# FoREST: Frame of Reference Evaluation in Spatial Reasoning Tasks

Anonymous ACL submission


## Abstract

Spatial reasoning is a fundamental aspect of human intelligence. One key concept in spatial cognition is the Frame of Reference (FoR), which identifies the perspective of spatial expressions. Despite its significance, FoR has received limited attention in AI models that need spatial intelligence. There is a lack of dedicated benchmarks and in-depth evaluation of large language models (LLMs) in this area. To address this issue, we introduce the **F**rame **o**f **R**eference **E**valuation in **S**patial Reasoning **T**asks (FoREST) benchmark, designed to assess FoR comprehension in LLMs. We evaluate LLMs on answering questions that require FoR comprehension and layout generation in text-to-image models using FoREST. Our results reveal a notable performance gap across different FoR classes in various LLMs, affecting their ability to generate accurate layouts for text-to-image generation. This highlights critical shortcomings in FoR comprehension. To improve FoR understanding, we propose Spatial-Guided prompting, which improves LLMs' ability to extract essential spatial concepts. Our proposed method improves overall performance across spatial reasoning tasks.


## 1 Introduction

Spatial reasoning plays a significant role in human cognition and daily activities. It is also a crucial aspect in many AI problems, including language grounding (Zhang and 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., 2024), and image generation (Gokhale et al., 2023). One key concept in spatial reasoning is the Frame of Reference (FoR), which identifies the perspective of spatial expressions. FoR has been studied extensively in cognitive linguistics (Edmonds-Wathen, 2012; Vukovic and Williams, 2015). Levinson (2003) initially defines three FoR classes: *relative*, based on the observer's perspective; *intrinsic*, based on an inherent feature of the reference object; and *absolute*, using environmental cues like cardinal directions -See Figure 1. This framework was expanded by Tenbrink (2011) to create a more comprehensive framework, serving as the basis of our work. Understanding FoR is important for many applications, especially in embodied AI. In such applications, an agent must simultaneously comprehend multiple perspectives, including the one from the instruction giver and from the instruction follower, to communicate and perform tasks effectively. However, recent spatial evaluation benchmarks have largely overlooked FoR. For example, the text-based benchmarks Shi et al. (2022); Mirzaee and Kordjamshidi (2022); Rizvi et al. (2024) and text-to-images benchmarks (Gokhale et al., 2023; Huang et al., 2023; Cho et al., 2023a,b) assume a fixed perspective for all spatial expressions. This inherent bias limits situated spatial reasoning, restricting adaptability in interactive environments where perspectives can change.

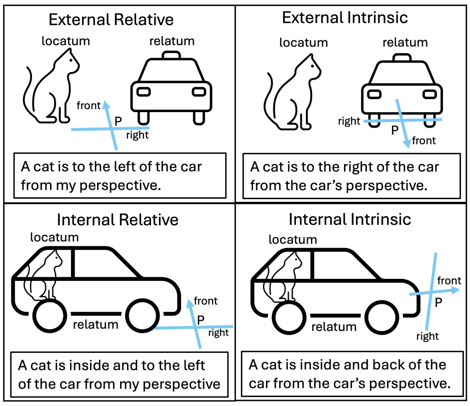



Figure 1: Illustration of FoR classes. The cat is the locatum, the car is the relatum, and the arrow indicates the perspective.

To systematically investigate the role of FoR in spatial understanding and create a new resource, that is, **F**rame **o**f **R**eference **E**valuation in **S**patial

Reasoning **T**asks (FoREST), FoREST is designed to evaluate models' ability to comprehend FoR from textual descriptions and extend this evaluation to grounding and visualization. Our benchmark includes spatial expressions with FoR ambiguity—where multiple FoRs may apply to the described situation—and spatial expressions with only a single valid FoR. This design allows evaluation of the models' understanding of FoR in both scenarios. We evaluate several LLMs in a QA setting that require FoR understanding and apply the FoR concept in text-to-image models. Our findings reveal performance differences across FoR classes and show that LLMs exhibit bias toward specific FoRs when handling ambiguous cases. This bias extends to layout-diffusion models, which rely on LLM-generated layouts in the image generation pipeline. To enhance FoR comprehension in LLMs, we propose Spatial-Guided prompting, which enables models to analyze and extract additional spatial information, including directional, topological, and distance relations. We demonstrate that incorporating spatial information improves question-answering and layout generation, ultimately enhancing text-to-image generation performance.

Our contribution[1] are summarized as follows, 1. We introduce the FoREST benchmark to systematically evaluate LLMs' FoR comprehension in a QA setting. 2. We analyze the impact of FoR information on text-to-image generation using multiple diffusion models. 3. We propose a prompting approach that generates spatial information, which can be incorporated into QA and layout diffusion to enhance performance.

## 2 Spatial Primitives

We review three semantic aspects of spatial information expressed in language: Spatial Roles, Spatial Relations, and Frame of Reference.

