Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies,” *Trans. Assoc. Comput. Linguistics*, pp. 346–361, 2021.
- [46] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, and C. Gan, “See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning,” *CoRR*, 2023.
- [47] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, “Multimodal Chain-of-Thought Reasoning in Language Models,” *Trans. Mach. Learn. Res.*, 2024.
- [48] B. Luan, H. Feng, H. Chen, Y. Wang, W. Zhou, and H. Li, “TextCoT: Zoom In for Enhanced Multimodal Text-Rich Image Understanding,” *CoRR*, 2024.
- [49] Z. Jia, J. Liu, H. Li, Q. Liu, and H. Gao, “DCoT: Dual Chain-of-Thought Prompting for Large Multimodal Models,” in *Proc. of ACML*, pp. 1064–1079, 2024.
- [50] T. Gao, P. Chen, M. Zhang, C. Fu, Y. Shen, Y. Zhang, S. Zhang, X. Zheng, X. Sun, L. Cao, and R. Ji, “Cantor: Inspiring Multimodal Chain-of-Thought of MLLM,” in *Proc. of ACM MM*, pp. 9096–9105, 2024.
- [51] G. Xu, P. Jin, H. Li, Y. Song, L. Sun, and L. Yuan, “LLaVA-CoT: Let Vision Language Models Reason Step-by-Step,” *CoRR*, 2024.
- [52] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, “Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning,” in *Proc. of NeurIPS*, 2024.
- [53] Z. Li, R. Luo, J. Zhang, M. Qiu, and Z. Wei, “VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models,” in *Proc. of NAACL*, 2025.
- [54] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei, “Mind’s Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models,” in *Proc. of NeurIPS*, 2024.
- [55] F. Meng, H. Yang, Y. Wang, and M. Zhang, “Chain of Images for Intuitively Reasoning,” *CoRR*, 2023.
- [56] C. Li, W. Wu, H. Zhang, Y. Xia, S. Mao, L. Dong, I. Vulic, and F. Wei, “Imagine while Reasoning in Space: Multimodal Visualization-of-Thought,” *CoRR*, 2025.
- [57] D. A. Hudson and C. D. Manning, “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering,” in *Proc. of CVPR*, pp. 6700–6709, 2019.
- [58] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, “CLEVRER: Collision Events for Video Representation and Reasoning,” in *Proc. of ICLR*, 2020.
- [59] C. H. Song, V. Blukis, J. Tremblay, S. Tyree, Y. Su, and S. Birchfield, “RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics,” *CoRR*, 2024.

- [60] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie, “Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces,” *CoRR*, 2024.
- [61] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. V. Thapliyal, J. Bradbury, and W. Kuo, “PaLI: A Jointly-Scaled Multilingual Language-Image Model,” in *Proc. of ICLR*, 2023.
- [62] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” in *Proc. of ICML*, pp. 19730–19742, 2023.
- [63] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, “PaLM-E: An Embodied Multimodal Language Model,” in *Proc. of ICML*, pp. 8469–8488, 2023.
- [64] S. Wu, H. Fei, L. Qu, W. Ji, and T. Chua, “NEX-T-GPT: Any-to-Any Multimodal LLM,” in *Proc. of ICML*, 2024.
- [65] A. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, “SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models,” in *Proc. of NeurIPS*, 2024.
- [66] E. Zhou, J. An, C. Chi, Y. Han, S. Rong, C. Zhang, P. Wang, Z. Wang, T. Huang, L. Sheng, and S. Zhang, “RoboRefer: Towards Spatial Referring with Reasoning in Vision-Language Models for Robotics,” *CoRR*, 2025.
- [67] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proc. of ICML*, pp. 8748–8763, 2021.
- [68] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proc. of ICLR*, 2021.
- [69] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid Loss for Language Image Pre-Training,” in *Proc. of ICCV*, pp. 11941–11952, 2023.
- [70] C. Yu and Z. Su, “Symmetrical Gaussian Error Linear Units (SGELUs),” *CoRR*, 2019.
- [71] A. Gu and T. Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” in *Proc. of CoLM*, 2024.
- [72] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, “ShareGPT4V: Improving Large Multi-modal Models with Better Captions,” in *Proc. of ECCV*, pp. 370–387, 2024.
- [73] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” in *Proc. of ICCV*, pp. 2641–2649, 2015.
- [74] A. Masry, D. X. Long, J. Q. Tan, S. R. Joty, and E. Hoque, “ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning,” in *Proc. of ACL Findings*, pp. 2263–2279, 2022.
- [75] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, “A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge,” in *Proc. of ECCV*, pp. 146–162, 2022.
- [76] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei, “Visual7W: Grounded Question Answering in Images,” in *Proc. of CVPR*, pp. 4995–5004, 2016.
- [77] A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi, “A Diagram is Worth a Dozen Images,” in *Proc. of ECCV*, pp. 235–251, 2016.

