Spatial Task-Explicity Matters in Prompting Large Multimodal Models for Spatial Planning

Ivan Majic
ivan.majic@univie.ac.at
University of Vienna
Vienna, Austria

Zhangyu Wang
zhangyuwang@ucsb.edu
University of California Santa Barbara
Santa Barbara, CA, United States

Krzysztof Janowicz
krzysztof.janowicz@univie.ac.at
University of Vienna
Vienna, Austria

Mina Karimi
mina.karimi@univie.ac.at
University of Vienna
Vienna, Austria

Abstract

The advance in large multimodal models (LMMs) gives rise to autonomous bots that perform complex tasks using human-like reasoning on their own. The ability of large models to understand spatial relations and perform spatial operations, however, is known to be limited. This gap hinders the development of autonomous GIS analysts, travel planning assistants, and other possibilities of spatial bots. In this paper, we explore the impact of modality on the performance of LMMs in spatial planning tasks - specifically, retrieving a target brick by first removing all other bricks on top of it. Experiments demonstrate that what matters is not only the modality of the prompts (text or image), but also how informative the spatial descriptions are for the LMMs to complete the task. We propose novel concepts of *task-implicit* and *task-explicit* spatial descriptions to qualitatively quantify the task-specific informativity of prompts. Furthermore, we develop simple techniques to increase the spatial task-explicity of image prompts, and the accuracy of spatial planning increases from 26% to 100% accordingly.

CCS Concepts

• **Theory of computation** → *Semantics and reasoning*; • **Computing methodologies** → **Planning and scheduling**; **Spatial and physical reasoning**.

Keywords

task-explicity, multi-modal prompts, large multimodal models (LMM), spatial reasoning, GeoAI

ACM Reference Format:
Ivan Majic, Zhangyu Wang, Krzysztof Janowicz, and Mina Karimi. 2024. Spatial Task-Explicity Matters in Prompting Large Multimodal Models for Spatial Planning. In *7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI'24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3687123.3698293

GeoAI'24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1176-3/24/10
https://doi.org/10.1145/3687123.3698293

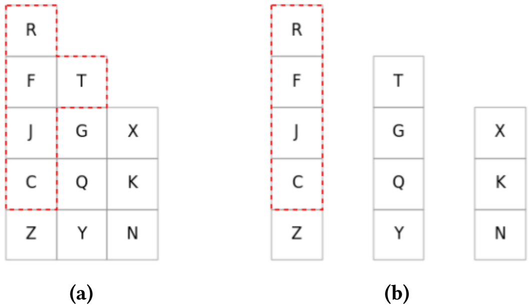
1 Introduction

The GIS community has long been speculating about an autonomous GIS analyst (a *GeoMachina*) that can perform simple spatial data science tasks entirely by itself [10]. This goal becomes increasingly tangible given the rapid advance in foundation models, especially large multimodal models [12], since spatial data can be expressed in multiple modality (e.g., natural language description vs an image depiction of a route). Instead of training task-specific models from scratch, LMMs allow researchers to build intelligent bots by prompting and decomposing complex tasks into steps via Chain-of-Thought (CoT) reasoning [16].

Conventionally, prompts for spatial reasoning tasks consist of images that describe the spatial environments and texts that describe the goal of the reasoning [1]. For example, one may present a photo of a living room to an LMM and ask it to plan a route to get a bottle of water in the corner. Whereas this setting of prompt modality seems most intuitive, researchers find that text-only prompts also work well in certain tasks [9, 11]. It is yet unclear to what extent the modality of prompts affects the performance of spatial reasoning, and one step further, which modality suits which spatial reasoning task best. This paper will make some explorations.

We focus on a specific spatial reasoning task: planning the order of removal of bricks in a brick world, where a brick can be removed only after all other bricks that are placed on top of it have been removed first [9]. This is a simple case of the conceptual abstraction of many real-world applications, where an AI system needs to find a sequence of locations in space (i.e., bricks) given a set of rules (i.e., the criteria for blocking, which in our case is "there exists a block on top"). We generate prompts for both 1D (i.e., one column) and 2D (i.e., multiple columns) brick arrangements in both textual and image modalities and use them to evaluate the planning accuracy of several AI systems (e.g., GPT-4o).

