Orchestrating Visual and Linguistic Modalities for Robust Spatial Intelligence in LVLMs

Jie-Hao Lim, Carter Ross, Gavin Walker

Singapore Institute of Management

Abstract. The ability of large vision-language models (LVLMs) to understand and reason about complex spatial relationships within visual scenes is critical for advancing artificial intelligence, particularly in domains like robotics and augmented reality. Despite their impressive general capabilities, current LVLMs often struggle with fine-grained spatial grounding, exhibiting limitations in precisely describing relative object positions, sizes, and distances, and in performing multi-step spatial reasoning. This paper introduces **Multi-Granularity Spatial-Relational Graph Transformer (MGS-RGT) Training**, a novel two-stage learning paradigm designed to significantly enhance LVLMs' spatial intelligence. Our method first involves **Hierarchical Spatial Graph Prediction (HSGP)** pre-training, which rigorously trains the visual encoder to represent multi-scale spatial relationships (fine-grained, object-level, and scene-level) through explicit graph learning. Following this, the full LVLM undergoes **Spatially-Grounded Language Generation (SGLG)** fine-tuning, enriched with **Chain-of-Thought (CoT)** integration, guiding the model to articulate its spatial reasoning process. Comprehensive experiments on standard VQA benchmarks and our new Spatially-Grounded Interaction Dataset (SGID) demonstrate that MGS-RGT consistently and substantially outperforms state-of-the-art baselines across Spatial VQA Accuracy, Relationship Prediction F1-Score, and Task Success Rate for embodied tasks. Ablation studies confirm the critical contributions of both HSGP pre-training and CoT integration. Qualitative analysis and human evaluations further corroborate MGS-RGT's ability to generate highly accurate, detailed, and coherent spatial descriptions, validating its superior spatial reasoning capabilities.

1 Introduction

The ability of artificial intelligence systems to understand and interact with the physical world hinges critically on their capacity for **spatial reasoning**. Traditional computer vision models have achieved remarkable success in object recognition, segmentation, and detection [1,2], but their understanding often remains at a semantic level, failing to grasp the intricate 3D spatial relationships between objects, their relative positions, sizes, and distances. This limitation becomes particularly evident in complex, real-world scenarios where inferring spatial layouts and predicting physical interactions are paramount for effective

decision-making, such as in robotics, augmented reality, and autonomous navigation. The recent advancements in **Large Vision-Language Models (LVLMs)** have shown impressive capabilities in integrating visual perception with linguistic understanding, enabling more natural human-AI interaction and sophisticated visual question answering [2,3]. However, as highlighted by recent works like "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities" [4], even these powerful LVLMs often struggle with explicit and fine-grained spatial reasoning, frequently defaulting to semantic associations rather than robust spatial grounding. This gap underscores a fundamental challenge: how can we imbue LVLMs with a truly comprehensive and actionable understanding of spatial relationships, moving beyond mere co-occurrence to genuine geometric and physical awareness? The primary **challenge** lies in the inherent nature of how LVLMs process visual information. Current architectures typically flatten images into sequences of patches, or extract abstract features, which can inadvertently discard crucial low-level spatial and topological cues. While positional embeddings provide some sense of location, they are often insufficient for encoding complex, hierarchical spatial relationships (e.g., "a book on top of a table, which is next to a chair, and all are within the living room"). Furthermore, existing training methodologies for LVLMs, even when exposed to "spatially enhanced" datasets, often do not explicitly enforce the learning of such structured spatial knowledge. This leads to LVLMs sometimes hallucinating spatial details or providing coarse, inaccurate spatial descriptions that lack the precision required for real-world applications. Our **motivation** for this research stems from the belief that explicitly modeling and integrating hierarchical spatial relationships into the LVLM's learning process is crucial for unlocking truly intelligent spatial reasoning. We aim to bridge the gap between abstract semantic understanding and concrete spatial grounding, enabling LVLMs to accurately describe, reason about, and predict outcomes based on intricate spatial arrangements. In this paper, we propose a novel approach for enhancing LVLMs with superior spatial reasoning capabilities, termed **Multi-Granularity Spatial-Relational Graph Transformer (MGS-RGT) Training**. Our method introduces a **two-stage training paradigm** designed to systematically teach LVLMs hierarchical spatial understanding. In the **first stage**, we pre-train the visual encoder of the LVLM on a novel task called **Hierarchical Spatial Graph Prediction (HSGP)**. For each image, we automatically construct multi-granularity spatial graphs that capture relationships at pixel-level adjacency, object-level relative positions (e.g., "left of," "above," "overlaps with"), and scene-level structural layouts. The HSGP task encourages the visual encoder to learn rich, spatially-aware representations that explicitly encode these intricate relationships, moving beyond mere abstract features. In the **second stage**, the entire LVLM, incorporating this spatially-aware visual encoder, undergoes fine-tuning on a diverse set of **Spatially-Grounded Language Generation (SGLG)** tasks. This involves leveraging a newly curated dataset where image-text pairs are augmented with explicit spatial prompts and require precise spatial reasoning outputs. Crucially, we integrate **Chain-of-Thought (CoT) reasoning** prompts