- [78] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, T. Duerig, and V. Ferrari, “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale,” *CoRR*, 2018.
- [79] J. Cao and J. Xiao, “An Augmented Benchmark Dataset for Geometric Question Answering through Dual Parallel Text Encoding,” in *Proc. of COLING*, pp. 1511–1520, 2022.
- [80] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.
- [81] P. Lu, S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering,” in *Proc. of NeurIPS*, 2022.
- [82] F. Liu, G. Emerson, and N. Collier, “Visual Spatial Reasoning,” *Trans. Assoc. Comput. Linguistics*, pp. 635–651, 2023.
- [83] M. Mathew, D. Karatzas, and C. V. Jawahar, “DocVQA: A Dataset for VQA on Document Images,” in *Proc. of WACV*, pp. 2199–2208, 2021.
- [84] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, “People in Social Context (PISC) Dataset,” 2017.
- [85] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, “CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning,” in *Proc. of CVPR*, pp. 1988–1997, 2017.
- [86] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, “Llava-next: Improved reasoning, ocr, and world knowledge,” January 2024.
- [87] A. D. Lindström and S. S. Abraham, “CLEVR-Math: A Dataset for Compositional Language, Visual and Mathematical Reasoning,” in *Proc. of IJCLR*, pp. 155–170, 2022.
- [88] A. Gupta, P. Dollár, and R. B. Girshick, “LVIS: A Dataset for Large Vocabulary Instance Segmentation,” in *Proc. of CVPR*, pp. 5356–5364, 2019.
- [89] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, “Efficient Multimodal Learning from Data-centric Perspective,” *CoRR*, 2024.
- [90] A. O’Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. E. Wang, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. P. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Wang, H. Su, H. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Sünderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Bajjal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine,

- S. Lin, S. Moore, S. Bahl, S. Dass, S. D. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, L. Xu, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, and Z. Lin, “Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration,” in *Proc. of ICRA*, pp. 6892–6903, 2024.
- [91] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, “Qwen2.5-VL Technical Report,” 2025.
- [92] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. L. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, and D. Sherburn, “GPT-4o System Card,” *CoRR*, 2024.
- [93] A. Yang, B. Yu, C. Li, D. Liu, F. Huang, H. Huang, J. Jiang, J. Tu, J. Zhang, J. Zhou, J. Lin, K. Dang, K. Yang, L. Yu, M. Li, M. Sun, Q. Zhu, R. Men, T. He, W. Xu, W. Yin, W. Yu, X. Qiu, X. Ren, X. Yang, Y. Li, Z. Xu, and Z. Zhang, “Qwen2.5-1M Technical Report,” *CoRR*, 2025.
- [94] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. P. Lillicrap, J. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, and et al., “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” *CoRR*, 2024.
- [95] Y. Niu, M. Ning, M. Zheng, B. Lin, P. Jin, J. Liao, K. Ning, B. Zhu, and L. Yuan, “WISE: A World Knowledge-Informed Semantic Evaluation for Text-to-Image Generation,” *CoRR*, 2025.
- [96] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, “MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities,” in *Proc. of ICML*, 2024.
- [97] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” in *Proc. of ICLR*, 2022.
- [98] Y. Zhao, A. Gu, R. Varma, L. Luo, C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, A. Desmaison, C. Balioglu, P. Damania, B. Nguyen, G. Chauhan, Y. Hao, A. Mathews, and S. Li, “PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel,” *Proc. VLDB Endow.*, pp. 3848–3860, 2023.
- [99] Anthropic, “Claude 3.5 Haiku.” <https://www.anthropic.com/news/3-5-models-and-computer-use>, 2024.