One issue we noticed in our preliminary experiments is that 2D image-based prompts often show a horizontal shift in their solutions where the AI system falsely recognizes bricks from an adjacent column as being on top of the target brick; see Figure 1. This phenomenon reduces the planning accuracy of image prompts to be anti-intuitively lower than text prompts. We hypothesize that this is because the text prompts, which use relational expressions such as *on top of* and *to the left of*, are spatial descriptions **explicitly** related to the solution of the planning task, whereas for image prompts the AI models need to **implicitly** infer the touching relations from



(a) (b)

Figure 1: Chatgpt-4o prediction of which bricks need to be removed (outlined in red) to reach brick Z. To reach any brick, all bricks on top of it must be removed first. When the input image has no spacing between columns (a), the system predicts that brick T has to be removed which is incorrect. When the spacing between columns is introduced (b), the system correctly predicts that bricks R, F, J, and C need to be removed to retrieve brick Z.

raw pixels. We propose a pair of concepts – *task-implicit* and *task-explicit* – to qualitatively quantify how straightforward and clear the spatial descriptions of prompts are for solving the task. For simplicity, we say a prompt is spatially task-implicit/task-explicit if its spatial description is task-implicit/task-explicit.

As the wording suggests, whether the a prompt is spatially task-explicit depends on the properties of the task. For example, in our block removal task, the most useful information is the vertical touching relations; thus the spatially task-explicit prompts are those who represent vertical touching relations as clear as possible. To verify the usefulness of our proposed concepts, we conduct two sets of experiments: (1) we generate spatially task-implicit text prompts by only describing the raw coordinates of bricks, and the planning accuracy drops significantly as expected; (2) we generate spatially task-explicit image prompts by adding horizontal spacing between brick columns, i.e., giving hints to the LMMs that vertical relations are more important, and the planning accuracy improves significantly as expected. We further investigate how different spacing strategies affect the planning accuracy. Experiments demonstrate that the larger the horizontal spacing is than the vertical spacing, the higher the planning accuracy, supporting our argument that hints of focusing on vertical relations make image prompts more spatially task-explicit, resulting in better performance.

The research contributions of this paper are as follows:

- We extend the existing study of spatial reasoning of LLM from [9] to include image-based prompts;
- We find that 2D spatial planning is very different from 1D in both text and image modalities, and identify the issue of column shift in the brick removal sequence;
- Propose novel theoretical concepts of *task-implicit* and *task-explicit* spatial descriptions to explain this phenomenon;
- For the vertical brick removal task, we demonstrate in our ablation studies that the key is to let the model focus on the vertical touching relations.

2 Related Works

Spatial reasoning and task planning have long been studied in GIScience and AI, with various approaches developed to enable autonomous spatial decision-making. Early studies explored symbolic reasoning and logic-based methods to automate spatial analysis, focusing on rules and relationships between spatial entities [5, 6]. With advancements in deep learning and computer vision, the integration of multimodal data, such as images and texts, has gained traction for complex spatial tasks, including object recognition, navigation, and path planning [2, 8]. Recent research highlights the potential of LMMs to enhance spatial reasoning capabilities, leveraging textual and visual prompts to guide AI behavior [13, 15].

The emergence of large multimodal models (LMMs) and their application in spatial reasoning tasks have gained significant attention recently. A key study by Hu et al. (2024) introduced Chain-of-Symbol Prompting, which simplifies user interactions with LLMs by transitioning from chain-of-thought to chain-of-symbol reasoning. This approach reduces token usage and improves performance in spatial reasoning tasks, particularly in textual prompts that utilize topological descriptions of spatial relations [9]. While their work explored various spatial reasoning tasks, our study extends their approach by incorporating metrical descriptions (using coordinates) and image-based prompts, specifically within the context of brick-world examples in 1D and 2D setups.

Gao et al. [7] further explored the potential of LLMs in representing textual descriptions of geometries and spatial relations using Well-Known Text (WKT) formats to assess how models interpret and reason about spatial objects from textual inputs. However, their focus was limited to earlier models like GPT-3, whereas our study not only includes GPT-3.5 but also examines the performance of newer models like GPT-4. This broader evaluation provides deeper insights into how different modalities and model versions affect spatial reasoning accuracy by extending the evaluation to include both metrical and image-based prompts.

Cohen's recent foundational contributions to spatial reasoning, particularly involving RCC8 spatial relations, underscore the growing interest in integrating spatial resoning with LLMs [3]. Although our study does not directly address RCC8 relations, Cohen's work highlights critical considerations in spatial reasoning, such as the ability of models to understand and manipulate spatial relationships, which are crucial for spatial planning tasks like those addressed in our research.