during training, guiding the LVLM to articulate its step-by-step spatial inference process [5]. This not only enhances the accuracy of spatial answers but also provides valuable insights into the model's reasoning, promoting interpretability. For our experiments, we utilize a comprehensive dataset combining existing large-scale visual question answering benchmarks [6] with our newly constructed **Spatially-Grounded Interaction Dataset (SGID)**. The SGID is meticulously annotated with fine-grained spatial relationships, object affordances, and temporal sequences of spatial changes derived from simulated and real-world robot interaction logs [7]. We evaluate our proposed MGS-RGT model against state-of-the-art LVLMs across a spectrum of challenging spatial reasoning tasks. Our evaluation metrics include accuracy on various spatial question types (e.g., relative position, size comparison, distance estimation), task success rates in simulated embodied environments, and a novel **Spatial Coherence Score (SCS)**, which quantifies the logical consistency and correctness of the generated spatial descriptions and reasoning chains. The preliminary results demonstrate that our MGS-RGT approach significantly outperforms existing baselines, achieving notable improvements across all spatial reasoning benchmarks. Specifically, our model shows superior performance in understanding complex 3D spatial arrangements, accurately interpreting subtle spatial cues, and generating coherent spatial narratives. This substantiates the effectiveness of our hierarchical spatial graph learning strategy and its profound impact on enhancing LVLM's spatial intelligence. Our contributions are summarized as follows:

- We propose a novel two-stage training paradigm, MGS-RGT Training, that explicitly integrates multi-granularity spatial-relational graph learning into LVLMs, significantly enhancing their spatial reasoning capabilities.
- We introduce Hierarchical Spatial Graph Prediction (HSGP) as a pre-training task for the visual encoder, enabling it to learn and encode intricate, multi-scale spatial relationships directly from visual inputs.
- We demonstrate that combining Spatially-Grounded Language Generation (SGLG) tasks with Chain-of-Thought (CoT) reasoning during fine-tuning leads to superior and more interpretable spatial understanding in LVLMs, validating our approach through comprehensive experimental evaluation on diverse spatial benchmarks.

2 Related Work

2.1 Large Vision-Language Models

The rapid advancements in deep learning have led to the emergence of powerful Large Vision-Language Models (LVLMs), which aim to bridge the gap between visual perception and natural language understanding. These models typically integrate a high-capacity visual encoder with a sophisticated language model, enabling them to process and reason across multimodal inputs. Early foundational work in this domain laid the groundwork for aligning visual features with textual embeddings. For instance, [1] pioneered a contrastive learning approach

that effectively learned transferable visual representations from natural language supervision, enabling zero-shot image classification and retrieval based on text queries. The development of robust ranking and retrieval models is crucial for these capabilities, with related advancements being extensively studied in areas like long document and robust text retrieval [8,9]. Subsequent research focused on enhancing the capabilities of these models for more intricate interactions and few-shot learning. The Flamingo model, as presented in [2], demonstrated significant strides in enabling LVLMs to perform few-shot learning by effectively conditioning a pre-trained language model on visual inputs. Similarly, the paradigm of visual in-context learning has been specifically explored to empower models to learn new visual concepts on the fly from just a few examples [10]. Further expanding the scope, models like [11] introduced architectures designed for large-scale vision and language understanding, excelling in tasks that require deep semantic understanding of both modalities, from conventional question answering to creative generation tasks like sketch-based storytelling [12]. Similarly, [13] proposed a bootstrapping approach for language-image pre-training, aiming for unified vision-language understanding and generation, highlighting the importance of robust pre-training strategies. More recently, the focus has shifted towards making these large models more efficient, adaptable, and capable of following complex instructions. Works such as [14] explored methods for enhancing vision-language understanding by aligning visual encoders with powerful large language models, leveraging their reasoning capabilities. Recent efforts have pushed the boundaries even further by investigating how to achieve strong generalization from weaker models [15] and by rethinking core architectural assumptions, such as visual dependency in long-context reasoning, to handle more complex scenarios [16]. The concept of instruction tuning has also gained prominence, as exemplified by [17], which investigates how explicit textual instructions can guide LVLMs to perform a wide array of tasks more effectively. Furthermore, to address the computational challenges of fine-tuning these colossal models, research like [18] introduced efficient adapter-based methods, allowing for the fine-tuning of large language models for visual instruction following with significantly fewer trainable parameters. While these LVLMs exhibit impressive general reasoning, they often implicitly learn spatial relationships from vast data rather than possessing an explicit, structured understanding, which is a key motivation for our proposed MGS-RGT approach.

2.2 Spatial Reasoning

Spatial reasoning is a fundamental cognitive ability that enables humans and intelligent systems to understand, navigate, and interact with the physical world. In artificial intelligence, developing robust spatial reasoning capabilities is crucial for tasks ranging from robotic manipulation to complex visual question answering. Early work on computational spatial reasoning often focused on symbolic representations and qualitative spatial relations, aiming to model human-like common sense about space [19]. These methods provided foundational frameworks for describing topological, directional, and distance relationships between

objects without relying on precise metric coordinates. With the advent of deep learning, approaches to spatial reasoning have increasingly leveraged data-driven methods, particularly within the domain of computer vision. A significant line of work has concentrated on inferring spatial relationships directly from images and videos. For instance, scene graph generation models, as exemplified by [20], have become instrumental in explicitly representing objects and their pairwise relationships (including spatial ones) within a visual scene. These graphs provide a structured semantic understanding that can be consumed by downstream reasoning modules. The principle of leveraging structured knowledge graphs for complex reasoning has also been proven effective in natural language processing, for instance, in improving zero-shot cross-lingual question answering [21]. Beyond static images, reasoning about dynamic spatial changes in temporal sequences has also gained traction, with research like [19] exploring how models can learn to track and infer spatial transformations over time in videos. More recently, the integration of spatial reasoning with large vision-language models has become a key research frontier. While general LVLMs can implicitly learn some spatial cues from vast amounts of data, they often struggle with explicit, fine-grained spatial grounding and common-sense spatial reasoning [22]. This highlights a gap between general visual understanding and precise spatial intelligence. Furthermore, the challenges of spatial reasoning extend to embodied AI, where agents must understand their physical environment to plan and execute actions. Surveys like [23] emphasize the critical role of robust spatial reasoning for effective human-robot interaction and collaboration in real-world settings. Our work aims to bridge this gap by explicitly injecting hierarchical spatial knowledge into LVLMs, enabling more precise and explainable spatial reasoning.

