

# FOREST: FRAME OF REFERENCE EVALUATION IN SPATIAL REASONING TASKS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Spatial cognition is one fundamental aspect of human intelligence. A key factor in spatial cognition is understanding the frame of reference (FoR) that identifies the perspective of spatial relations. However, the AI research has paid very little attention to this concept. Specifically, there is a lack of dedicated benchmarks and in-depth experiments analyzing large language models’ (LLMs) understanding of FoR. To address this issue, we introduce a new benchmark, **Frame of Reference Evaluation in Spatial Reasoning Tasks** (FoREST) to evaluate LLMs ability in understanding FoR. We evaluate the LLMs in identifying the FoR based on textual context and employ this concept in text-to-image generation. Our results reveal notable differences and biases in the FoR identification of various LLMs. Moreover, the bias in FoR interpretations impacts the LLMs’ ability to generate layouts for text-to-image generation. To deal with these biases, we propose Spatial-Guided prompting, which guides the model in exploiting the types of spatial relations for a more accurate FoR identification. This approach reduces FoR bias in LLMs and improves the overall performance of FoR identification. Eventually, using FoR information in text-to-image generation leads to a more accurate visualization of the spatial configuration of objects.

## 1 INTRODUCTION

Spatial reasoning plays a significant role in human cognition and conducting daily activities. It is also a crucial aspect in many AI problems, including language grounding (Zhang & Kordjamshidi, 2022; Yang et al., 2024), navigation (Yamada et al., 2024), computer vision (Liu et al., 2023; Chen et al., 2024), medical domain (Gong et al., 2023), and image generation (Gokhale et al., 2023). One key concept in spatial cognition is the frame of reference (FoR), which identifies the perspective of spatial expressions. Levinson (2003) initially defines three basic FoR classes: intrinsic, relative, and absolute. The intrinsic FoR describes spatial expressions based on the viewer’s perspective, while the relative FoR uses the object’s perspective. The last type is the absolute FoR, which uses environmental cues such as cardinal directions. The framework from Tenbrink (2011), which is the main FoR framework of our work, expanded these basics. These concepts have been studied extensively in cognitive linguistics (Edmonds-Wathen, 2012; Vukovic & Williams, 2015). However, only limited studies investigate how AI models understand FoR. Recent benchmarks for evaluating spatial understanding primarily focus on reasoning on objects and their spatial relations. For instance, Shi et al. (2022), Mirzaee & Kordjamshidi (2022), and Rizvi et al. (2024) propose benchmarks that assess comprehension of complex spatial scenes. While concentrating on complex reasoning tasks, they give minimal attention to the notion of FoR. They often limit evaluations to intrinsic FoR, using one object as the center of coordinates. Similarly, recent text-to-image benchmarks (Gokhale et al., 2023; Huang et al., 2023; Cho et al., 2023a;b) suffer from the same limitation, where spatial expressions are evaluated in terms of the relative FoR, i.e., camera perspective. Consequently, this potentially restricts the situated spatial reasoning abilities in dynamic environments and interactive settings where the perspective can change. Nonetheless, studies in computer vision (Liu et al., 2023) and robotics (Liu et al., 2010; Kang & Han, 2023) have begun exploring FoR understanding.

To systematically investigate the concept of FoR in spatial understanding and provide new resources, we introduce **Frame of Reference Evaluation in Spatial Reasoning Tasks** (FoREST) benchmark to assess models’ ability to understand FoR classes from textual descriptions and extend this to grounding and visualization. Our dataset consists of two splits: ambiguous (A-split) and clear (C-split). The

A-split contains spatial expressions with FoR ambiguity, meaning multiple valid FoRs can apply to the explained situation. In contrast, the C-split has spatial expressions with only one valid FoR. This design allows us to evaluate models’ understanding of spatial expressions in ambiguous and clear contexts. We conduct experiments with large language models (LLMs) to identify FoR classes in spatial expressions and employ this concept in text-to-image models. Our findings reveal performance differences across FoR classes and show that LLMs tend to be biased toward particular FoRs when spatial expressions with ambiguous FoRs are provided. The bias is also evident in diffusion models that use LLM-generated layouts in the image generation pipeline. These diffusion models tend to perform better in one specific FoR class. To address these biases, we propose Spatial Guided (SG) prompting, which encourages models to consider general types of spatial relations: direction, topology, and distance, in their reasoning process. We hypothesize that these relations provide essential information to help accurately identify FoR classes. Our results confirm this, showing improved identification of FoR classes, reduced bias, and enhanced layout generation, ultimately benefiting downstream tasks like text-to-image generation.

To summarize our contributions, 1. We introduce the FoREST benchmark to systematically evaluate large language models’ abilities to identify FoR classes from textual spatial expressions, experimenting with various in-context learning approaches for FoR identification. 2. We assess the impact of using FoR information on text-to-image generation using diffusion models, including stable and layout diffusion models. 3. We propose a new prompting approach that considers the types of spatial relations in its reasoning process and improves FoR identification and image generation quality.

## 2 PRIMITIVES

We review three aspects of spatial information expressed in language: spatial roles, spatial relations, and frame of reference.

**Spatial Roles.** We use the main conceptual roles defined in spatial language literature (Kordjamshidi et al., 2010; Tenbrink, 2011). These roles include Locatum (L), Relatum (R), and Perspective. The **locatum** represents the object described in the spatial expression. While the **relatum** represents another object used to describe the location of the locatum. Lastly, **perspective** is defined as the origin of a coordinate system used as the basis for determining the direction. For example, “a cat is to the left of a dog from the owner.” In this example, a cat is the locatum, a dog is a relatum, and the perspective is the owner’s coordinate.

**Spatial Relations.** When dealing with spatial knowledge representation and reasoning, often three main relations categories are considered: directional, topological, and distance (Hernández, 1994; Cohn & Renz, 2008; Kordjamshidi et al., 2010).

1. **Directional:** These relations define one object’s direction from another based on specific coordinates. Examples of relations include left, right, above, and below.

2. **Topological:** These relations describe the containment between two objects, such as inside.

3. **Distance:** These relations provide qualitative and quantitative relations between entities. Examples of qualitative distance relations are near and far, and quantitative distance relations are 3km.

**Spatial Frame of Reference.** We use the four frames of reference investigated in-depth in the cognitive linguistic studies (Tenbrink, 2011) and are defined as follows.

1. *external intrinsic.* It describes a spatial relation based on the relatum’s perspective, which does not contain the locatum. The top-right image in Figure 1 illustrates this scenario with the sentence, “A cat is to the right of the car from the car’s perspective.”

2. *external relative.* It presents a spatial relation based on the observer’s perspective, which may not be presented in the context. The top-left image in Figure 1 shows an example with the sentence, “A cat is to the left of a car from my perspective.”

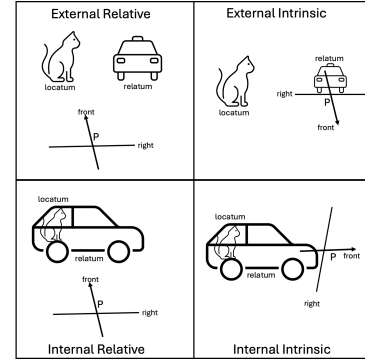


Figure 1: Illustration of FoR classes. The Cat is the locatum, the Car is the relatum, and the arrow indicates the perspective.

3. *internal intrinsic*. It expresses a spatial relation based on the relatum’s perspective, which contains the locatum. The bottom-right image in Figure 1 illustrates this circumstance with the sentence, “The cat is inside and back of a car from a car’s point of view.”

4. *internal relative*. It describes a spatial relation from the observer’s perspective where the locatum is inside the relatum. The bottom-left image in Figure 1 displays this relation with the sentence, “A cat is inside and to the left of the car from my perspective.”