Recent studies by [18] on spatial foundation models and spatial embeddings represents a significant advancement in creating domain-specific representations that enhance LLM performance in spatial contexts. Their research focuses on pre-training and fine-tuning LLMs with spatial data to improve spatial reasoning capabilities. In contrast, our approach evaluates general-purpose LMMs without specific spatial adaptations, providing a baseline for understanding how these models perform spatial tasks without specialized training.

Additionally, from a computer vision perspective, spatial reasoning has been a key focus, especially in robotics and autonomous systems. For example, Zhangyu et al. (2024) [17] explored how computer vision models integrate spatial reasoning in robotics, using visual data to navigate and manipulate environments. Their work demonstrated that image-based prompts, such as those used in our study, can effectively guide spatial tasks, such as object retrieval or navigation. This research is particularly relevant to our investigation of how visual input impacts LLM performance in spatial planning tasks, reinforcing the importance of evaluating multimodal approaches.

These studies collectively form a foundation for understanding the evolving landscape of spatial reasoning with LLMs and computer vision techniques, highlighting the need for comprehensive evaluations across multiple input modalities and models to better understand their applicability and limitations in complex spatial tasks.

However, existing studies predominantly focus on simple environments or specific applications, with limited exploration into the comparative performance of different prompt modalities. Additionally, previous works often overlook the nuanced interactions between task-implicit and task-specific information in spatial reasoning, an aspect that becomes crucial when scaling from controlled, single-dimensional tasks to more complex, multi-dimensional environments [4, 14]. This gap underscores the need for a systematic examination of how LMMs process spatial prompts and how prompt design can impact task accuracy, particularly in scenarios where precise spatial relationships are critical.

3 Methodology

To test the effect of the input modality on the correctness of spatial reasoning of LMMs, we propose a methodology that is based on the brick world examples from [9], but extends further. This section presents the details of the brick removal task, the data generation process, the prompt types and strategies used, and the evaluation metrics. All code and data relating to our methodology and results are available on GitHub[1].

3.1 Task definition

The task is to retrieve a specified brick from a defined arrangement of bricks. Bricks can be arranged in 1D where they are stacked on top of each other, or in 2D where they are stacked both vertically, i.e., on top of each other, and horizontally, i.e., next to each other. In this study, we keep the 2D examples limited to three columns of bricks for simplicity. One brick is specified as a brick that needs to be retrieved. However, to retrieve the desired brick, one must first remove all the bricks on top of it, one brick at a time. The arrangement of bricks can be presented both as a text or an image.

3.2 Data generation

The data for this task, i.e., the 1D and 2D brick arrangements, are generated programmatically, where the position, label, and colors of each brick are selected randomly. The basis for our brick generation script is the code published by [9]. In both 1D and 2D cases, we first generate the coordinates of each brick. We call a textual description that describes the brick arrangement through their coordinates a *task-implicit* textual description. Then, we also generate a *task-explicit* textual description using the spatial relation terms to describe the relations between the bricks. In the 1D cases, the spatial relation term that is used is *"on top of"* - e.g., brick A is *on top of* brick B. In 2D cases, there are two additional spatial relation terms used - *to the left of* and *to the right of.* Each brick is mentioned exactly once in a description, giving its name and relation to one other brick or its name and color only if all other bricks have already been mentioned.

In addition to textual descriptions, we also generate image representations of each brick arrangement. Bricks are depicted as squares that are stacked on top of each other and have one column (or stack) in 1D cases, or three columns in 2D cases. Each square is labeled with the name of the corresponding brick, represented by any randomly selected uppercase letter from A to Z, ensuring that each letter is unique. The images are created in both color and black and white (bw) versions to test if the color influences the success of the task solving. For 2D cases, we also generate images where spacing is introduced between columns to test if this highlights their topological properties and alleviates the issue of the model removing bricks from adjacent columns. We call these kinds of images *task-explicit*, while we call images without spacing between bricks *task-implicit.* Table 1 shows examples of 1D and 2D brick arrangements with the corresponding task-implicit and task-implicit textual and image descriptions.

3.3 Prompting

We employ different prompts to instruct LMMs in solving the specified brick retrieval task, tailored to the input modality (text or image) and the representation of the brick arrangement (task-implicit or task-explicit). All prompts in this study are zeroshot, meaning that the model has only one attempt at answering. We also did not perform any fine-tuning or pre-training on the models. Text-based task-implicit prompts remain mostly unchanged from the ones used by [9], while the other prompts are newly generated.