3 Method

Our proposed approach, **Multi-Granularity Spatial-Relational Graph Transformer (MGS-RGT) Training**, enhances LVLMs with robust spatial reasoning capabilities through a novel two-stage learning paradigm. The core of our method lies in transforming a standard generative LVLM into a spatially-aware reasoning agent by meticulously integrating hierarchical spatial knowledge. While the ultimate goal is **generative** (e.g., producing spatially coherent descriptions or task plans), the learning process incorporates strong **discriminative** signals for spatial relationship prediction, ensuring that the model not only generates text but does so based on a deeply grounded understanding of the visual scene's geometry and topology.

3.1 Model Architecture

The MGS-RGT model is built upon a foundation of a pre-trained Large Vision-Language Model, which typically comprises a visual encoder and a language

model connected by a sophisticated cross-modal attention mechanism. This architecture is designed to seamlessly integrate visual perception with linguistic understanding, forming a cohesive system capable of complex multimodal reasoning.

Visual Encoder (E_V) The visual encoder, denoted as E_V, is responsible for extracting rich, spatially relevant features from an input image $I \in \mathbb{R}^{H \times W \times 3}$. We adopt a **Vision Transformer (ViT)** architecture for E_V, renowned for its ability to capture global dependencies across image regions. The input image is first meticulously divided into N_P non-overlapping patches $p_1, p_2, \ldots, p_{N_P}$, where each patch $p_i \in \mathbb{R}^{P_h \times P_w \times 3}$. These individual patches are then linearly embedded into a higher-dimensional space and combined with learnable **positional embeddings** to preserve spatial layout information, forming the initial sequence of visual tokens v_0. This sequence is then processed through a series of L_V transformer layers, each layer refining the contextual understanding of these visual tokens:

$$v_0 = [\text{LinearEmbed}(p_1) + \text{PosEmbed}_1, \ldots, \text{LinearEmbed}(p_{N_P}) + \text{PosEmbed}_{N_P}] \tag{1}$$

$$v_l = \text{TransformerLayer}_V(v_{l-1}) \quad \forall l \in \{1, \ldots, L_V\} \tag{2}$$

The output of the visual encoder is a set of spatially-rich visual features $F_V = v_{L_V} \in \mathbb{R}^{N_P \times D_V}$, where D_V is the dimension of the visual features, encoding not just what is present, but also where it is.

Language Model (M_L) The language model, denoted as M_L, serves as the linguistic backbone of our LVLM, handling the processing of textual inputs and the generation of coherent, contextually appropriate textual outputs. We employ a large transformer-based architecture, typically a decoder-only or an encoder-decoder model, which has demonstrated formidable capabilities in various natural language processing tasks. Given an input text sequence $T = \{t_1, t_2, \ldots, t_M\}$, the language model computes contextualized embeddings $h_k = \text{LMEmbed}(t_k)$ for each token. During the generative phase, M_L predicts the probability distribution over the next token, conditioning its prediction on the preceding tokens and the integrated visual features:

$$P(t_{k+1}|t_1, \ldots, t_k, F_V) = \text{Softmax}(\text{Linear}(M_L(h_k, F_V))) \tag{3}$$

This allows the model to generate responses that are not only grammatically correct but also semantically and spatially aligned with the visual content.

Cross-Modal Fusion Module (F_{CM}) The cross-modal fusion module is the crucial bridge connecting the visual and linguistic modalities. This module effectively integrates the visual features F_V from the encoder into the linguistic context maintained by the language model. We leverage advanced cross-attention

mechanisms, where visual features act as keys and values, and language embeddings act as queries (or vice-versa, depending on the specific fusion strategy). For a given language token embedding h_k, its fused representation $\tilde{h}_k$ is computed by attending over the visual features, dynamically selecting relevant visual information for each linguistic context:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{4}$$

$$\tilde{h}_k = \text{LayerNorm}(h_k + \text{MultiHeadAttention}(h_k, F_V, F_V)) \tag{5}$$

This refined, fused representation $\tilde{h}_k$ is then passed to the subsequent layers of the language model, enabling deep, visually grounded text generation. The effectiveness of this module is paramount for the LVLM to truly "see" and "understand" the visual scene as it processes and generates language.

3.2 Two-Stage Learning Paradigm: MGS-RGT Training

Our MGS-RGT training strategy is a meticulously designed, sequential two-stage process. This paradigm ensures that the LVLM not only understands semantic content but also develops a profound, explicit understanding of hierarchical spatial relationships.

Stage 1: Hierarchical Spatial Graph Prediction (HSGP) The foundational objective of the HSGP pre-training stage is to imbue the visual encoder E_V with an explicit, structured understanding of multi-granularity spatial relationships. For each input image I, we automatically construct a set of hierarchical spatial graphs $G = \{G_{fine}, G_{obj}, G_{scene}\}$, representing spatial information at different levels of abstraction.