### 3 FOREST DATASET CONSTRUCTION

We propose a new problem setting to identify the FoR in linguistic expressions to evaluate the LLMs’ understanding on spatial frames of reference(FoR).. In this setting, the language model receives a textual spatial explanation as input, denoted as  $T$ , and the model outputs an FoR class in  $FoR = \{\text{external intrinsic, external relative, internal relative, internal intrinsic}\}$  according to the primitives defined in Section 2. We introduce the **Frame of Reference Evaluation in Spatial Reasoning Tasks (FoREST)** benchmark to evaluate models’ performance on this problem. We should note that identifying FoR is challenging and, in some cases, inherently ambiguous. For example, in “a cat is to the left of a dog.”, It has two correct interpretations. The first one is *external relative* FoR interpretation, “a cat is to the left of a dog from the camera’s perspective.” Another valid interpretation for *external intrinsic* FoR is “a cat is to the left of a dog from the dog’s perspective.” To distinguish clear from ambiguous cases, we create two splits for our FoREST dataset: ambiguous (A-split) and clear (C-split). Spatial expressions in the A-split can have more than one valid FoR, while C-split expressions only have one valid FoR.

#### 3.1 FOR CATEGORIES BASED ON RELATUM TYPE

Using the FoR classes defined in Section 2, we found that two properties of relatum cause FoR ambiguity. The first property is the relatum’s intrinsic direction. It creates ambiguity between intrinsic and relative FoR classes since spatial relations can originate from both the relatum’s and observer’s perspectives. The second is the relatum’s affordance as a container. It introduces the ambiguity between internal and external FoR classes since spatial relations can refer to the inside and outside of the relatum. We use the combination of these two properties to define four cases of relatum: the cow case, box case, car case, and pen case. We use these cases to divide the A-split of our dataset into four subsets. Then, we create clear counterparts of these cases to generate the C-split of our dataset. There are two types of clear cases. The first type is inherently clear from the context, such as “a pencil is to the right of a pen.” In this case, there are no different interpretations about the spatial configuration of the two objects. However, another type needs additional information to be clear, such as “A cat is to the left of the dog.” In this type, we add a clause clarifying the perspective or topology. For example, “the cat is to the left of the dog from the dog’s perspective.” In the following, we further clarify the four ambiguous cases based on the properties of the relatum.

**Case 1: Cow Case.** We create a cow case as a subset of our A-split. We select a relatum with intrinsic directions but without affordance as the container. The obvious example is a cow, which should not be a container but has a front and back. In such a case, the relatum potentially provides a perspective for spatial relations. Thus, the applicable FoR classes are  $FoR = \{\text{external intrinsic, external relative}\}$ . We explicitly augment such cases with perspective information to resolve the ambiguity and add their clear counterparts to the C-split. To specify the perspective, we use templates for augmenting clauses, such as “from {relatum}’s perspective” for *external intrinsic* or “from my perspective” for *external relative*. An example of A-split context is “a cat is to the right of the cow.” The counterparts included in the C-split are “a cat is to the right of the cow from cow’s perspective.” for *external intrinsic* and “a cat is to the right of the cow from my perspective” for *external intrinsic*.

**Case 2: Box Case.** We create a box-case subset as part of the A-split. Unlike the cow case, the relatum selected in this subset can be a container but lacks intrinsic directions. For example, a box can serve as a container without having intrinsic directions. An internal FoR can be established since the relatum can be a container. Accordingly, the applicable FoR classes of this context are  $FoR = \{\text{external relative, internal relative}\}$ , causing the ambiguity. To include their unambiguous counterparts in the C-split, we explicitly specify the topology between locatum and relatum by adding “inside” for *internal relative* and “outside” for *external relative* in the spatial expression.

An example of the A-split context is “A cat is to the right of the box.” The counterpart for *internal relative* is “a cat is inside and to the right of the box.” The counterpart for *external relative* is “a cat is outside and to the right of the box.” We add both counterparts in the C-split.

**Case 3: Car Case.** We introduce the third case subset of A-split, Car case. We select the relatum with intrinsic direction and affordance as a container for this case. With these two properties, the relatum can provide the perspective for spatial relations and have the locatum inside, allowing both intrinsic and internal FoR classes. An obvious example is a car that can be a container with intrinsic directions. Therefore, the applicable frames of reference classes are  $FoR = \{ \textit{external relative}, \textit{external intrinsic}, \textit{internal intrinsic}, \textit{internal relative} \}$ , which introduces FoR ambiguity. We resolve this ambiguity by including perspective and topology information in the context to create clear counterparts for the C-split. The template for augment clauses is reused from the Cow case and Box case for perspective and topology information, respectively. A proper example of context in A-split is “a person is in front of the car.” The four counterparts to include in the C-split are “a person is outside and in front of the car from the car itself” for *external intrinsic*, “a person is outside and in front of the car from the observer” for *external relative*, “a person is inside and in front of the car from the car itself” for *internal intrinsic*, and “a person is inside and in front of the car from the observer” for *internal relative*.

**Case 4: Pen Case.** We called the last subset of A-split with the Pen case. The last case covers the circumstance that the relatum neither has the intrinsic direction nor the affordance as a container. An obvious example is a pen that does not have a left or right direction nor the ability to be a container. Lacking these two properties, the created context should be clear and have one applicable FoR,  $FoR = \{ \textit{external relative} \}$ . There is no ambiguity to clarify since there is only one valid FoR class. Therefore, we can reuse it in the C-split without modifications. An example of such a context is “the book is to the left of a pen.”

### 3.2 CONTEXT VISUALIZATION

As a part of the dataset, we include the image visualizations of spatial expressions. In intrinsic FoR classes, the relatum’s perspective influences how we position the locatum when visualizing spatial expressions, leading to visualization ambiguity. For example, given the expression “a cow is to the right of a car relative to the car,” with the car’s position fixed in the scene, the cows can be placed in different positions depending on the car’s orientation. To address this issue, we extend the context in both splits of FoREST by adding the relatum’s orientation information. To specify the relatum’s orientation, we use templates such as “facing forward.” For instance, “a cat is to the left of a dog” is extended to “a cat is to the left of a dog, facing forward.” In this way, we obtain I-A-split from A-split and I-C-split from C-split. We restrict I-A-split and I-C-split to external FoR classes to avoid occlusion in the visualization since one object can become invisible in internal FoR classes. We then create scene configurations based on the spatial expressions in I-A-split and I-C-split, as illustrated in Figure 2. We use the Unity-3D simulator<sup>1</sup> to process scene configurations and generate four visualizations for each one. The detail on the simulation is provided in the Appendix B.

### 3.3 RELATUM/LOCATUM SELECTION

We selected nine object sets to support the four FoR cases defined above. For instance, an example set of objects is “small objects with intrinsic direction.” Selected objects in this group, such as dogs and cats, are guaranteed to have intrinsic direction without the affordance of being containers. This set is used to create the Cow Case context and visualization. All sets of objects are in the Appendix B. The total number of selected objects is 20, enough to cover all defined FoR cases.

### 3.4 DATASET CREATION PROCEDURE

The pipeline is illustrated in Figure 2 to combine all the above-explained procedures. First, we select a set of locatum and relatum based on the FoR cases defined in Section 3.1 to form A-split spatial expressions. We substitute the actual locatum and relatum objects in the Spatial Relation template, “<locatum> <spatial relation> <relatum>.” In the figure, left is the spatial relation, locatum is a horse, and relatum is a cow. After obtaining the A-split contexts, we create their counterparts using

<sup>1</sup><https://unity.com>

the perspective/topology clauses described in Section 3.1 represented in yellow text. Next, we apply the orientation template described in Section 3.2 to prepare the context for the visualization. We then create the scene configuration from modified spatial expression and send it to the simulator to finalize visualizations. The dataset statistic is in Appendix A, and the complete sets of all patterns and entities are included in Appendix B.