Text-based prompts are first presented with a message that instructs the model to work on the problem step by step, thereby triggering CoT reasoning:

Let's think step by step, and provide the answer in the format of a sequence of bricks by a comma in the last sentence.

This is followed by a question template that specifies the task to the LLM. Each task instance has a different brick arrangement and target brick, so these elements are filled in for each task instance:

Question: There is a set of bricks. [brick arrangement description] Now we have to get a specific brick. The bricks must now be grabbed from top to bottom, and if the lower brick is to be grabbed, the upper brick must be removed first. How to get brick [target brick]?

This prompt remains the same for 1D and 2D cases because the same rules apply, and the only difference is in the spatial relation terms they use in the textual description of the brick arrangements. The *[brick arrangement description]* can be either task-implicit or task-specific (see Table 1).

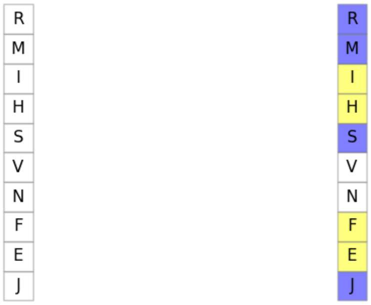
Image-based prompts use images instead of text to describe the brick arrangements. Again, there is no differentiation in the image prompt between different types of images used. Image prompts are also presented with the chain of thought instruction, and then

[1] https://github.com/ivan-majic/llm_modality_reasoning.git

Table 1: 1D (top) and 2D (bottom) brick arrangements with corresponding textual descriptions and images.

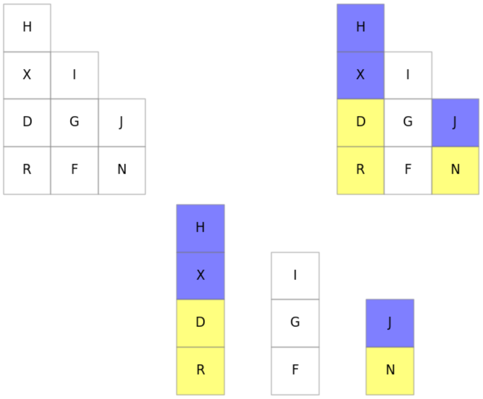
Text task-implicit:
“Brick J is at position (0, 0). Brick E is at position (0, 1). Brick F is at position (0, 2). Brick N is at position (0, 3). Brick V is at position (0, 4). Brick S is at position (0, 5). Brick H is at position (0, 6). Brick I is at position (0, 7). Brick M is at position (0, 8). Brick R is at position (0, 9).”

Text task-explicit:
“The brick V is on top of the brick N. The brick S is on top of the brick V. The brick F is on top of the brick E. For the brick J, the color is blue. The brick H is on top of the brick S. The brick R is on top of the brick M. The brick M is on top of the brick I. The brick I is on top of the brick H. The brick N is on top of the brick F. The brick E is on top of the brick J.”



Text task-implicit:
“Brick R is at position (0, 0). Brick D is at position (0, 1). Brick X is at position (0, 2). Brick H is at position (0, 3). Brick F is at position (1, 0). Brick G is at position (1, 1). Brick I is at position (1, 2). Brick N is at position (2, 0). Brick J is at position (2, 1).”

Text task-explicit:
“There is a set of bricks. The brick H is on top of the brick X. The brick J is on top of the brick N. The brick N is to the right of the brick F. The brick G is on top of the brick F. The brick X is on top of the brick D. The brick I is to the right of the brick X. The brick D is on top of the brick R. For the brick R, the color is yellow. The brick F is to the right of the brick R.”



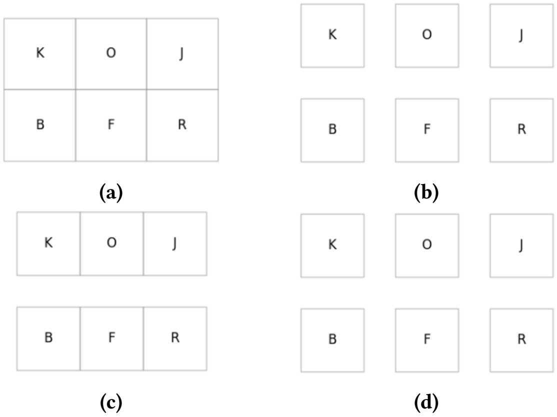


Figure 2: Example of no spacing (a), and 0.5 spacing between bricks in the horizontal (b), vertical (c), and both (d) dimensions.

followed by a specific textual prompt that explains the task and specifies the target brick for that specific task instance:

Let’s think step by step, and provide the answer in the format of a sequence of bricks by a comma in the last sentence.