Spatial Graph Construction Process: The construction of these graphs is a critical preliminary step, providing the ground truth for our visual encoder's learning task.

- **Fine-grained Graph (G_{fine}):** This graph captures minute, local spatial relationships. Nodes N_{fine} correspond to small, semantically meaningful image regions, typically obtained through superpixel segmentation or grid-based partitioning. Edges E_{fine} represent fine-grained relationships such as adjacency, connectivity, and precise local relative positions (e.g., "directly above," "touching left"). For any two adjacent or interacting superpixels $s_i, s_j \in N_{fine}$, an edge $(s_i, s_j) \in E_{fine}$ is formed with an associated relation type $r_{ij} \in \{\text{adjacent, top-left-of, below}, \ldots\}$.
- **Object-level Graph (G_{obj}):** Building upon robust object detection mechanisms, this graph models spatial relationships between identified objects $O = \{o_1, \ldots, o_K\}$ within the scene. Nodes N_{obj} represent these detected objects, uniquely identified by their bounding boxes and semantic labels.

Edges E_{obj} denote a rich set of spatial relations such as "left of," "right of," "above," "below," "contains," "inside," "overlaps with," "near," "far," and even implicit functional relationships derived from common sense knowledge (e.g., "cup on table" implying support). The determination of a relationship $R(o_i, o_j)$ between objects o_i and o_j is based on their geometric properties, specifically their bounding box coordinates (x_i, y_i, w_i, h_i) and (x_j, y_j, w_j, h_j). For instance, "left of" could be formally defined as $x_i + w_i < x_j - \delta$, where δ is a small threshold.

– **Scene-level Graph (G_{scene}):** This graph captures overarching global spatial layouts and structural elements of the entire scene. Nodes N_{scene} might represent dominant scene components (e.g., "wall," "floor," "ceiling," "furniture cluster," "open space"), and edges E_{scene} describe their high-level topological relationships (e.g., "floor supports table," "door leads to hallway"). This can be extracted using advanced scene parsing techniques, 3D scene graph generation methods, or coarse semantic segmentation combined with layout estimation.

HSGP Task and Loss Function: To enable the visual encoder to predict these complex spatial relationships, a dedicated **Graph Prediction Head (P_G)** is attached to the output of E_V. For each output visual token $f_i \in F_V$, P_G predicts the likelihood of various spatial relationships originating from or terminating at the image region corresponding to f_i, across all granularities. Let A_G be the ground-truth adjacency matrix (or a collection of tensors, each representing a specific relation type) for a constructed spatial graph. The predicted graph structure $\hat{A}_G$ is derived from the transformed visual features F_V. The HSGP loss L_{HSGP} is a composite loss function, meticulously designed to cover and combine the prediction accuracies across all three graph granularities:

$$L_{HSGP} = \sum_{g \in \{\text{fine, obj, scene}\}} \mathcal{L}_{CE}(P_G(F_V^{(g)}), A_G^{(g)}) \tag{6}$$

Here, $\mathcal{L}_{CE}$ denotes the standard cross-entropy loss, and $F_V^{(g)}$ represents the subset or transformation of visual features most relevant to predicting the relations within graph g. This multi-faceted loss stringently pushes the visual encoder to generate features that are inherently rich in multi-level spatial relationship information, far beyond simple object presence.

Stage 2: Spatially-Grounded Language Generation (SGLG) Following the comprehensive pre-training of the visual encoder, the entire LVLM, now equipped with its spatially-aware E_V, undergoes fine-tuning on a diverse and challenging set of Spatially-Grounded Language Generation (SGLG) tasks. This stage ensures that the model can effectively leverage the learned spatial knowledge to generate natural language responses that precisely describe, explain, or reason about spatial relationships within an image, guided by specific linguistic prompts.

SGLG Task Formulation: In the SGLG task, the model is presented with an image I and a textual prompt Q, and its objective is to generate an accurate and spatially coherent natural language response A. For tasks demanding explicit spatial reasoning, Q is carefully crafted to contain specific spatial queries (e.g., "Where is the bottle relative to the book?"). The dataset for SGLG consists of meticulously curated triplets (I, Q, A), where A is the ground-truth spatial description, a direct answer to the spatial question, or a complete reasoning chain.

Chain-of-Thought (CoT) Integration: To foster transparent and interpretable spatial reasoning, we rigorously integrate **Chain-of-Thought (CoT)** prompting during the SGLG fine-tuning. The model is specifically trained to first generate an explicit, step-by-step "thought process" (T_{CoT}) that articulates its spatial inferences, before producing the final succinct answer (A_{final}). This transforms the generation task from simply A to $T_{CoT} \oplus A_{final}$ (where $\oplus$ denotes concatenation). The input prompt Q is potentially augmented to Q' to explicitly solicit this thought process. The model's generation probability is then maximized over this entire sequence:

$$P(T_{CoT}, A_{final}|I, Q') = \prod_{k=1}^{|T_{CoT} \oplus A_{final}|} P(\text{token}_k|\text{token}_{<k}, F_V(I), Q') \tag{7}$$

The standard language modeling loss (cross-entropy) is rigorously applied over the entire generated sequence ($T_{CoT} \oplus A_{final}$), ensuring both the correctness of the final answer and the logical coherence of the intermediate reasoning steps.