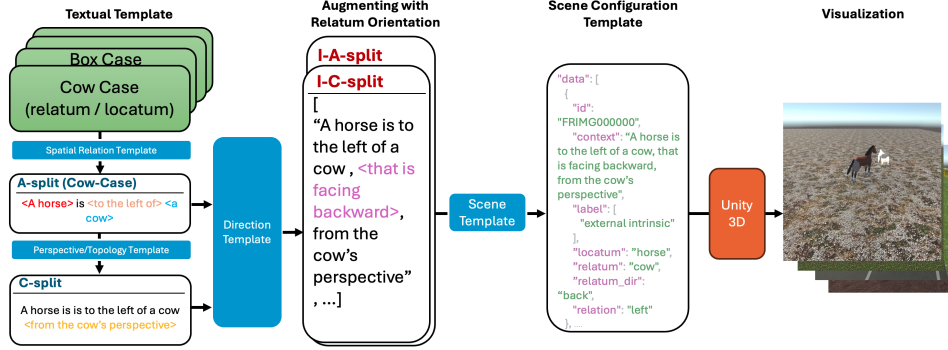


Figure 2: The pipeline of creating FoREST dataset starts by selecting the locatum and relatum based on defined FoR cases. Next, a spatial template is applied to generate the A-split, which is then extended into the C-split by applying a topology/perspective template. Afterward, the I-C-split and I-A-split are created by including a direction template into the C-split and A-split. Finally, scene configurations are generated from the I-C-split and I-A-split to create visualizations using Unity3D.

## 4 MODELS AND TASKS

### 4.1 FOR IDENTIFICATION

**Task.** We evaluate the LLMs’ performance in recognizing the FoR classes from given spatial expressions. The LLMs receive aspatial expression, denoted as  $T$ , and output one FoR class,  $F$ , from the valid set of FoR classes,  $F \in FoR = \{\text{external relative, external intrinsic, internal relative}\}$ . All in-context learning examples are in the Appendix C.

**Zero-shot model.** We follow the regular setting of *zero-shot* prompting. We only provide instruction to LLM with spatial context. The instruction prompt briefly explains each class of the FoR and candidate answers for the LLM. We called the LLM with the instruction prompt and  $T$  to find  $F$ .

**Few-shot model.** We manually craft four spatial expressions for each FoR class. To avoid creating bias, each spatial expression is ensured to fit in only one FoR class. These expressions serve as examples of our *few-shot* setting. We provide these examples in addition to the instruction as a part of the prompt, followed by  $T$  and query  $F$  from the LLM.

**Chain-of-Thought (CoT) model.** To create CoT (Wei et al., 2023) examples, we modify the prompt to require reasoning before answering. Then, we manually crafted reasoning explanations with the necessary information for each example used in few-shot. Finally, we call the LLMs, adding modified instructions to updated examples, followed by  $T$  and query  $F$ .

**Spatial-Guided Prompting (SG) model.** We hypothesize that the general spatial relation types defined in Section 2 can provide meaningful information for recognizing FoR classes. For instance, a topological relation, such as “inside,” is intuitively associated with an internal FoR. Therefore, we propose Spatial-Guided Prompting to direct the model in identifying the type of relations before querying  $F$ . We revise the prompting instruction to guide the model in considering these three aspects. Then, we manually explain these three aspects. We specify the relation’s origin from the context for direction relations, such as “the left direction is relative to the observer.” We hypothesize that this information helps the model distinguish between intrinsic and relative FoR. Next, we specify whether the locatum is inside or outside the relatum for topological relations. This information should help distinguish between internal and external FoR classes. Lastly, we provide the potential quantitative distance, e.g., far. This quantitative distance further encourages identifying the correct topological and directional relations. Eventually, we insert these new explanations in examples and call the model with the updated instructions followed by  $T$  to query  $F$ .

## 4.2 TEXT-TO-IMAGE (T2I)

**Task.** The input to the text-to-image is a spatial expression,  $T$ , and output from the model is a generated image, denoted as  $I$ , corresponding to given  $T$ . This task aims to determine the diffusion models’ ability to consider FoR by assessing their generated images.

**Stable Diffusion models.** We evaluate the performance of the stable diffusion models for the simplest baseline of T2I models. This model only needs the scene description as input. Therefore, we provide  $T$  to the model and expect an output image of  $I$ .

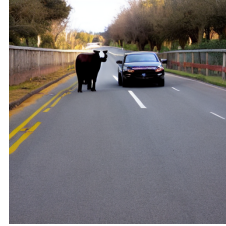
**Layout Diffusion models.** We evaluate the Layout Diffusion model for more advanced T2I models. The layout diffusion model has two phases: text-to-layout and layout-to-image. As the LLMs can be used to generate the bounding box layout (Cho et al., 2023b; Lian et al., 2024), we provide  $T$  to LLMs with the instruction to generate the layout including bounding box coordinates for each object in the format of  $\{\text{object: } [x, y, w, h]\}$ , where  $x$  and  $y$  represent the starting point of the bounding box and  $h$  and  $w$  represent the height and width of the bounding box. After generating the bounding box coordinates, they are provided with  $T$  as an additional input for the layout-to-image model to create the output image,  $I$ .

**Spatial-Guide Layout Diffusion models.** We propose Spatial-Guide Layout Diffusion pipeline for image generation, which introduces an additional step before the text-to-layout phase. This step involves obtaining the FoR information from  $T$ , denoted as  $S(T)$ . We guide LLMs to extract direction, topology, and distance information from  $T$  to generate  $S(T)$ . Following the SG prompting procedure, we create examples for this step. Then, we provide examples to help the model understand the task and generate  $S(T)$ . Once  $S(T)$  is generated, it is used as supplementary information to guide the LLMs in generating bounding box coordinates. This model allows us to consider FoRs in image generation and assess their impact on the T2I task. After obtaining the bounding box coordinates, we follow the same outline in Layout Diffusion to generate the final image.

## 5 EXPERIMENTAL RESULTS



(a) An image generated from SD-2.1.



(b) An image generated from Llam3-8B + GLIGEN.

Figure 3: Two images generated from the ambiguous spatial expression “A car is to the right of a cow.” (a) is correct by intrinsic FoR interpretation, while (b) is correct by relative FoR interpretation.

### 5.1 EVALUATION METRICS

**FoR Identification.** We report the accuracy of the model on the multi-class classification task. Note that the expressions in A-split can have multiple correct answers. Therefore, we consider the prediction correct when it is in one of the valid FoR classes for the given spatial expression.

**T2I.** To evaluate the generated images, we assess the generated objects and their spatial relationships. To do so, inspired by *spatialEval* (Cho et al., 2023b), we detect the spatial relation in images. However, we modify their approach to consider the given FoR when evaluating spatial relations. In particular, we convert all relations based on their FoR to be expressed from camera view and then pass it to *spatialEval* evaluation since *spatialEval* assumes the camera perspective. When evaluating the generated image from a context with FoR ambiguity, we consider it correct if it fits one of the valid FoRs for the given situation. See Figure 3 where context with FoR ambiguity produces two correct images in different FoR interpretations. We report the evaluation score in terms of  $\text{VISOR}_{\text{cond}}$  and  $\text{VISOR}_{\text{uncond}}$  (Gokhale et al., 2023). VISOR score is a metric designed to compare

the spatial understanding abilities of T2I models. The  $\text{VISOR}_{\text{cond}}$  evaluates the spatial relations and only includes the cases with both objects mentioned in the spatial expression correctly appearing in the generated image. In other words, it ignores cases with object errors and focuses on how well the model interprets spatial relations, which is the target of our work. While the  $\text{VISOR}_{\text{uncond}}$  evaluates the model’s overall performance, including object creation errors.