The image shows a set of bricks that can be placed on top of each other. Now we have to get a specific brick. The bricks must now be grabbed from top to bottom, and if the lower brick is to be grabbed, the upper brick must be removed first. How to get brick [target brick]?

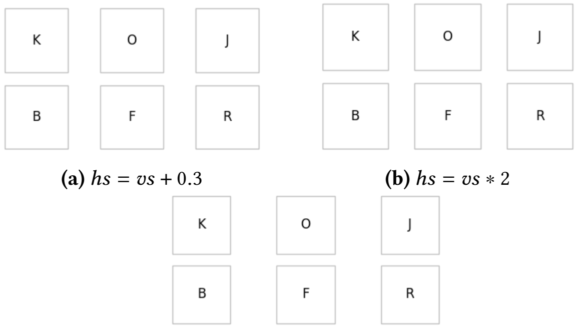
3.4 Ablation study of brick spacing

To test the effect of task-implicit versus task-explicit representation of the brick world in image prompts, we carried out an ablation study. We fine-tune the images in our prompts by introducing and increasing spacing between bricks to make the model focus more on the relative positions of objects rather than their adjacency in the image (pixel) space. In the first instance, we gradually increased the spacing between bricks in an image horizontally, then vertically, and then in both directions (Figure 2). We consider only 2D image prompts in this ablation study and increase the spacing from 0 to 1 in increments of 0.1 units. Since the bricks in our images are depicted as squares with a width of 1, the maximum spacing is the same as the width of a brick.

In the second instance, we tested an asymmetrical spacing between bricks that is larger in the horizontal dimension than the vertical. By doing so, we are effectively directing the reasoning of the AI system towards the vertical direction, which is the directions that is relevant for our task. Here, we express the horizontal spacing (hs) as a function of vertical spacing (vs). We define three such functions in this ablation study: $hs = vs + 0.3$, $hs = vs * 2$, and $hs = vs * 4$ (Figure 3).

3.5 Evaluation

The answers are first parsed to extract the part that represents the brick removal sequence provided by the LMM. These sequences are then evaluated against the ground truth using three evaluation metrics: accuracy, precision, and recall. Following [9], accuracy



(a) $hs = vs + 0.3$ **(b)** $hs = vs * 2$

(c) $hs = vs * 4$

Figure 3: Example of asymmetrical spacing between bricks, where horizontal spacing (hs) is a function of vertical spacing (vs). In this example, vs=0.2.

is defined as the success rate in achieving the final goal, i.e., the ratio of predicted sequences that exactly match the ground truth sequences. The precision and recall are calculated using the Longest Common Sequence (LCS) between the predicted and ground truth sequences to measure their similarity. Thus, precision is defined as the ratio of the LCS and the length of the predicted sequence. Recall is defined as the ratio of the LCS and the ground truth sequence.

4 Results

Here we present the results of our prompting experiments, based on different prompt characteristics. For this study, we generated 50 1D and 50 2D brick arrangements. First, we analyze the performance of all prompt dimensions, modalities, descriptions, and colors. Then we analyze the effects of brick spacing on the accuracy of 2D image based prompts in an ablation study.

4.1 Prompting results

b

We executed our prompts on three AI models from OpenAI using their API service[2]: gpt-4o-2024-08-06, gpt-4o-mini, gpt-3.5-turbo. The "4o" family of GPT models are the latest and most advanced LMM models published to date by OpenAI, and support image and text modalities as input for their prompts. The gpt-3.5-turbo is a legacy LLM model that supports only text-based prompts. We test prompts with both task-implicit and task-explicit brick representations, and for the image-based prompts, we test both the black and white and color versions.

The results of the brick removal planning task are evaluated in Table 2. As one can expect, it can be seen that all three models perform better when dealing with the simpler 1D cases versus 2D when corresponding categories are compared. Further, the results show that image-based prompts can achieve better accuracy over text-based prompts, especially in 2D cases where text-based prompts perform poorly for all models.