- [100] remyxai, “VQASynth,” 2024. GitHub repository.
- [101] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, Y. Zhao, Y. Ao, X. Min, T. Li, B. Wu, B. Zhao, B. Zhang, L. Wang, G. Liu, Z. He, X. Yang, J. Liu, Y. Lin, T. Huang, and Z. Wang, “Emu3: Next-Token Prediction is All You Need,” *CoRR*, 2024.
- [102] P. Tong, E. Brown, P. Wu, S. Woo, A. Iyer, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, X. Pan, R. Fergus, Y. LeCun, and S. Xie, “Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs,” in *Proc. of NeurIPS*, 2024.
- [103] A. Kamath, J. Hessel, and K. Chang, “What’s “up” with vision-language models? Investigating their struggle with spatial reasoning,” in *Proc. of EMNLP*, pp. 9161–9175, 2023.
- [104] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering,” in *Proc. of CVPR*, pp. 6325–6334, 2017.
- [105] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, “Towards VQA Models That Can Read,” in *Proc. of CVPR*, pp. 8317–8326, 2019.
- [106] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. Wen, “Evaluating Object Hallucination in Large Vision-Language Models,” in *Proc. of EMNLP*, pp. 292–305, 2023.
- [107] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, “MMBench: Is Your Multi-modal Model an All-Around Player?,” in *Proc. of ECCV*, pp. 216–233, 2024.
- [108] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *Proc. of ICLR*, 2019.
- [109] I. M. Alabdulmohsin, X. Zhai, A. Kolesnikov, and L. Beyer, “Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design,” in *Proc. of NeurIPS*, 2023.

A Hyperparameters

Detailed hyperparameter configurations are provided in Table 9.

Table 9: Training hyper-parameters of our proposed SSR.

Configuration	Stage 1	Stage 2	Configuration	Stage 1	Stage 2
Vision Encoder	Clip-ViT-Large-Patch14-336 [67]		Optimizer	AdamW [108]	
Depth Encoder	Siglip-So400M-Patch14-384 [69, 109]		Learning Rate	0.00002	
Mamba	Mamba 130M [71]		Numerical Precision	BFloat16	
LLM	Qwen2.5 3B [38]	N/A	Epoch	2	1
VLM	N/A	Qwen2.5-VL 3B [91]	Global Batch Size	32	32
Question Length		256	Learning Schedule	Cosine Decay	
Rational Length	1024	N/A	Warm-up Ratio	0.02	
Answering Length	N/A	256	Number of Latent Tokens	10	

B SSRBENCH Results in Score Metrics

The evaluation metrics employed for SSRBENCH include both accuracy and a quantitative score ranging from 0 to 5. Quantitative results are presented in Table 10, with detailed descriptions of the assessment methodology provided in Appendix D. These scores are generally consistent with the accuracy trends presented in Table 3.

Table 10: Score performance comparison on SSRBENCH. SSRBENCH^G and SSRBENCH^S denote general and spatial tasks, respectively.

Method	Size		SSRBENCH ^G			SSRBENCH ^S		
	Mamba	VLM	Existence	Attribute	Action	Counting	Position	Object
PROPRIETARY								
GPT-4o-mini [92]	N/A	N/A	4.05	2.95	3.46	3.12	2.87	2.66
Claude-3.5-Haiku [99]	N/A	N/A	3.48	2.99	2.71	2.75	2.56	2.31
OPEN-SOURCE								
SpatialVLM [10, 100]	N/A	3B	2.34	3.32	3.55	2.34	3.56	3.24
LLaVA-1.5 [2]	N/A	7B	4.17	3.72	3.66	2.71	3.87	3.56
LLaVA-NeXT [86]	N/A	7B	4.23	3.59	3.79	2.66	3.69	3.41
LLaVA-NeXT [86]	N/A	13B	4.30	3.82	3.79	2.76	3.78	3.12
SpatialBot [17]	N/A	3B	3.97	3.47	3.82	2.66	3.96	3.47
Emu3 [101]	N/A	8B	3.07	2.35	2.39	1.71	3.04	2.28
Qwen2.5-VL [91]	N/A	3B	3.56	3.42	3.56	2.41	3.43	3.00
Qwen2.5-VL [91]	N/A	7B	4.07	3.55	3.71	2.85	3.50	3.16
SSR (Ours)	130M	3B	4.44	4.28	3.95	3.17	4.40	4.02
SSR (Ours)	130M	7B	4.65	4.17	4.10	3.71	4.43	4.16