## 5.2 EXPERIMENTAL SETTING

**FoR Identification.** We selected five different LLMs including Llama3-8B, Llama3-70B (Llama, 2024), Gemma2-9B (Gemma, 2024), GPT-3.5-turbo (Brown et al., 2020), and GPT-4o (OpenAI, 2024) as the backbones for prompt engineering. The version of GPT-3.5-turbo is "gpt-3.5-turbo-0125," and GPT-4o is "gpt-4o-2024-05-13". We set the temperature of all models to be 0 to make the experiments reproducible. For each model, we apply several in-context learning (ICL) approaches including, *zero-shot*, *few-shot*, *CoT*, and our technique of Spatial-Guided Prompting (SG) as described in Section 4.1. For *few-shot*, *CoT*, and *SG*, we provide four examples to the models. The procedures for creating examples for each ICL are described in Section 4.1. The data splits used in these experiments are A-split and C-split.

**T2I.** We select Stable Diffusion 1.5 (SD-1.5) and Stable Diffusion 2.1 (SD-2.1) (Rombach et al., 2021) for stable diffusion models. For the backbone of layout-to-image, we choose GLIGEN (Li et al., 2023). We utilize Llama3-8B and Llama3-70B to handle the transition from spatial description to the textual bounding box information. The bounding box format is described in Section 4.2. To generate FoR information, we use the same selection of LLMs for the Spatial-Guided Layout Diffusion (SG Layout Diffusion), explained in Section 4.2. We generated four images per spatial expression to evaluate performance and calculated the VISOR score, following the original paper in Gokhale et al. (2023). The number of inference steps for all text-to-image models was set to 50. The data splits used in these experiments are I-A-split and I-C-split. For the evaluation, we select grounding DINO (Liu et al., 2024) and DPT (Ranftl et al., 2021), following VPEval Cho et al. (2023b), to detect objects and depth map, respectively.

We conduct all experiments and evaluations on GPU A6000, taking roughly 300 GPU hours.

## 5.3 RESULTS

Model	A-split	C-Split				
		Avg.	ER-C-Split	EI-C-Split	II-C-Split	IR-C-Split
Gemma2-9B (0-shot)	<b>94.17</b>	60.45	<b>94.24</b>	35.98	53.91	57.66
Gemma2-9B (4-shot)	59.58	64.29(↑ 3.84)	55.89(↓ 38.34)	72.61(↑ 36.63)	74.22(↑ 20.31)	54.44(↓ 3.23)
Gemma2-9B (CoT)	60.49	65.64(↑ 5.20)	60.49(↓ 33.74)	60.54(↑ 24.57)	87.50(↑ 33.59)	54.03(↓ 3.63)
Gemma2-9B (SG)(Our)	72.67	70.13(↑ 9.68)	65.87(↓ 28.37)	65.54(↑ 29.57)	53.12(↓ 0.78)	<b>95.97(↑ 38.31)</b>
Llama3-8B (0-shot)	59.58	65.73	60.36	83.80	56.25	62.50
Llama3-8B (4-shot)	59.58	63.32(↓ 2.41)	58.68(↓ 1.68)	61.74(↓ 22.07)	81.25(↑ 25.00)	51.61(↓ 10.89)
Llama3-8B (CoT)	66.19	68.31(↑ 2.58)	66.19(↑ 5.83)	56.63(↓ 27.17)	99.22(↑ 42.97)	51.21(↓ 11.29)
Llama3-8B (SG) (Our)	72.73	67.08(↑ 1.35)	69.88(↑ 9.52)	49.24(↓ 34.57)	100.00(↑ 43.75)	49.19(↓ 13.31)
Llama3-70B (0-shot)	77.33	44.62	35.04	32.39	57.81	53.23
Llama3-70B (4-shot)	59.78	63.81(↑ 19.20)	59.78(↑ 24.74)	66.52(↑ 34.13)	77.34(↑ 19.53)	51.61(↓ 1.61)
Llama3-70B (CoT)	66.00	70.88(↑ 26.27)	68.01(↑ 32.97)	65.65(↑ 33.26)	91.41(↑ 33.59)	58.47(↑ 5.24)
Llama3-70B (SG) (Our)	74.94	83.33(↑ 38.71)	78.17(↑ 43.13)	70.87(↑ 38.48)	100.00(↑ 42.19)	84.27(↑ 31.05)
GPT3.5 (0-shot)	60.88	62.04	60.62	62.50	74.22	50.81
GPT3.5 (4-shot)	59.58	72.68(↑ 10.65)	39.64(↓ 20.98)	99.89(↑ 37.39)	100.00(↑ 25.78)	51.21(↑ 0.40)
GPT3.5 (CoT)	59.13	70.65(↑ 8.61)	59.52(↓ 1.10)	74.67(↑ 12.17)	100.00(↑ 25.78)	48.39(↓ 2.42)
GPT3.5 (SG) (Our)	77.59	82.01(↑ 19.97)	69.62(↑ 9.00)	97.93(↑ 35.43)	100.00(↑ 25.78)	60.48(↑ 9.68)
GPT4o (0-shot)	59.90	77.85	60.43	99.35	100.00	51.61
GPT4o (4-shot)	59.78	82.32(↑ 4.47)	59.91(↓ 0.52)	<b>100.00(↑ 0.65)</b>	100.00	69.35(↑ 17.74)
GPT4o (CoT)	64.31	81.49(↑ 3.65)	63.99(↑ 3.56)	99.89(↑ 0.54)	100.00	62.10(↑ 10.48)
GPT4o (SG) (Our)	69.88	<b>85.78(↑ 7.94)</b>	70.08(↑ 9.65)	99.67(↑ 0.33)	<b>100.00</b>	73.39(↑ 21.77)

Table 1: Accuracy results report from FoR Identification with LLMs. The correct prediction is one of the valid FoR classes for the given spatial expression. All FoR classes are external relative (ER), external intrinsic (EI), internal intrinsic (II), and internal relative (IR).

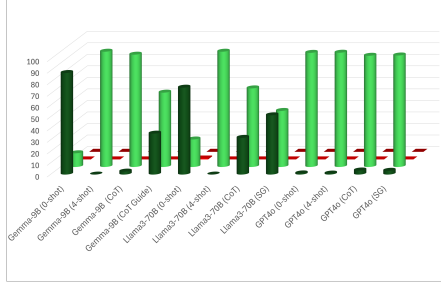
### 5.3.1 FoR BIAS IN LLMs

**C-split.** The *zero-shot* setting reflects the LLMs’ inherent bias in identifying FoR. Table 1 presents the accuracy for each FoR class in C-split, where sentences explicitly include information about topology and perspectives. We found that some models strongly recognize specific FoR classes.

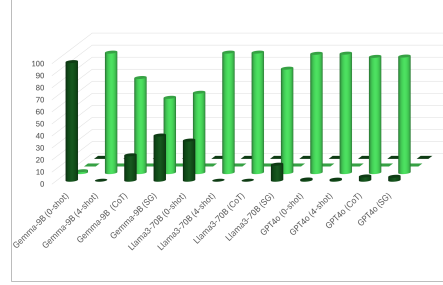
Notably, Gemma2-9B achieves a near-perfect accuracy of 94.24% on external relative FoR but performs poorly on other classes, especially external intrinsic. In contrast, GPT4o shows exceptional performance in intrinsic FoR classes, with 99.35% for external intrinsic and 100% for internal intrinsic, showing an opposite behavior to Gemma2.

**A-split.** We examine the FoR bias in the A-split. Based on the results in Table 1, we plotted the top-3 models’ results (Gemma2-9B, Llama3-70B, and GPT4o) for a more precise analysis in Figures 4. The plots show the frequencies of each FoR category. According to the plot, Gemma and GPT have strong biases toward external relative and external intrinsic, respectively. This bias helps Gemma2 perform well in the A-split since all spatial expressions can be interpreted as external relative. However, GPT4o’s bias leads to errors when intrinsic FoRs aren’t valid, as in the Box and Pen cases (see plots (c) and (d)). Llama3 exhibits different behavior, showing a bias based on the relatum’s properties, specifically the relatum’s affordance as a container. In cases where relatum cannot serve as containers, i.e., Cow and Pen cases, Llama3 favors external relative. Conversely, Llama3 tends to favor external intrinsic when the relatum has the potential to be a container.