Regarding the brick descriptions used in prompts, the results show that task-explicit brick descriptions always outperform task-implicit brick descriptions. The difference is largest in 2D image-based prompts, where the accuracy has increased by 74% for gpt-4o-2024-08-06 and 46% for gpt-4o-mini. Task-implicit text-based prompts perform particularly badly with gpt-4o-mini achieving the highest accuracy of 22% across all models and dimensions.

The effect of color on image-based prompts is not consistent across models. For gpt-4o-2024-08-06 colored images show slightly better performance in 1D and slightly worse performance in 2D cases. For the gpt-4o-mini model, there is a stronger effect visible where black and white images achieve 18% higher accuracy in 1D, and 14% in 2D cases.

4.2 Ablation study results

We performed an ablation study using the 2D image-based prompts with black and white images, and the gpt-4o-2024-08-06 model. We start with the task-implicit brick description (i.e., no spacing between bricks) and gradually increase the spacing. For this experiment, 50 2D brick arrangements were used with 10 different spacing sizes (i.e., 0.1 - 1) across three different spacing directions (i.e., horizontal, vertical, both). This results in the total of 1500 images.

Results are shown in Figure 4a and indicate that horizontal spacing has a very positive effect where the accuracy increases from 26% for no spacing to over 93% for images with horizontal spacing of 0.3. Further increases in horizontal spacing show much lesser effect on performance, but the largest spacing of 1 achieves the perfect accuracy.

Spacing in vertical or both directions does not seem to provide a meaningful increase in the performance. Vertical spacing actually consistently lowers the accuracy, making it less than 5% with the spacing of 0.5 or more. Spacing in both directions shows little effect on accuracy, with accuracy increasing to 31% when the spacing is 0.4, but ultimately decreasing to 18% when the spacing is 1 in both directions.

In the second ablation study, we tested the effect of asymmetrical spacing between bricks on the performance, where the horizontal spacing (hs) is always larger than the vertical spacing (vs). We test three different functions for the hs, each with 10 different spacing values and 50 different brick arrangements, which again results in the total of 1500 image-based prompts.

Results of this study are shown in Figure 4b and demonstrate mixed effect of asymmetrical spacing on the accuracy. In the case of the fixed difference between hs and vs ($hs = vs + 0.3$) the accuracy shows a large decline from 90% to 45% as soon as the vertical spacing is introduced. Further increases in hs and vs using this equation do not have a meaningful effect on accuracy. When hs is expressed as a multiplication of the vs, we see an increase in accuracy. However, this is increase is quite modest when $hs = vs * 2$ is used, as the maximum accuracy achieved is 52%. When a larger multiplication is used in $hs = vs * 4$, the increase in accuracy is more pronounced and achieves the maximum accuracy of 90% when $vs = 0.7$ and $hs = 2.8$.

[2] https://platform.openai.com

Table 2: Overall task performance expressed as accuracy, precision, and recall by dimension, modality, color, and model.

					gpt-4o-2024-08-06			gpt-4o-mini			gpt-3.5-turbo	
				accuracy	precision	recall	accuracy	precision	recall	accuracy	precision	recall
	modality	description	color									
1D	text	task-implicit	–	12.0	85.5	24.8	22.0	67.6	43.8	2.0	82.6	17.5
		task-explicit	–	92.0	99.6	98.3	38.0	86.6	63.8	54.0	82.0	71.0
	image	task-implicit	bw	96.0	100.0	96.6	82.0	88.7	85.5	–	–	–
			color	100.0	100.0	100.0	54.0	71.8	63.6	–	–	–
2D	text	task-implicit	–	6.0	57.3	37.4	8.0	43.8	44.2	8.0	61.6	39.4
		task-explicit	–	38.0	75.0	78.0	22.0	65.6	68.4	14.0	37.7	46.8
	image	task-implicit	bw	26.0	44.9	72.0	0.0	29.5	48.3	–	–	–
			color	20.0	40.7	64.3	0.0	27.2	43.6	–	–	–
		task-explicit	bw	100.0	100.0	100.0	46.0	62.5	67.5	–	–	–
			color	98.0	98.4	98.4	38.0	63.0	62.8	–	–	–

5 Discussion

Our experiments have tested the effect of both prompt modality and spatial description on the performance of large multimodal models (LMMs) in a spatial planning task. We first show that all models perform significantly better in 1D versus 2D cases, especially when the text modality and task-explicit image modality are used (Table 2). This result aligns with our expectations, given that the task requires removing all bricks positioned on top of the target brick, implying a clear vertical hierarchy. In 1D cases, there is only one column of bricks, and they only span in the vertical dimension, eliminating potential confusion from neighboring columns, as seen in the 2D cases.