Overall Loss Function for Fine-tuning: The total loss for fine-tuning the LVLM in Stage 2, denoted as L_{SGLG}, is primarily a cross-entropy loss that maximizes the likelihood of generating the target text sequence given the visual input and prompt:

$$L_{SGLG} = -\frac{1}{N_{\text{tokens}}} \sum_{i=1}^{N_{\text{tokens}}} \log P(\text{token}_i|\text{token}_{<i}, F_V(I), Q) \tag{8}$$

where N_{tokens} is the total length of the target sequence ($T_{CoT} \oplus A_{final}$). This loss vigorously encourages the model to generate accurate, spatially coherent descriptions and sophisticated reasoning. To prevent any potential catastrophic forgetting of the finely tuned spatial graph representations learned in Stage 1, we judiciously incorporate a small component of the L_{HSGP} loss during this fine-tuning stage. This ensures a continuous reinforcement of the visual encoder's spatial understanding:

$$L_{\text{Total}} = L_{SGLG} + \lambda L_{HSGP} \tag{9}$$

Here, λ is a carefully chosen hyperparameter that balances the primary objective of language generation with the crucial need to maintain and refine the

model's underlying spatial knowledge. This dual-objective optimization ensures that the visual encoder provides consistently richer, spatially-structured features to the cross-modal fusion module, empowering the language model to perform exceptionally accurate, nuanced, and explainable spatial reasoning.

4 Experiments

To rigorously evaluate the efficacy of our proposed **Multi-Granularity Spatial-Relational Graph Transformer (MGS-RGT) training** approach, we conducted a series of comprehensive experiments. Our primary objective was to demonstrate that MGS-RGT significantly enhances the spatial reasoning capabilities of Large Vision-Language Models (LVLMs) compared to existing state-of-the-art methods. We designed experiments to assess performance across various facets of spatial understanding, including fine-grained object localization, comprehension of complex spatial relationships, and execution planning for multi-step embodied tasks.

4.1 Experimental Setup

Dataset Our experiments primarily utilized two main datasets to provide a holistic evaluation. The first is a large-scale public benchmark for Visual Question Answering (VQA), such as the **VQAv2 dataset**, which includes a diverse range of images and associated questions, serving as a general evaluation of LVLM capabilities. The second, and crucial for assessing fine-grained and hierarchical spatial reasoning, is our meticulously constructed **Spatially-Grounded Interaction Dataset (SGID)**. The SGID comprises over 100,000 image-text pairs, each explicitly annotated with hierarchical spatial relationships (at fine-grained, object-level, and scene-level granularities) and accompanied by **Chain-of-Thought (CoT)** reasoning paths. This dataset was further augmented with simulated robotic interaction logs from environments like **AI2-THOR**, providing dynamic spatial information and task execution sequences crucial for embodied reasoning tasks. For training our models, we allocated 80% of the combined dataset, reserving 10% for validation and the remaining 10% for final testing to ensure an unbiased performance assessment.

Baselines We established a clear performance benchmark by comparing MGS-RGT against several prominent baseline LVLMs, each representing different strategies for integrating visual and linguistic information. This ensures a comprehensive and fair comparison.

- **Vanilla Large Vision-Language Model (LVLM-Base):** This baseline refers to a standard LVLM architecture, such as a pre-trained ViT-L/14 visual encoder coupled with a 7B-parameter Transformer-decoder language model, fine-tuned on general large-scale image-text pairs. It lacks any specific architectural or training enhancements for explicit spatial reasoning beyond implicit learning from vast data.

- **LVLM with Enhanced Positional Encoding (LVLM-EPE):** This baseline extends the LVLM-Base by incorporating more sophisticated absolute and relative positional encodings for visual tokens, such as those derived from 2D pixel coordinates or attention bias mechanisms. The aim here is to provide the model with better inherent spatial awareness without relying on external graph representations.
- **LVLM with Integrated Scene Graph Features (LVLM-SGF):** This method represents a strong existing approach for incorporating structured spatial information. It utilizes an off-the-shelf, high-performing scene graph generation model (e.g., based on GATs or Transformer-based scene parsers) to extract object-level relationships and attributes. These extracted scene graph features are then explicitly fed as additional tokens or cross-attention inputs into the LVLM during training.

All baseline models were fine-tuned on the same combined VQAv2 and SGID dataset splits, with their hyperparameters meticulously optimized through grid search to ensure their best possible performance under the given setup. For baselines not natively supporting CoT, training was conducted without the explicit CoT paths.

Evaluation Metrics To provide a multi-faceted and comprehensive evaluation of spatial reasoning, we employed a diverse set of quantitative metrics:

- **Spatial VQA Accuracy (%):** This metric quantifies the exact match accuracy on a subset of the VQAv2 test set specifically curated for spatial reasoning questions, as well as dedicated spatial queries within the SGID. It measures the ability to provide precise answers to "where," "what's next to," or "how far" questions.
- **Relationship Prediction F1-Score:** This metric rigorously evaluates the model's capability to correctly identify and classify fine-grained and object-level spatial relationships (e.g., "left-of," "on-top-of," "contained-in") within a scene. It is calculated by comparing predicted relationships against ground-truth annotations in the SGID.
- **Task Success Rate (%):** For embodied tasks (simulated within the AI2-THOR environment as part of SGID), this metric assesses the percentage of tasks correctly completed based on the model's spatial understanding, multi-step planning, and execution robustness.
- **Spatial Coherence Score (SCS):** A newly introduced objective metric that quantifies the logical consistency, factual correctness, and completeness of generated spatial descriptions and reasoning chains. SCS is computed via a sophisticated combination of automated linguistic checks (e.g., grammatical correctness, absence of contradictions) and a semantic similarity measure against expert-annotated ground-truth spatial narratives.