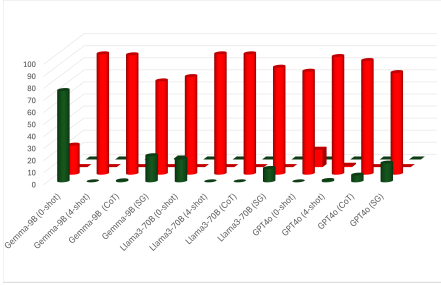
### 5.3.2 BEHAVIOR WITH ICL VARIATIONS



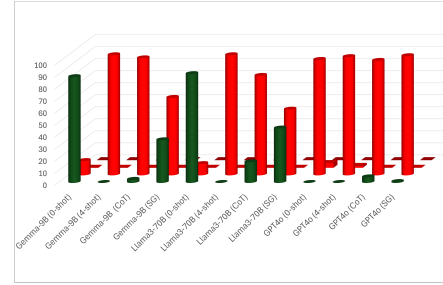
(a) Results of Cow Case in A-Split.



(b) Results of Car Case in A-Split.



(c) Results of Box Case in A-Split.



(d) Results of Pen Case in A-Split.

Figure 4: Red shows the wrong FoR identifications, and green shows the correct ones. The dark color is for relative FoRs, while the light color is for intrinsic FoRs. The round shape is for the external FoRs, while the square is for internal FoRs. The depth of the plots shows the four FoRs, i.e., *external relative*, *external intrinsic*, *internal intrinsic*, and *internal relative*, **from front to back**.

**C-split.** We observe the models’ behavior under various in-context learning (ICL) methods. In the C-split results from Table 1, the *few-shot* method outperforms *zero-shot* for most LLMs in intrinsic FoR classes. However, this approach often leads to decreased performance in relative FoRs. Then, we notice that applying *CoT* prompting generally improves performance in larger LLMs. However, *CoT* causes performance drops in Gemma2-9B and GPT-3.5, similar to *few-shot*. The decline is less severe for external relative FoR but more pronounced for internal relative FoR. Our proposed Spatial-Guide (SG) prompting significantly outperforms *CoT*, especially for larger models. GPT4o, using SG prompting, achieves state-of-the-art (SOTA) performance compared to other baselines. However, applying SG prompting to smaller models like Gemma2-9B and Llama3-8B has drawbacks. For Gemma2-9B, performance decreases in external relative and internal intrinsic FoR while improving in external intrinsic and internal relative. Interestingly, this effect is reversed for Llama3-8B. The same dropping trend can be seen when prompting requiring a longer explanation is applied to Llama3-8B.

**A-split.** We use the same Figure 4 to observe the behavior when applying ICL. The A-split shows minimal improvement with ICL variations, though some notable changes are observed. With *few-shot*, all models show a strong bias toward external intrinsic FoR, even when the relatum lacks intrinsic directions, i.e., Box and Pen cases. This bias appears even in Gemma2-9B, which usually behaves differently. This suggests that the models pick up biases from the examples despite efforts to avoid such patterns. However, *CoT* reduces some bias, leading LLMs to revisit relative, which is generally valid across scenarios. In Gemma2, the model predicts relative FoR where the relatum has intrinsic directions, i.e., Cow and Car cases. Llama3 behaves similarly in cases where the relatum cannot act as a container, i.e., Cow and Pen cases. GPT4o, however, does not depend on the relatum’s properties and shows slight improvements across all cases. Unlike *CoT*, our SG prompting is effective in all scenarios. It significantly reduces biases while following a similar pattern to *CoT*. Specifically, SG prompting increases external relative predictions for Car and Cow in Gemma2-9B, and for Cow and Pen in Llama3-70B. Nevertheless, GPT4o shows only a slight bias reduction. However, Our proposed method improves the overall performance of most models, as shown in Table 1. The Llama3-70B behaviors are also seen in Llama3-8B and GPT3.5. The plots for these LLMs are in Appendix E due to lack of space.

### 5.3.3 FoR IMPACT ON IMAGE GENERATION

Model	VISOR(%)					
	cond (I)	cond (R)	cond (avg)	cond (I)	cond (R)	cond (avg)
	I-A-Split			I-C-Split		
SD-1.5	72.72	48.95	68.72	53.92	53.77	53.83
SD-2.1	79.46	54.10	75.39	<b>60.06</b>	59.64	59.83
Llama3-8B + GLIGEN	79.45	66.08	77.38	57.51	65.98	62.12
Llama3-70B + GLIGEN	83.94	68.68	81.43	56.47	69.53	63.49
Llama3-8B + SG + GLIGEN (Our)	85.42	<b>71.14</b>	83.17	58.84	<b>70.36</b>	<b>65.15</b>
Llama3-70B + SG + GLIGEN (Our)	<b>87.13</b>	66.56	<b>83.75</b>	56.77	70.04	64.06

Table 2:  $VISOR_{cond}$  score on the I-A-Split and I-C-Split where *I* refer to the Cow Case and Car Case where relatum has intrinsic directions, and *R* refer to the Box Case and Pen case where relatum lacks intrinsic directions, *avg* is mirco-average of *I* and *R*. *cond* are explained in Section 5.1.

We evaluate SG layout diffusion to assess the impact of using FoR on image generation. We focus on  $VISOR_{cond}$  as it better reflects the model’s spatial understanding than the overall performance measured by  $VISOR_{uncond}$ . Due to space limitations,  $VISOR_{uncond}$  results are reported in Appendix D. Table 2 shows that adding FoR information (Llama3 + SG + GLIGEN) improves performance across all splits compared to the baseline models (Llama3 + GLIGEN). The most significant gains occur when the relatum lacks intrinsic direction, making external relative FoR the only valid option. This demonstrates an improved understanding of SG layout diffusion regarding relative FoRs. The performance gap in relative FoR understanding is further evident when comparing SD-2.1 with Llama3 + GLIGEN. GLIGEN models consistently perform better due to their ability to use spatial configurations based on bounding boxes. However, SD-2.1 surpasses all GLIGEN-based models, including ours, when FoR is intrinsic, as seen in the *cond(I)* of the I-C split in Table 2. This limitation likely arises from the reliance on bounding boxes for generating spatial configurations, which complicates handling intrinsic FoR due to the lack of object properties and orientation. This challenge is further highlighted in Table 3, which analyzes the I-A split in the Cow and Car case. In this case, GLIGEN favors external relative interpretations more than SD-2.1. Also, the results suggest Llama may be biased toward external intrinsic FoR when generating layouts that align with the FoR class identification in SG prompting. Incorporating FoR enhances intrinsic FoR understanding, showing improvements in our method compared to the layout diffusion baseline in both Cow and Car case splits.

To further explain these improvements, we assess the generated bounding boxes in the I-C split for left and right relations relative to the camera since these can be evaluated using only bounding boxes without depth information. As seen in Table 4, our SG prompting improved

Model	EI	ER
SD-1.5	51.11	21.61
SD-2.1	57.97	21.49
Llama3-8B + GLIGEN	53.67	25.78
Llama3-70B + GLIGEN	54.49	29.45
Llama3-8B + SG + GLIGEN (Our)	57.46	27.96
Llama3-70B + SG + GLIGEN (Our)	56.54	30.59

Table 3: The separate accuracy visualized Cow and Car Case in I-A-Split for external relative (ER) and external intrinsic interpretations (EI) of FoR.