As a solution to this column-shift problem in 2D cases, we assumed that the relational nature of the brick removal task would mean that emphasizing this information in the prompts, particularly with the task-explicit spatial descriptions, would lead to a significant increase in accuracy. While this is generally true across both text-based and image-based prompts, the findings are not entirely straightforward. As Table 2 illustrates, the results for 2D image-based task-explicit prompts, where only horizontal spacing is applied, show a substantial improvement over task-implicit prompts (i.e., no spacing between bricks), with accuracy rising from 26% to 100%. However, our ablation studies tested the extent and direction of spacing required for this effect, and the results offer a more complex picture.

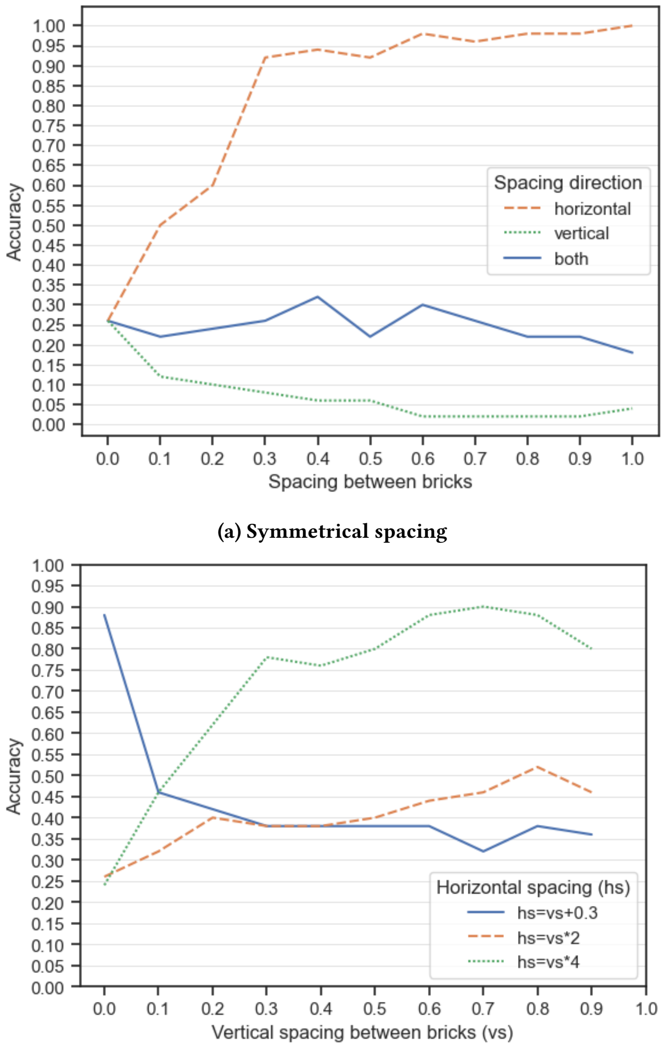
When we applied only horizontal spacing, even a modest spacing of 0.3 resulted in a significant improvement in accuracy (Figure 4a). Yet, when vertical spacing was introduced, this benefit was negated, and in some cases, accuracy declined. Contrary to our expectations, symmetrical spacing in both the horizontal and vertical dimensions did not enhance accuracy at all. This means that generic conversion of image prompts by introducing equal spacing in all directions does not completely solve the column-shift issue that occurs in 2D cases. This suggests that the improvement in accuracy from spacing is highly task-specific: it works effectively only when the spacing helps the models to focus on the spatial relations that matter for completing the task. We do not know what the effect of spacing would be in tasks that consider horizontal hierarchy or are isotropic and this would be an interesting study in the future.

We can interpret the adjustments we made in our experiments as a form of prompt fine-tuning, rather than model fine-tuning. This let us assess the capability of generic AI systems to perform a very specific spatial planning task with various levels of help or hinting provided in the task definition itself. Future studies could explore whether similar accuracy improvements could be achieved by maintaining the task-implicit spatial descriptions in the prompts (i.e., text prompts with coordinates and image prompts without spacing), but fine-tuning the AI system itself specifically for this task. Such a study could reveal whether fine-tuning the AI system directly could match or even surpass the benefits we have observed from optimizing the prompt design and compare the feasibility of both approaches.