4.2 Quantitative Results

Our experimental results unequivocally demonstrate the superior performance of MGS-RGT across all assessed spatial reasoning benchmarks. The strategic in-

tegration of hierarchical spatial graph learning, coupled with explicit CoT training, significantly enhances the LVLM's ability to understand, reason about, and articulate complex spatial relationships.

Main Comparison Results Table 1 comprehensively summarizes the quantitative performance of MGS-RGT against the selected baselines on various spatial reasoning tasks.

Table 1. Comparison of Spatial Reasoning Performance across Different LVLM Models

Model	Spatial VQA Acc.	Rel. Pred. F1	Task Success Rate	Spatial Coherence Score
LVLM-Base	68.2	0.52	55.1	0.65
LVLM-EPE	71.5	0.58	58.7	0.70
LVLM-SGF	75.3	0.67	64.9	0.76
MGS-RGT	**81.7**	**0.78**	**75.2**	**0.88**

As meticulously detailed in Table 1, MGS-RGT consistently and substantially outperforms all established baselines across every spatial reasoning metric. Notably, our method achieves an impressive **81.7% Spatial VQA Accuracy**, representing a significant improvement of 6.4 percentage points over the strongest baseline (LVLM-SGF at 75.3%). The F1-score for Relationship Prediction also shows a substantial gain of 0.11 points compared to LVLM-SGF, unequivocally highlighting MGS-RGT's enhanced capability in discerning accurate and fine-grained spatial relationships. Furthermore, for the challenging embodied tasks, our model's **Task Success Rate of 75.2%** demonstrates a more robust and actionable understanding of spatial configurations required for successful execution in dynamic environments. The consistently high Spatial Coherence Score underscores the exceptional quality and logical consistency of the spatial descriptions and reasoning generated by MGS-RGT. These results unequivocally affirm the efficacy of our multi-granularity spatial graph learning approach.

4.3 Ablation Study

To rigorously validate the individual contributions and synergistic effects of the core components within our MGS-RGT framework, we conducted an extensive ablation study. This systematic analysis helps isolate the precise impact of the **Hierarchical Spatial Graph Prediction (HSGP)** pre-training stage and the **Chain-of-Thought (CoT)** integration during fine-tuning.

Table 2 presents the illuminating results of our ablation study. Removing the **Hierarchical Spatial Graph Prediction (HSGP)** pre-training stage leads to a substantial degradation in performance across all metrics, with Spatial VQA Accuracy plummeting from 81.7% to 74.5%. This striking decline unequivocally demonstrates the critical and indispensable role of explicitly pre-training the visual encoder on multi-granularity spatial relationships. Without HSGP, the visual encoder demonstrably struggles to provide the granular and structured

Table 2. Ablation Study on Key Components of MGS-RGT

Model Variant	Spatial VQA Acc.	Rel. Pred. F1	Task Success Rate	Spatial Coherence Score
w/o HSGP Pre-training	74.5	0.65	63.8	0.75
w/o Chain-of-Thought (CoT) Integration	78.9	0.73	70.1	0.81
MGS-RGT	**81.7**	**0.78**	**75.2**	**0.88**

spatial cues that are essential for the LVLM to perform accurate and nuanced spatial reasoning. Similarly, ablating the **Chain-of-Thought (CoT)** integration, while yielding a slightly less drastic performance reduction than HSGP, still results in a noticeable drop (e.g., Spatial VQA Accuracy from 81.7% to 78.9%). This confirms that guiding the model to articulate its reasoning process step-by-step through CoT training leads to more robust, accurate, and arguably more transparent spatial understanding, likely by fostering a deeper internal representation of spatial logic and causal inference. The full MGS-RGT model consistently achieves the highest performance across all metrics, thereby validating the powerful synergistic benefits derived from combining both HSGP pre-training and CoT integration.

4.4 Qualitative Analysis and Human Evaluation

Beyond rigorous quantitative metrics, we conducted a meticulous qualitative analysis and a dedicated human evaluation to assess the perceptual quality, interpretability, and practical utility of MGS-RGT's generated spatial descriptions and reasoning processes from a human perspective.

Qualitative Observations Our in-depth qualitative analysis revealed that MGS-RGT generates spatial descriptions that are remarkably precise, highly detailed, and perceptually intuitive. For instance, when presented with an image depicting a book partially hidden behind a laptop on a desk, a typical baseline model might generate a vague statement like "The book is somewhere on the desk." In stark contrast, MGS-RGT consistently produces a more accurate and nuanced description such as "The book is positioned directly behind the laptop, slightly to its right, resting on the polished wooden desk surface." This exceptional level of detail, including understanding occlusions, precise relative positioning, and surface interactions, strongly suggests a superior and more granular comprehension of the scene's spatial geometry. Furthermore, the CoT reasoning chains explicitly generated by MGS-RGT frequently mirrored logical human thought processes. These chains often sequentially identified key objects, meticulously assessed their inter-relationships (e.g., "First, identify the red sphere. Second, locate the blue cube. Third, determine the directional relationship and approximate distance."), and then coherently synthesized this information into a final, actionable answer.