Llama3-70B’s by 3.48%, while Llama3-8B saw a slight decrease of 0.22%. This evaluation was conducted on all generated layouts from the I-C split, which differs from the evaluation subset of images used for  $\text{VISOR}_{\text{cond}}$  in Table 2. For a consistent evaluation, we report the  $\text{layout}_{\text{cond}}$  score in the same table.  $\text{Layout}_{\text{cond}}$  shows that Llama3-8B improves within the same evaluation subset with  $\text{VISOR}_{\text{cond}}$ . Overall, by incorporating FoR information through SG layout diffusion, Llama3 generates better spatial configurations, enhancing image generation performance.

Model	Layout	$\text{Layout}_{\text{cond}}$
Llama3-8B	85.26	88.84
Llama3-8B + SG	85.04	88.86
Llama3-70B	88.47	93.16
Llama3-70B + SG	91.95	95.45

Table 4: Layout accuracy where spatial relations are left or right relative to the camera. Layout is evaluated for all generated layouts in I-C split while  $\text{Layout}_{\text{cond}}$  uses the same testing examples as  $\text{VISOR}_{\text{cond}}$ .

## 6 RELATED WORKS

Understanding situated spatial expressions requires knowledge of the frame of reference (FoR), which defines the coordinate system used to describe objects’ positions. A detailed study of the FoR on multiple natural languages was conducted in (Levinson, 2003), which categorizes the FoR into three basic categories: intrinsic, relative, and absolute. Inspired by this basic framework, Tenbrink 2011 proposed a more comprehensive framework for specifying the FoR, used as the primary reference of our study. Their frameworks extended the basics with other spatial relation concepts, such as topology and temporal. Cognitive studies have increasingly focused on how humans perceive spatial FoR. Many findings in these studies suggest that humans favor specific FoR classes (Edmonds-Wathen, 2012; Vukovic & Williams, 2015; Shusterman & Li, 2016; Ruotolo et al., 2016). For instance, Ruotolo et al. 2016 investigated how the FoR affects the human’s ability to memorize and describe the scene within a limited time. They found that participants were better at describing and answering questions when the spatial relations were based on participants’ position, as opposed to using other objects as reference points. This highlights a gap between the relative and intrinsic FoR.

Several benchmarks have been developed across various domains to evaluate the spatial understanding of computation models. In the text-based domain, recent benchmarks focus on navigating with spatial instructions (Yamada et al., 2024) or question-answering tasks (Shi et al., 2022; Mirzaee & Kordjamshidi, 2022; Rizvi et al., 2024). These benchmarks are developed to assess the spatial reasoning capability without paying much attention to FoR. Existing research often lacks explicit consideration of FoR, and the benchmarks do not include FoR annotations. Consequently, evaluating FoR understanding remains a research gap in spatial reasoning-related work. Similarly, text-to-image (T2I) benchmarks (Gokhale et al., 2023; Huang et al., 2023; Cho et al., 2023a;b) face the same issue. They usually focus on correctly placing two objects based on spatial relation from the camera perspective and relative FoR. Nevertheless, few works in vision-text domains are starting to recognize the importance of a FoR (Chen et al., 2024; Liu et al., 2023). One notable study is provided by Liu et al. 2023. They provide a case study on the FoR and results showing that making the model capable of understanding the FoR affects downstream performance on visual question answering. However, their study is limited in terms of FoR categories. In our work, we extend the coverage of benchmarks into more diverse frames of reference for the FoR recognition tasks. Moreover, we are the first to study the impact of FoR identification on text-to-image generation as a downstream task.

## 7 CONCLUSION

Given the significance of spatial reasoning in AI applications and the importance of understanding spatial frame of reference (FoR), we introduce **Frame of Reference Evaluation in Spatial Reasoning Tasks (FoREST)** benchmark to assess FoR comprehension in text-based spatial expressions and its impact on grounding in visual modality by diffusion models. Our benchmark results reveal notable differences in FoR identification in various LLMs. Moreover, the bias in FoR interpretations impacts the LLMs’ ability to generate layouts for text-to-image generation. To deal with these biases, we propose Spatial-Guided prompting, which guides the model in considering the type of spatial relations: topology, distance, and direction for a more accurate FoR identification. This approach reduces the FoR biases in LLMs and improves the overall performance of the FoR identification task. Eventually, it enhances text-to-image generation performance by providing more accurate spatial configurations.

## REFERENCES

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities, 2024. URL <https://arxiv.org/abs/2401.12168>.
- Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models, 2023a. URL <https://arxiv.org/abs/2202.04053>.
- Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation, 2023b. URL <https://arxiv.org/abs/2305.15328>.
- Anthony G. Cohn and Jochen Renz. Chapter 13 qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter (eds.), *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pp. 551–596. Elsevier, 2008. doi: [https://doi.org/10.1016/S1574-6526\(07\)03013-1](https://doi.org/10.1016/S1574-6526(07)03013-1). URL <https://www.sciencedirect.com/science/article/pii/S1574652607030131>.
- Cris Edmonds-Wathen. False friends in the multilingual mathematics classroom. In :, pp. 5857–5866, 2012. URL <http://www.icme12.org/>.
- Gemma. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation, 2023. URL <https://arxiv.org/abs/2212.10015>.
- Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adaptor: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation, 2023.
- Daniel Hernández (ed.). *Reasoning with qualitative representations*, pp. 55–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. ISBN 978-3-540-48425-7. doi: 10.1007/BFb0020333. URL <https://doi.org/10.1007/BFb0020333>.
- Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, 2023. URL <https://arxiv.org/abs/2307.06350>.
- Soo-Han Kang and Ji-Hyeong Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, 15(4):631–641, Apr 2023. ISSN 1875-4805. doi: 10.1007/s12369-021-00842-1. URL <https://doi.org/10.1007/s12369-021-00842-1>.
- Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Task definition and annotation scheme. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2010/pdf/846\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/846_Paper.pdf).
- Stephen C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press, 2003.

- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation, 2023. URL <https://arxiv.org/abs/2301.07093>.
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2024. URL <https://arxiv.org/abs/2305.13655>.
- Changsong Liu, Jacob Walker, and Joyce Y. Chai. Ambiguities in spatial language understanding in situated human robot dialogue. In *Dialog with Robots, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*, volume FS-10-05 of *AAAI Technical Report*. AAAI, 2010. URL <http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2292>.
- Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. URL <https://arxiv.org/abs/2205.00363>.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL <https://arxiv.org/abs/2303.05499>.
- Llama. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Roshanak Mirzaee and Parisa Kordjamshidi. Transfer learning with synthetic corpora for spatial role labeling and reasoning, 2022. URL <https://arxiv.org/abs/2210.16952>.
- OpenAI. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. URL <https://arxiv.org/abs/2103.13413>.
- Md Imbesat Hassan Rizvi, Xiaodan Zhu, and Iryna Gurevych. Sparc and sparp: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models, 2024. URL <https://arxiv.org/abs/2406.04566>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- Francesco Ruotolo, Tina Iachini, Gennaro Ruggiero, Ineke J. M. van der Ham, and Albert Postma. Frames of reference and categorical/coordinate spatial relations in a “what was where” task. *Experimental Brain Research*, 234(9):2687–2696, Sep 2016. ISSN 1432-1106. doi: 10.1007/s00221-016-4672-y. URL <https://doi.org/10.1007/s00221-016-4672-y>.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts, 2022. URL <https://arxiv.org/abs/2204.08292>.
- Anna Shusterman and Peggy Li. Frames of reference in spatial language acquisition. *Cognitive Psychology*, 88:115–161, 2016. ISSN 0010-0285. doi: <https://doi.org/10.1016/j.cogpsych.2016.06.001>. URL <https://www.sciencedirect.com/science/article/pii/S0010028516301190>.
- Thora Tenbrink. Reference frames of space and time in language. *Journal of Pragmatics*, 43(3):704–722, 2011. ISSN 0378-2166. doi: <https://doi.org/10.1016/j.pragma.2010.06.020>. URL <https://www.sciencedirect.com/science/article/pii/S037821661000192X>. The Language of Space and Time.
- Nikola Vukovic and John N. Williams. Individual differences in spatial cognition influence mental simulation of language. *Cognition*, 142:110–122, 2015. ISSN 0010-0277. doi: <https://doi.org/10.1016/j.cognition.2015.05.017>. URL <https://www.sciencedirect.com/science/article/pii/S0010027715001146>.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.

Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models, 2024. URL <https://arxiv.org/abs/2310.14540>.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701, 2024. doi: 10.1109/ICRA57147.2024.10610443.

Yue Zhang and Parisa Kordjamshidi. Lovis: Learning orientation and visual signals for vision and language navigation, 2022.

## A DATASET STATISTICS

The FoREST dataset statistic is provided in the Table 5.

Case	A-Split	I-A-Split	FoR class	C-Split	I-C-Split
Cow Case	792	3168	External Relative	1528	4288
Box Case	120	120	External Intrinsic	920	3680
Car Case	128	512	Internal Intrinsic	128	0
Pen Case	488	488	Internal Relative	248	0
Total	1528	4288	Total	2824	7968

Table 5: Dataset Statistic of FoREST dataset.

## B DETAILS CREATION OF FOREST DATASET

We define the nine categories of objects selected in our dataset as indicated below in Table 6. We select sets of locatum and relatum based on the properties of each class to cover four cases of frame of reference defined in Section 3.1. Notice that we also consider the appropriateness of the container; for example, the car should not contain the bus.

Based on the selected locatum and relatum. To create an A-split spatial expression, we substitute the actual locatum and relatum objects in the Spatial Relation template. After obtaining the A-split contexts, we create their counterparts using the perspective/topology clauses to make the counterparts in C-split. Then, we obtain the I-A and I-C split by applying the directional template to the first occurrence of relatum when it has intrinsic directions. The directional templates are "that is facing towards," "that is facing backward," "that is facing to the left," and "that is facing to the right." All the templates are in the Table 7. We then construct the scene configuration from each modified spatial expression and send it to the simulator developed using Unity3D. Eventually, the simulator produces four visualization images for each scene configuration.

Category	Object	Intrinsic Direction	Container
small object without intrinsic directions	umbrella, bag, suitcase, fire hydrant	✗	✗
bog object with intrinsic directions	bench, chair	✓	✗
big object without intrinsic direction	water tank	✗	✗
container	box, container	✗	✓
small animal	chicken, dog, cat	✓	✗
big animal	deer, horse, cow, sheep	✓	✗
small vehicle	bicycle	✓	✗
big vehicle	bus, car	✓	✓
tree	tree	✗	✗

Table 6: All selected objects with two properties: intrinsic direction, affordance of being container

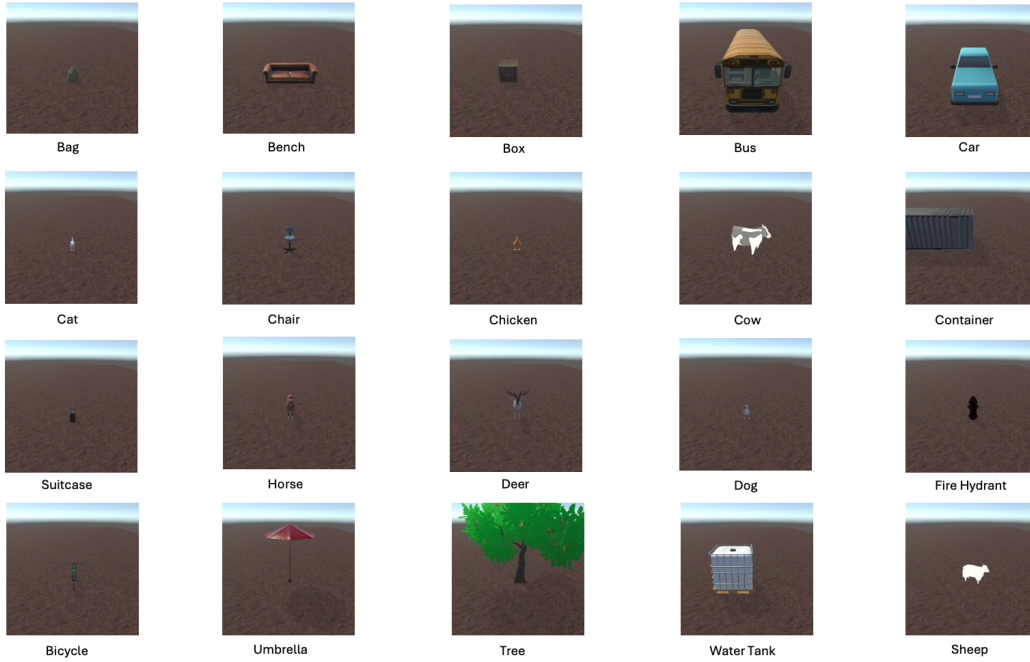


Figure 5: All 3d models used to generate visualizations for FoREST.

### B.1 SIMULATION DETAILS

The simulation starts with randomly placing the relatum into the scene with the orientation based on the given scene configuration. We randomly select the orientation by given scene configuration,  $[-40, 40]$  for front,  $[40, 140]$  for left,  $[140, 220]$  for back, and  $[220, 320]$  for right. Then, we create the locatum from the relatum position and move it in the spatial relation provided. If the frame of reference is relative, we move the locatum based on the camera’s orientation. Otherwise, we move it from the relatum’s orientation. Then, we check the camera’s visibility of both objects. If one of them is not visible, we repeat the process of generating the relatum until the correct placement is achieved. After getting the proper placement, we randomly choose the background from 6 backgrounds. Eventually, we repeat the procedures four times for one configuration.

### B.2 OBJECT MODELS AND BACKGROUND

For the object models and background, we find it from the unity asset store<sup>2</sup>. All of them are free and available for download. All of the 3D models used are shown in Figure 5.

### B.3 TEXTUAL TEMPLATES

All the templates used to create FoREST are given in Table 7.

## C IN-CONTEXT LEARNING

We provide the prompting for each in-context learning. The prompting for *zero-shot* and *few-shot* is provided in Listing 1. The instruction answer for these two in-context learning is “Answer only the category without any explanation. The answer should be in the form of {Answer: Category.}”

For the Chain of Thought (CoT), we only modified the instruction answer to “Answer only the category with an explanation. The answer should be in the form of {Explanation: Explanation Answer: Category.}” Similarly to CoT, we only modified the instruction answer to “Answer only

<sup>2</sup><https://assetstore.unity.com>

Spatial Relation Templates	{locatum} is in front of {relatum} {locatum} is on the left of {relatum} {locatum} is to the left of {relatum} {locatum} is behind of {relatum} {locatum} is back of {relatum} {locatum} is on the right of {relatum} {locatum} is to the right of {relatum}
Topology Templates	within {relatum} and inside {relatum} and outside of {relatum}
Perspective Templates	from {relatum}'s view relative to {relatum} from {relatum}'s perspective from my perspective from my point of view relative to observer
Directional Templates	that is facing toward that is facing backward that is facing to the left that is facing to the right

Table 7: All templates used to create FoREST dataset.

the category with an explanation regarding topological, distance, and direction aspects. The answer should be in the form of {Explanation: Explanation Answer: Category.}", respectively. The example responses are provided in Listing 2 for Spatial Guided prompting.

```

1 # Instruction to find frame of reference class of given context
2 ""
3 Instruction:
4 You specialize in language and spatial relations, specifically in the
5 reference frame of context. Identify the following context into the
6 frame of reference categories (external intrinsic, internal intrinsic
7 , external relative, internal relative) based on the information.
8
9 External intrinsic is the context that uses spatial relation to describe
10 the relative position of the object by referring to the referenced
11 object's intrinsic directions, and both objects do not contain one
12 another.
13
14 Internal intrinsic is the context that uses spatial relation to describe
15 the relative position of the object by referring to the referenced
16 object's intrinsic directions, and one object is inside another one.
17
18 External relative is the context that uses spatial relation to describe
19 the relative position of the object by referring to the referenced
20 observer's intrinsic directions, and both objects do not contain one
21 another.
22
23 Internal relative is the context that uses spatial relation to describe
24 the relative position of the object by referring to the referenced
25 observer's intrinsic directions, and one object is inside another one
26 .
27
28 {Instruction answer}
29
30 Context: {spatial exprssion}
31
32 ""
33
34 # Instruction for generate bounding box
35 ""