Discussion with Meteor. Meteor [21] is an approach similar to ours, designed to compress rationales using efficient large language models. However, unlike our method, Meteor does not separate the reasoning module from the large language model during response generation. Due to this tight coupling, Meteor must be trained end-to-end from scratch, a process that demands extensive datasets and significant computational resources. In contrast, our method specifically focuses on enhancing the spatial awareness and reasoning abilities of Vision-Language Models (VLMs), leveraging their inherent capabilities to a greater extent. Consequently, our approach substantially reduces the complexity and resource requirements related to training VLMs from scratch. Moreover, we focus on directionally enhancing the depth perception and spatial reasoning capabilities of existing VLMs in this paper. Therefore, our comparative analysis primarily emphasizes evaluating model performance before and after applying these enhancements.

C SSR-CoT

As detailed in Section 4, the SSR-CoT is constructed from four distinct data sources, with spatially-aware CoT rationales generated for each data sample. Representative samples are shown in Figure 4.

Additionally, a comprehensive description of the rationale-generation pipeline specific to each data source is presented in Appendix C.2, C.1, and C.3.

C.1 Visual-CoT

For Visual-CoT [52], each data sample includes a bounding box that serves as a CoT rationale to guide the generation of the corresponding answer. We utilize this bounding box, which is closely related to the target answer, as an intermediate step to query spatial information about the selected object using SpatialRGPT [65], a spatial question-answering model tailored for vertical domains. Subsequently, we aggregate the obtained spatial question-and-answer information using a powerful LLM such as GPT-4o [92]. The resulting text serves as the CoT rationale for the Visual-CoT data source within SSR-CoT.

Spatial Query for Visual-CoT

1. What is the object in [bbox]? Think step by step, and avoid repetition.
2. Can you estimate the height and width of [bbox]? Think step by step, and avoid repetition.
3. What is the object to the left of [bbox], and what is its height and width? Think step by step, and avoid repetition.
4. What is the object to the left of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
5. What is the object to the right of [bbox], and what is its height and width? Think step by step, and avoid repetition.
6. What is the object to the right of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
7. What is the object in front of [bbox], and what is its height and width? Think step by step, and avoid repetition.
8. What is the object in front of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
9. What is the object behind [bbox], and what is its height and width? Think step by step, and avoid repetition.
10. What is the object behind [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
11. What is the object below [bbox], and what is its height and width? Think step by step, and avoid repetition.
12. What is the object below [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
13. What is the object above [bbox], and what is its height and width? Think step by step, and avoid repetition.
14. What is the object above [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.

Rationale Generation for Visual-CoT

Please generate an image description in continuous paragraphs using these strict guidelines:

Coordinate Usage Rules:

1. ONLY use coordinates that are explicitly defined in this mapping table:
 - Region [0]: [bbox]
 - ...
2. Do NOT create or infer any new coordinates
3. Each coordinate can only be used ONCE in the description

4. Coordinates must be written in [x1,y1,x2,y2] format without spaces

Content Rules:

1. Place coordinate immediately after describing its corresponding object
2. Integrate coordinates naturally within complete sentences
3. Include all provided measurements and spatial relationships
4. Maintain narrative flow while incorporating technical details
5. Focus on visual elements and their relationships
6. Embed coordinates from the mapping table immediately after their corresponding region objects (e.g., "a dog [x1,y1,x2,y2]")
7. Maintain paragraph continuity by integrating coordinates within complete sentences
8. Preserve strict region-coordinate mapping from the provided table
9. Use only [x1,y1,x2,y2] format without spaces
10. Exclude technical metadata and region index numbers from final text
11. Automatically resolve spatial contradictions using coordinate data
12. Ensure coordinate annotations flow naturally after object nouns

Input Data:

Spatial Query and Response for Visual-CoT

C.2 VoCoT

VoCoT [53] includes multiple bounding boxes per data sample, more than Visual-CoT [52], to clearly outline reasoning paths involving multiple objects within an image. Similar to the process used for Visual-CoT, we perform spatial queries on each object associated with a bounding box. Additionally, we capture the relative spatial relationships between every pair of objects to comprehensively utilize available spatial context and support accurate reasoning. Finally, we aggregate this spatially derived question-and-answer information using a robust language model, such as GPT-4o [92].