**Spatial Roles.** We focus on two main spatial roles (Kordjamshidi et al., 2010) of *Locatum*, and *Relatum*. The locatum is the object described in the spatial expression, while the relatum is the other object used to describe the position of the locatum. An example is *a cat is to the left of a dog*, where the *cat* is the locatum, and the *dog* is the relatum.

**Spatial Relations.** When dealing with spatial knowledge representation and reasoning, three main relations categories are often considered, that is, directional, topological, and distance (Hernández, 1994; Cohn and Renz, 2008; Kordjamshidi et al., 2011). *Directional* describes an object's direction based on specific coordinates. Examples of relations include left and right. *Topological* describes the containment between two objects, such as inside. *Distance* describes qualitative and quantitative relations between entities. Examples of qualitative are far, and quantitative are 3km.

**Spatial Frame of Reference.** We use four frames of references investigated in the cognitive linguistic studies (Tenbrink, 2011). These are defined based on the concept of *Perspective*, which is the origin of a coordinate system to determine the direction. The four frames of reference are defined as follows.

1. *External Intrinsic* describes a spatial relation from the relatum's perspective, where the relatum does not contain the locatum. The top-right image in Figure 1 illustrates this with the sentence, *A cat is to the right of the car from the car's perspective.*
2. *External Relative* describes a spatial relation from the observer's perspective. 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.*
3. *Internal Intrinsic* describes a spatial relation from the relatum's perspective, where the relatum contains the locatum. The bottom-right image in Figure 1 show this with the sentence, *A cat is inside and back of the car from the car's perspective.*
4. *Internal Relative* describes a spatial relation from the observer's perspective where the locatum is inside the relatum. The bottom-left image in Figure 1 show this FoR with the sentence, *A cat is inside and to the left of the car from my perspective.*

## 3 FoREST Dataset Construction

To systematically evaluate LLM on the frame of reference (FoR) recognition, we introduce the **F**rame **o**f **R**eference **E**valuation in **S**patial Reasoning **T**asks (FoREST) benchmark. Each instance in FoREST consists of a spatial context ($T$), a set of corresponding FoR ($FoR$) which is a subset of {*external relative*, *external intrinsic*, *internal intrinsic*, *internal relative*}, a set of questions and answers ($\{Q, A\}$), and a set of visualizations ($\{I\}$). An example of $T$ is *A cat is to the right of a dog. A dog is facing toward the camera.* The FoR of this expression is {*external intrinsic*, *external relative*}. A possible question-answer is $Q$ = *Based on the camera's perspective, where is the cat from the dog's position?*, $A$ = {left, right}. There is an ambiguity in the FoR for this expression. Thus, the

[1] code and dataset available at anonymous repository.

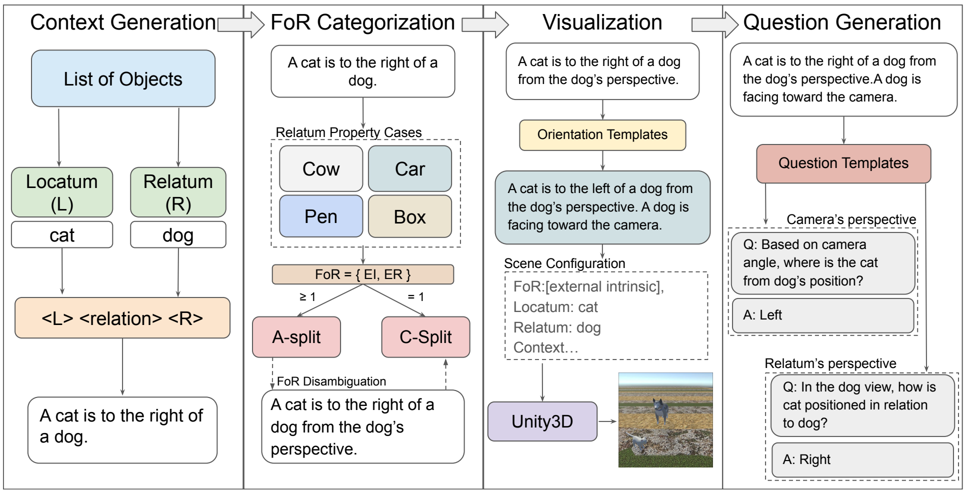



Figure 2: Pipelind for dataset creation, starting from selecting a locatum and relatum from available objects and then applying a spatial template to generate the spatial expression ($T$). FoRs are assigned based on the relatum's properties. $T$ is then categorized based on the number of FoRs. For example, *A cat is to the right of a dog* (with two possible FoRs: external intrinsic and external relative) belongs to the A-split. Then, its disambiguated version (A cat is to the right of a dog from the dog's perspective) is added to the C-split. Next, if applicable, a relatum's orientation is included for visualization and question generation. Finally, Unity3D generates scene configurations, and question-answer pairs are created from $T$.

answer will be *left* if the model assumes the external relative. Conversely, it will be *right* if the model assumes the external intrinsic. The visualization of this example is in Figure 2.