6 Conclusion and future work

This paper investigated the effect of modality on the ability of AI systems to solve a spatial reasoning task. Specifically, we considered the modality of prompts (text versus image) and the spatial descriptions of the environment. We adopted a brick removal task from [9] and extended it by generating more complex 2D cases and introducing image-based prompts, which were not considered before. We also employed two types of spatial descriptions: task-implicit and task-explicit. Our experiments were conducted using several large multimodal models, and we tested the effectiveness of these different modalities through various prompt designs.

Through our experiments, we show that aligning the spatial description of the prompt with the properties of the task significantly increases the accuracy of AI systems in spatial reasoning. In text-based prompts, this is done by describing the environment with spatial-relation terms (i.e., relative positions) instead of the absolute positions of blocks. In images, spatial task-explicity is achieved by introducing spacing between blocks whose relative positions are not useful for LMMs to solve the task. Before introducing task-explicit spatial descriptions, the maximum achieved accuracy for 2D text-based and image-based tasks was only 8% and 26%, respectively. After the task-explicit spatial descriptions were introduced, the accuracy increased to 38% for text and 100% for



(a) Symmetrical spacing

(b) Asymmetrical spacing

Figure 4: The accuracy of 2D image-based brick removal prompts in relation to different spacings between bricks in the image. Prompts were evaluated using gpt-4o-2024-08-06.

image prompts. Thus, we can conclude that task-explicit spatial descriptions outperform task-implicit spatial descriptions, and that image-based prompts significantly outperform text-based prompts in the more complex 2D brick world.

In contrast to the prompt manipulation investigated here, future work could explore AI system fine-tuning, specific to the spatial reasoning task at hand, as an alternative strategy to improve the system's performance. Additionally, more complex spatial scenarios and other spatial reasoning tasks that focus on different directions need to be investigated to better understand the potential and limitations of spatial reasoning with AI systems.

Acknowledgments

The authors acknowledge support by the National Science Foundation under Grant 2033521 A1: KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

[1] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14455–14465.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017). https://doi.org/10.1109/TPAMI.2017.2699184

[3] Anthony Cohen. 2024. Recent Work on Spatial Relations and Large Language Models. https://arxiv.org/abs/2406.16528

[4] Max J. Egenhofer and Robert D. Franzosa. 1991. Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems* 5, 2 (1991), 161–174. https://doi.org/10.1080/02693799108927841

[5] Max J. Egenhofer and John R. Herring. 1990. A Mathematical Framework for the Definition of Topological Relationships. In *Proceedings of the Fourth International Symposium on Spatial Data Handling*. 803–813.

[6] Andrew U. Frank. 1992. Spatial Concepts, Geometric Data Models, and Geometric Data Structures. *Computers & Geosciences* 18, 4 (1992), 409–417. https://doi.org/10.1016/0098-3004(92)90047-D

[7] Song Gao and Yuhan Ji. 2023. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations. In *Proceedings of the 12th International Conference on Geographic Information Science (GIScience 2023) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 277)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 43:1–43:6. https://doi.org/10.4230/LIPIcs.GIScience.2023.43

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[9] Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2024. Chain-of-Symbol Prompting Elicits Planning in Large Langauge Models. https://doi.org/10.48550/arXiv.2305.10276 arXiv:2305.10276 [cs]

[10] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. 2020. GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond. *International Journal of Geographical Information Science* 34, 4 (April 2020), 625–636. https://doi.org/10.1080/13658816.2019.1684500

[11] Yuhan Ji and Song Gao. 2023. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations. In *12th International Conference on Geographic Information Science (GIScience 2023) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 277)*, Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:6. https://doi.org/10.4230/LIPIcs.GIScience.2023.43

[12] Chunyuan Li. 2023. Large multimodal models: Notes on cvpr 2023 tutorial. *arXiv preprint arXiv:2306.14895* (2023).

[13] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zheng Xie, Yixuan Wei, and Jian Sun. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2021.00123

[14] Daniel R. Montello. 1993. Scale and Multiple Psychologies of Space. In *Spatial Information Theory: A Theoretical Basis for GIS*, Andrew U. Frank and Irene Campari (Eds.). Springer, 312–321. https://doi.org/10.1007/3-540-57207-4_21

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[17] Li Zhangyu. 2024. Spatial Planning in Robotics: Insights from Computer Vision. https://arxiv.org/pdf/2407.10380

[18] Li Zhangyu and Li Gengchen. 2024. Spatial Foundation Models and Creating Embeddings for Spatial Representation. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=X2Wfl1UAAAAJ