Spatial Query for VoCoT

Query for each bounding box:

1. What is the object in [bbox]? Think step by step, and avoid repetition.
2. Can you estimate the height and width of [bbox]? Think step by step, and avoid repetition.
3. What is the object to the left of [bbox], and what is its height and width? Think step by step, and avoid repetition.
4. What is the object to the left of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
5. What is the object to the right of [bbox], and what is its height and width? Think step by step, and avoid repetition.
6. What is the object to the right of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
7. What is the object in front of [bbox], and what is its height and width? Think step by step, and avoid repetition.
8. What is the object in front of [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
9. What is the object behind [bbox], and what is its height and width? Think step by step, and avoid repetition.
10. What is the object behind [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
11. What is the object below [bbox], and what is its height and width? Think step by step, and avoid repetition.

12. What is the object below [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.
13. What is the object above [bbox], and what is its height and width? Think step by step, and avoid repetition.
14. What is the object above [bbox], and what is its distance to [bbox]? Think step by step, and avoid repetition.

Query for every two bounding box:

1. Which one is higher between [bbox1] and [bbox2]? Think step by step, and avoid repetition.
2. Can you estimate how far apart [bbox1] and [bbox2] are? Think step by step, and avoid repetition.
3. What direction is [bbox2] in relation to [bbox1]? Think step by step, and avoid repetition.
4. How far is [bbox1] from [bbox2] horizontally? Think step by step, and avoid repetition.
5. Does [bbox1] have a larger size compared to [bbox2]? Think step by step, and avoid repetition.
6. Does [bbox1] have a lesser width compared to [bbox2]? Think step by step, and avoid repetition.

Rationale Generation for VoCoT

Integrate all measurements values and spatial information from the conversation into answer to get detailed reasoning rationale with spatial details.
Then, extract the direct question and answer from question and answer respectively.

Content Rules:

1. Place coordinate immediately after describing its corresponding object first time, make sure each coordinate appear only once.
2. Avoid introducing other coordinates that do not appear in answer.
3. Add all provided measurements values and spatial relationships from the conversation to the rationale detailedly.
4. Ensure the rationale contains all the information from each sentence in the conversation, especially the measurements values and spatial relationships.
5. Automatically resolve spatial contradictions using coordinate data based on the image.

Output in the following json template:

```
{
  "question": <question>
  , "rationale": <rationale>
  , "answer": <answer>
}
```

Question: **Question**

Answer: **Answer**

Conversation: **Spatial Query and Response for VoCoT**

C.3 SpatialQA

Unlike Visual-CoT [52] and VoCoT [53], the SpatialQA [17] dataset does not provide intermediate CoT reasoning steps or bounding boxes for object identification. Therefore, we leverage GPT-4o [92], a powerful multi-modal large language model, to generate detailed synthetic rationale data. These

synthetic rationales supply the necessary intermediate reasoning processes to enable accurate answer generation.

Rationale Generation for SpatialQA

I have an image and a question that I want you to answer.

I need you to strictly follow the format with four specific sections: summary, caption, reasoning, and conclusion.

It is crucial that you adhere to this structure exactly as outlined and that the final answer in the conclusion matches the standard correct answer precisely.

To explain further:

- In summary, briefly explain what steps you'll take to solve the problem.
 - In caption, describe the contents of the image, specifically focusing on details relevant to the question.
 - In reasoning, outline a step-by-step thought process you would use to solve the problem based on the image.
 - In conclusion, give the final answer in a direct format, and it must match the correct answer exactly. If it's a multiple choice question, the conclusion should only include the option without repeating what the option is.
- Finally, integrate these sections into a natural thinking paragraph.