```

```

810 22 | Your task is to generate the bounding boxes of objects mentioned in the
811      | caption.
812 23 | The image is size 512x512. The bounding box should be in the format of (x
813      | , y, width, height). Please considering the frame of reference of
814      | caption and direction of reference object if possible. If needed, you
815      | can make the reasonable guess.
816 24 | ""

```

Listing 1: Prompting to find the frame of reference class of given context and generate bounding box layout

```

820 1 | ## This is example response used in few-shot, CoT, and SG
821 2 | # Example 1: The bird is in front of the car relative to the car.
822 3 | # Few-shot response: Answer: external intrinsic
823 4 | # CoT response: Explanation: The position of the bird refers to the front
824      | direction of the car. Answer: external intrinsic.
825 5 | # SG response: Explanation: Topological: A bird is outside of the car.
826      | Distance: A bird may be slightly far from the front of the car from
827      | the context. Direction: The car, the relative object, has the
828      | direction, and the bird is actually in front of the car. Answer:
829      | external intrinsic "
830 6 |
831 7 | # Example 2: The bird is inside of the car and left of the car from the
832      | car's perspective.
833 9 | # Few-shot response: Answer: internal intrinsic
834 10 | # CoT response: Explanation: The spatial relationship is described
835      | relative to the inside of the car. The bird is accuracy left of the
836      | car. Answer: internal intrinsic.
837 11 | # SG response: Explanation: Topological: A bird is inside the car.
838      | Distance: A bird should be in the same location as the car from the
839      | context. Direction: The car which is the relative object has the
840      | direction, and bird is in the front side of the car. Answer: internal
841      | intrinsic.
842 12 |
843 13 | # Example 3: The box is inside and at the back of the room.
844 14 | # Few-shot response: Answer: internal relative
845 15 | # CoT response: Explanation: The box is inside of the room, and the room
846      | doesn't have an intrinsic direction, which makes the back relation
847      | based on the observer's perspective of the room. Answer: internal
848      | relative.
849 17 | # SG response: Explanation: Topological: box is inside of the room.
850      | Distance: box should be in the same location with the room from the
851      | context. Direction: Room doesn't have the direction and context
852      | refere from observer perspective for box is in the back of the room.
853      | Answer: internal relative.
854 18 |
855 19 | # Example 4: A phone is on the left of a tablet from my perspective.
856 20 | # Few-shot response: Answer: internal relative
857 21 | # CoT response: Explanation: left relation is based on the observer's (me
858      | ) perspective to the room even if tablet has the intrinsic direction
859      | and a phone is not inside the tablet. Answer: external relative.
860 23 | # SG response: Explanation: Topological: phone is not inside of the
861      | tablet. Distance: phone may has some distance from tablet according
862      | to the context. Direction: even if tablet has direction, the context
863      | left relation is refered from observer's perspective that a phone is
864      | on the left side of the tablet location. Answer: external relative.

```

Listing 2: Spatial expression examples with the response for few-shots, Chain-of-Thought (CoT), and Spatial Guide (SG) prompting

## D VISOR SCORE

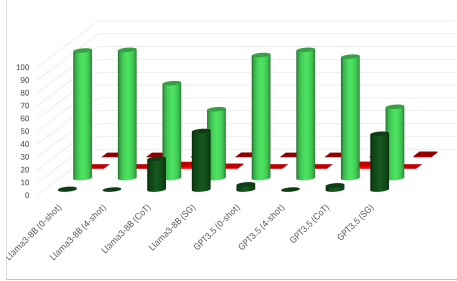
$\text{VISOR}_{\text{uncond}}$  provides the overall spatial relation score, including images with object generation errors. Since it is less focused on evaluating spatial interpretation than  $\text{VISOR}_{\text{cond}}$ , which assesses explicitly the text-to-image model’s spatial reasoning, we report  $\text{VISOR}_{\text{uncond}}$  results here in the Table 8 rather than in the main paper. The results are similar to the pattern observed in  $\text{VISOR}_{\text{uncond}}$  that the based model performs better in the relative frame of reference, while the model is better in the intrinsic frame of reference.

Model	VISOR(%)					
	uncond (I)	uncond (R)	uncond (avg)	uncond (I)	uncond (R)	uncond (avg)
	I-A-Split			I-C-Split		
SD-1.5	45.43	33.22	43.51	35.06	35.68	35.40
SD-2.1	<b>62.87</b>	43.90	<b>59.89</b>	<b>45.98</b>	46.59	<b>46.31</b>
Llama3-8B + GLIGEN	46.74	38.16	45.39	33.98	39.36	36.89
Llama3-70B + GLIGEN	54.33	46.89	53.17	38.04	46.04	42.37
Llama3-8B + SG + GLIGEN (Our)	51.83	43.24	50.48	36.28	44.43	40.70
Llama3-70B + SG + GLIGEN (Our)	58.92	<b>47.44</b>	57.12	38.23	<b>48.62</b>	43.86

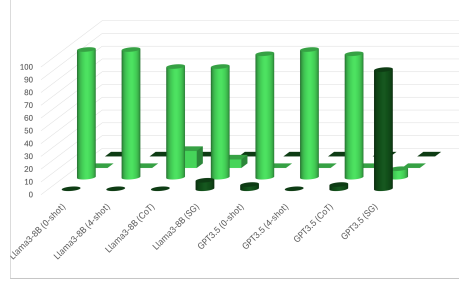
Table 8:  $\text{VISOR}_{\text{uncond}}$  score on the I-A-Split and I-C-Split where *I* refer to the Cow Case and Car Case where relatum has intrinsic directions, and *R* refer to the Box Case and Pen case where relatum lacks intrinsic directions, *avg* is mirco-average of *I* and *R*. cond and uncond are explained in Section 5.1.

## E A-SPLIT PLOT FOR FREQUENCY PREDICTIONS

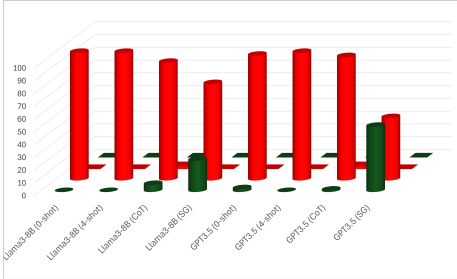
Due to a lack of space and the same behavior as explained for LLama3-70B in Section 5.3.2, we provide the plot for Llama3-8B and GPT3.5 here in Table 6 instead of the main paper.



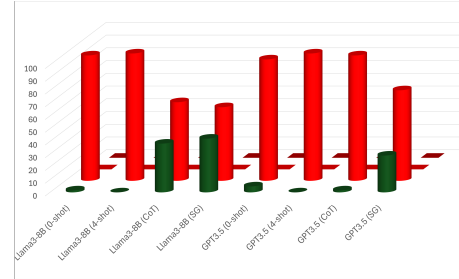
(a) Results of Cow Case in A-Split.



(b) Results of Car Case in A-Split.



(c) Results of Box Case in A-Split.



(d) Results of Pen Case in A-Split.

Figure 6: Red shows the wrong FoR identifications, and green shows the correct ones. The dark color is for relative FoRs, while the light color is for intrinsic FoRs. The round shape is for the external FoRs, while the square is for internal FoRs. The depth of the plots shows the four FoRs, i.e., external relative, external intrinsic, internal intrinsic, and internal relative, from front to back. This plot is the result of the rest of LLMs.