Human Evaluation To provide an unbiased and user-centric assessment of the generated outputs, we recruited a diverse panel of 10 human annotators. These

annotators were presented with a randomized set of image-question pairs sourced from the SGID, along with the corresponding answers generated by our MGS-RGT model and the best-performing baseline (LVLM with Integrated Scene Graph Features). They were instructed to rate each generated answer based on three critical criteria, using a Likert scale ranging from 1 (Very Poor) to 5 (Excellent): **Accuracy** (factual correctness of spatial information), **Completeness of Spatial Detail** (how thoroughly and precisely the spatial relationships are described), and **Fluency/Naturalness of Description** (the linguistic quality and readability of the response). The average scores for each model across these criteria are presented in Table 3.

Table 3. Human Evaluation of Generated Spatial Descriptions and Reasoning

Model	**Accuracy (1-5)**	**Completeness (1-5)**	**Fluency (1-5)**
LVLM-SGF	3.8	3.5	4.1
MGS-RGT	**4.6**	**4.5**	**4.4**

The compelling results from the human evaluation, as clearly depicted in Table 3, strongly corroborate our quantitative findings. MGS-RGT consistently received significantly higher average scores across all three evaluation criteria. Human annotators rated MGS-RGT's outputs as perceptually more **accurate** (average score of 4.6 compared to 3.8 for LVLM-SGF) and, crucially, demonstrating substantially superior **completeness of spatial detail** (average score of 4.5 compared to 3.5). The notable improvement in fluency (4.4 vs. 4.1) further indicates that the enriched and structured spatial understanding directly translates into more natural, informative, and human-like language generation. These comprehensive human-centric evaluations provide compelling evidence of the tangible benefits of our MGS-RGT approach in delivering more human-aligned and precise spatial reasoning capabilities to LVLMs, moving closer to true human-level spatial intelligence.

4.5 Further Analysis of Spatial Understanding

Beyond the core metrics, we conducted additional analyses to delve deeper into specific aspects of MGS-RGT's enhanced spatial understanding. These investigations aim to highlight how our method addresses subtle yet critical challenges in spatial reasoning.

Robustness to Visual Perturbations Real-world scenarios often involve visual ambiguities, occlusions, or slight viewpoint changes. We tested the robustness of MGS-RGT by introducing controlled visual perturbations (e.g., slight blurring, minor occlusions of objects) to the test images in the SGID and evaluating the models' ability to maintain spatial reasoning accuracy. Table 4 summarizes these results.

Table 4. Robustness to Visual Perturbations (Spatial VQA Accuracy %)

Model	No Perturbation	Minor Blurring	Partial Occlusion
LVLM-Base	68.2	62.5	58.1
LVLM-EPE	71.5	66.8	61.2
LVLM-SGF	75.3	69.1	65.5
MGS-RGT	**81.7**	**77.2**	**73.9**

As presented in Table 4, MGS-RGT demonstrates significantly higher robustness to visual perturbations. While all models show some performance degradation under challenging visual conditions, MGS-RGT maintains a notably higher accuracy. For instance, under partial occlusion, MGS-RGT's Spatial VQA Accuracy only drops by approximately 7.8 percentage points from its unperturbed performance, whereas LVLM-SGF drops by 9.8 percentage points. This resilience suggests that our hierarchical spatial graph representations provide a more stable and abstract understanding of spatial relationships, making them less susceptible to superficial visual distortions. The explicit encoding of global and local spatial structures helps the model infer relationships even when parts of objects or scenes are obscured.

Generalization to Novel Spatial Configurations A critical test for any spatial reasoning system is its ability to generalize to novel spatial configurations not explicitly seen during training. We curated a specific subset of the SGID test set containing object arrangements and spatial relationships that are synthetically generated and structurally distinct from the training data, while maintaining realism. The results are shown in Table 5.

Table 5. Generalization to Novel Spatial Configurations (Spatial VQA Accuracy %)

Model	Novel Config. Acc. (%)
LVLM-Base	55.0
LVLM-EPE	58.5
LVLM-SGF	63.2
MGS-RGT	**72.8**

Table 5 highlights MGS-RGT's superior generalization capabilities. Our method achieves a remarkable **72.8% accuracy** on novel spatial configurations, significantly outperforming LVLM-SGF (63.2%). This strong performance indicates that the HSGP pre-training enables the visual encoder to learn fundamental principles of spatial composition and decomposition, rather than simply memorizing specific arrangements. The hierarchical graph structures allow MGS-RGT to infer how unseen objects or layouts relate in space by leveraging learned pat-

terns of spatial grammar, making it more adaptable to new environments and scenarios.

Performance on Multi-Step Embodied Reasoning For real-world robotic applications, the ability to plan and execute multi-step tasks based on dynamic spatial understanding is crucial. We further analyzed the Task Success Rate on complex, multi-step embodied tasks within the AI2-THOR environment, where success depends on correctly inferring object affordances, predicting spatial changes, and planning sequential actions. Table 6 details the results for tasks requiring 2, 3, or 4 sequential spatial reasoning steps.

Table 6. Task Success Rate on Multi-Step Embodied Reasoning Tasks (%)

Model	2-Step Tasks	3-Step Tasks	4-Step Tasks
LVLM-Base	65.2	48.7	35.1
LVLM-EPE	68.5	53.9	39.5
LVLM-SGF	74.1	62.8	49.3
MGS-RGT	**85.0**	**78.2**	**68.5**

The results in Table 6 unequivocally demonstrate MGS-RGT's exceptional ability in multi-step embodied reasoning. As the complexity of tasks increases (from 2-step to 4-step), the performance gap between MGS-RGT and baselines widens significantly. For 4-step tasks, our method achieves a success rate of **68.5%**, which is a substantial 19.2 percentage points higher than the best baseline (LVLM-SGF at 49.3%). This superior performance is a direct consequence of MGS-RGT's explicit spatial reasoning capabilities, particularly its CoT integration, which allows the model to logically decompose complex spatial goals into manageable sub-goals and predict the spatial consequences of its actions. This level of robust, sequential spatial understanding is paramount for intelligent embodied agents.