Here's the question and answer:

Question: Question

Answer: Answer

D LLM-Assistant Evaluation

As discussed in Section 4, we utilize the LLM-Assistant evaluation method to assess the data quality of SSR-COT and measure the performance of SSRBENCH [7, 1, 95, 96]. Evaluation metrics include accuracy and a quantitative score ranging from 0 to 5; both metrics are determined by the LLM-Assistant powered by the Qwen2.5-14B-Instruct-1M [38, 93].

Prompt for LLM-Assistant VQA Evaluation

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs.

Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

##INSTRUCTIONS:

- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

Please evaluate the following image-based question-answer pair:

Question: question

Correct Answer: answer

Predicted Answer: response

Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match.

Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: {'pred': 'yes', 'score': 4.8}.

E Benchmarks

E.1 Benchmark Employed

We evaluate our method using various benchmarks: SpatialBench [17], CV-Bench [102], VSR [], What’s Up [], our proposed SSRBENCH, and multiple general VQA benchmarks []. Table 11 summarizes the statistics for the all benchmarks. Tables 3 and 6 present comparisons between SSR and baseline methods, along with comprehensive ablation studies conducted on SpatialBench and SSRBENCH. Furthermore, Table 4 summarizes the performance improvements observed across all spatial-related benchmarks.

Table 11: Statistics of benchmarks utilized in this paper.

Benchmark	Task	Size	Benchmark	Task	Size	
SpatialBench [17]	Position	34	CV-Bench [102]	Count	788	
	Existence	40		Relation	650	
	Counting	20		Depth	600	
	Size	40		Distance	600	
TOTAL		114	TOTAL		2638	
SSRBENCH	General	Existence	107	Spatial	Counting	41
		Attribute	168		Position	162
		Action	157		Object	154
TOTAL					789	
VSR [82]	Random	1222	Zero-Shot		2195	
TOTAL					3417	
What’s Up [103]	TOTAL	820	VQAv2 [104]	TOTAL	107k	
TextVQA [105]	TOTAL	5k	POPE [106]	TOTAL	9k	
MMBench [107]	TOTAL	3k	GQA [57]	TOTAL	12k	

E.2 SSRBENCH

To construct the SSRBENCH dataset, we first filter data samples from SSR-CoT. Subsequently, we feed each filtered sample into the multi-modal large language models GPT-4o [92] and Gemini-2.5-Pro [94], using the following prompt to classify the task category. As mentioned in Section 4 and shown in Figure 7, if the classification results from both models are consistent, the sample is added to SSRBENCH; otherwise, it is returned to SSR-CoT.

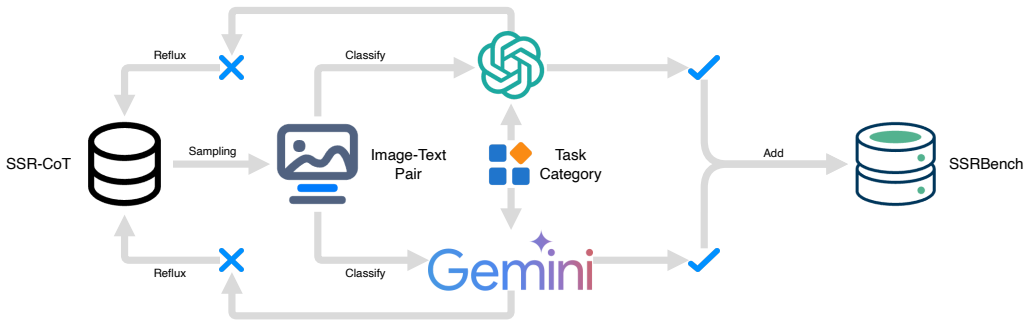


Figure 7: Schematic of SSRBENCH construction pipeline.

Task classification for SSRBENCH

You are an expert in image-based question classification.
You need to classify each input question into a specific task type based on the following

taxonomy.

Task Categories:

Spatial:

Explanation: Involve identifying and understanding the position, size, shape, and relative relationships of objects in an image.

Subtasks:

Count: Counting objects in the image (e.g., questions like “How many ...?”).

Relative Position Recognition: Determining spatial relations like “to the left of”, “above”, or “on the right”.

Position Based Object Recognition: Identifying an object based on its spatial relation to another object (e.g., “What is the object to the left of the dog?”).