5 Conclusion

In this paper, we introduced **MGS-RGT Training**, a novel and effective two-stage paradigm specifically engineered to elevate the spatial reasoning capabilities of Large Vision-Language Models. We identified that existing LVLMs, despite their general prowess, often lack a granular and explicit understanding of the complex spatial relationships inherent in the visual world. Our method directly addresses this limitation by first equipping the visual encoder with a profound sense of multi-level spatial structure through **Hierarchical Spatial Graph Prediction (HSGP)** pre-training. This initial stage ensures that the visual representations are rich with fine-grained, object-level, and scene-level spatial information. Subsequently, the entire LVLM is meticulously fine-tuned for

Spatially-Grounded Language Generation (SGLG), critically enhanced by **Chain-of-Thought (CoT)** integration. This not only pushes the model to produce accurate spatial descriptions but also fosters a transparent and logical reasoning process, enabling it to "think through" spatial problems.

Our extensive experimental results rigorously validate the efficacy of MGS-RGT. The quantitative comparisons against strong baselines unequivocally demonstrated our model's superior performance across a diverse range of spatial tasks, including VQA, relationship prediction, and complex embodied task execution. The consistent and significant improvements observed across these metrics underscore the profound impact of our proposed learning strategy. Furthermore, our detailed ablation studies provided clear evidence of the individual and synergistic contributions of both the HSGP pre-training and CoT integration, confirming that these components are indispensable for achieving advanced spatial intelligence. Beyond numerical metrics, qualitative analysis revealed MGS-RGT's ability to generate perceptually accurate and highly detailed spatial narratives, a testament to its deeper understanding of visual scenes. Complementary human evaluations further reinforced these findings, with human annotators consistently rating MGS-RGT's outputs as more accurate, complete, and natural.

Looking forward, this work opens several exciting avenues for future research. One direction involves exploring the application of MGS-RGT to more dynamic and interactive 3D environments, such as those found in virtual reality or advanced robotics, where real-time spatial understanding is paramount. Another promising path is to investigate how explicit spatial uncertainty can be modeled and conveyed in the generated language, reflecting the inherent ambiguities of perception. Finally, extending the hierarchical spatial graphs to incorporate temporal relationships could enable LVLMs to reason about changes in spatial configurations over time, leading to more comprehensive understanding of complex events. We believe MGS-RGT represents a significant step towards building truly spatially intelligent LVLMs, capable of robustly perceiving, reasoning about, and interacting within our intricate physical world.

References

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
2. Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Bel-

grave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html
3. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
4. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L.J., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 14455–14465. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01370, https://doi.org/10.1109/CVPR52733.2024.01370
5. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
6. Narayanan, A., Rao, A., Prasad, A., Subramanyam, N.: VQA as a factoid question answering problem: A novel approach for knowledge-aware and explainable visual question answering. Image Vis. Comput. **116**, 104328 (2021). https://doi.org/10.1016/J.IMAVIS.2021.104328, https://doi.org/10.1016/j.imavis.2021.104328
7. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(2), 230–244 (2022)
8. Zhou, Y., Shen, T., Geng, X., Tao, C., Shen, J., Long, G., Xu, C., Jiang, D.: Fine-grained distillation for long document retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19732–19740 (2024)
9. Zhou, Y., Shen, T., Geng, X., Tao, C., Xu, C., Long, G., Jiao, B., Jiang, D.: Towards robust ranker for text retrieval. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 5387–5401 (2023)
10. Zhou, Y., Li, X., Wang, Q., Shen, J.: Visual in-context learning for large vision-language models. In: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. pp. 15890–15902. Association for Computational Linguistics (2024)
11. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
12. Zhou, Y.: Sketch storytelling. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4748–4752. IEEE (2022)
13. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Mary-

land, USA. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022), https://proceedings.mlr.press/v162/li22n.html
14. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
15. Zhou, Y., Shen, J., Cheng, Y.: Weak to strong generalization for large language models with multi-capabilities. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=N1vYivuSKq
16. Zhou, Y., Rao, Z., Wan, J., Shen, J.: Rethinking visual dependency in long-context reasoning for large vision-language models. arXiv preprint arXiv:2410.19732 (2024)
17. Du, Y., Guo, H., Zhou, K., Zhao, W.X., Wang, J., Wang, C., Cai, M., Song, R., Wen, J.: What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025. pp. 8197–8214. Association for Computational Linguistics (2025), https://aclanthology.org/2025.coling-main.546/
18. Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
19. Ziaeetabar, F.: Spatio-temporal reasoning for semantic scene understanding and its application in recognition and prediction of manipulation actions in image sequences. Ph.D. thesis, Dissertation, Göttingen, Georg-August Universität, 2019 (2020)
20. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3097–3106. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.330, https://doi.org/10.1109/CVPR.2017.330
21. Zhou, Y., Geng, X., Shen, T., Zhang, W., Jiang, D.: Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5822–5834 (2021)
22. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
23. Sisbot, E.A., Marin, L.F., Alami, R.: Spatial reasoning for human robot interaction. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2281–2287. IEEE (2007)