General:

Explanation: Involve classifying, recognizing, or reasoning about visual content without necessarily focusing on spatial relations.

Subtasks:

Existence: Determining whether an object or feature is present (e.g., “Is there a cat?”).

Attribute Recognition: Identifying attributes like color, texture, size, or state (e.g., “What color is the apple?”).

Action Recognition: Recognizing what action or activity is occurring (e.g., “What is the man doing?”).

For each input question:

First determine whether the question belongs to the spatial or general category.

Then classify it into one of the three subtasks under that category.

If the question does not match any of the subtasks under either category, return None.

Output format:

{ "category": "spatial" or "general", "subtask": "subtask_name" or "None" }

Example Input: "Is there a bicycle in the image?"

Example Output: { "category": "general", "subtask": "existence" }

Now, let’s begin classification. Here’s the question:

Question: **Question**

F Case Studies

SpatialBench					SSRBench						
											
[Q] Are the hands of the man away from the banana? [A] Yes					[Q] What is placed on the wooden table? [A] Tablecloth						
Claude-3.5-Haiku	No	✗	LLaVA-NeXT-7B	No	✗	Claude-3.5-Haiku	Nothing	✗	LLaVA-NeXT-7B	Dishes	✗
GPT-4o-mini	No	✗	Qwen2.5-VL-3B	No	✗	GPT-4o-mini	None	✗	Qwen2.5-VL-3B	Chair	✗
SpatialBot	No	✗	SSR-3B (Ours)	Yes	✓	SpatialBot	Plant	✗	SSR-3B (Ours)	Tablecloth	✓

Figure 8: Two examples illustrating question-answering performance by baseline models compared to our SSR are presented.

To further illustrate the effectiveness of our proposed SSR, we provide two example cases in Figure 8, comparing the performance of SSR against five baseline models: Claude-3.5-Haiku [99], GPT-4o-mini [92], SpatialBot [17], LLaVA-NeXT-7B [86], and the backbone model Qwen2.5-VL-3B [91].

As shown, our SSR consistently produces correct answers, whereas all baseline models fail to provide accurate responses.

In the left example, the images depict only people and bananas. Consequently, the model must abandon conventional assumptions and carefully reason about the spatial relations explicitly present in the image to answer accurately. In the right example, complex relationships among numerous objects are depicted, and relevant features for answering the posed question are not immediately obvious. In this case, the model must thoroughly comprehend the correspondence between each object and the given question, as well as understand intricate spatial relations among these objects, to produce a correct response. These examples clearly demonstrate that our SSR effectively enhances the spatial awareness and reasoning capabilities of vision-language models, thereby significantly improving their ability to understand complex spatial relationships.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The claims we presented in the abstract and introduction are clearly stated and fully aligned with the contributions of this paper.

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: We discuss the limitations in the Section 6..

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[NA\]](#)

Justification: None of theoretical assumptions.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: We provide the implementation details to reproduce the main experimental results in Section 4.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: Instructions for reproducing our experiments are provided in Section 4. We will publicly release the data and code once they have been finalized and prepared.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: The experimental setup required for our study is described in Section 4.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [No]

Justification: Conducting experiment for statistical significance when training large-scale models typically demands an exponential increase in computational resources. To address this concern, we performed comprehensive multi-angle analyses on SSR, and our experimental results consistently validated its effectiveness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar rather than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [Yes]

Justification: Details regarding the computational resources used for all experiments are presented in Section 4.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: The research presented in this paper fully complies with the NeurIPS Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: We discuss impacts in the Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. Safeguards

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

Answer: [NA]

Justification: We used a publicly available visual instruction tuning dataset and pre-trained visual foundation models and large language models.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: We cited models and datasets we deal with in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

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

Answer: [Yes]

Justification: We propose a novel model SSR, curate a dataset SSR-COT, and construct a benchmark SSRBENCH.

Guidelines:

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

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

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

Answer: [NA]

Justification: We do not include human subjects in this paper.

Guidelines:

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

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

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

Answer: [NA]

Justification: We do not include human subjects in this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [\[Yes\]](#)

Justification: We utilized LLM for data synthesis and the reporting of performance indicators, which are comprehensively described throughout this paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.