### 3.1 Context Generation

We select two distinct objects—a relatum ($R$) and a locatum ($L$)—from a set of 20 objects and apply them to a Spatial Relation template, <$L$> <*spatial relation*> <$R$> to generate the context $T$. FoRs for $T$ are determined based on the properties of the selected objects. Depending on the number of possible FoRs, $T$ is categorized as ambiguous (A-split), where multiple FoRs apply, or clear (C-split), where only one FoR is valid. We further augment the C-split with disambiguated spatial expressions derived from the A-split, as shown in Figure 2.

### 3.2 Categories based on Relatum Properties

Using the FoR classes in Section 2, we identified two key relatum properties contributing to FoR ambiguity. The first property is the relatum's intrinsic direction. It creates ambiguity between intrinsic and relative FoR since spatial relations can originate from 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, as spatial relations can refer to the inside and outside of the relatum. Based on these properties, we define four distinct cases: *Cow Case, Box Case, Car Case, and Pen Case.*

**Case 1: Cow Case**. In this case, the selected relatum has intrinsic directions but does not have the affordance as the container for the locatum. An obvious example is a cow, which should not be a container but has a front and back. In such cases, the relatum potentially provides a perspective for spatial relations. The applicable FoR classes are $FoR$ = {*external intrinsic*, *external relative*}. We augment the C-split with expressions of this case but include the perspective to resolve their ambiguity. To specify the perspective, we use predefined templates for augmenting clauses, such as *from {relatum}'s perspective* for *external intrinsic* or *from the camera's perspective* for *external relative*. For example, if the context is *A cat is to the right of the cow*, in the A-split. 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.** The relatum in this category has the property of being a container but lacks intrinsic directions, making the internal FoR applicable. An example is a box. The applicable FoR classes are $FoR$ = {*external relative*, *internal relative*}. To include their unambiguous counterparts in the C-split, we specify the topological relation to the expressions, $T$, by adding *inside* for *internal relative* and *outside* for *external relative* cases. For example, for the sentence *A cat is to the right of the* *box.*, the unambiguous $T$ with *internal relative* FoR 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.*

**Case 3: Car Case.** A relatum with an intrinsic direction and container affordance falls into this case, allowing all FoR classes. An obvious example is a car that can be a container with intrinsic directions. The applicable FoR classes are $FoR$ = { *external relative*, *external intrinsic*, *internal intrinsic*, *internal relative*}. To augment C-split with this case's disambiguated counterparts, we add perspective and topology information similar to the Cow and Box cases. An example expression for this case is *A person is in front of the car.* The four disambiguated 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.** In this case, the relatum lacks both the intrinsic direction and the affordance as a container. An obvious example is a pen with neither left/right nor the ability to be a container. Lacking these two properties, the created context has only one applicable FoR, $FoR$ = { *external relative*}. Therefore, we can categorize this case into both splits without any modification. An example of such a context is *The book is to the left of a pen.*

### 3.3 Context Visualization

In our visualization, complexity arises when the relatum has an intrinsic direction within the intrinsic FoR, as its orientation can complicate the spatial representation. For example, for visualizing *A cat is to the right of a dog from the dog's perspective.*, the cat can be placed in different coordinates based on the dog's orientation. To address this issue, we add a template sentence for each direction, such as *<relatum> is facing toward the camera*, to specify the relatum's orientation of all applicable $T$ for visualization and QA. For instance, *A cat is to the left of a dog.* becomes *A cat is to the left of a dog. The dog is facing toward the camera.* To avoid occlusion issues, we generate visualizations only for external FoRs, as one object may become invisible in internal FoR classes. We use only expressions in C-split since those have a unique FoR interpretation for visualization. We then create a scene configuration by applying a predefined template, as illustrated in Figure 2. Images are generated using the Unity 3D simulator (Juliani et al., 2020), producing four variations per expression $T$ with different backgrounds and object positions. Further details on the simulation process are in Appendix B.

### 3.4 Question-Answering Generation

We generate questions for all generated spatial expressions ($T$). Note that we include the relatum orientation for cases where the relatum has an intrinsic direction, as mentioned in the visualization. Our benchmark includes two types of questions. The first type asks for the spatial relation between two given objects from the camera's perspective, following predefined templates such as, *Based on the camera's perspective, where is the locatum relative to the relatum's position?* Template variations are made based on GPT4o. The second type of question queries the spatial relation from the relatum's perspective. This question type follows the same templates but replaces the camera with the relatum. The first type of question is generated for all $T$, while the second type is only generated for $T$ where the relatum has intrinsic direction and a perspective can be defined accordingly. Question templates are provided in Appendix B.3. Answers are determined based on the corresponding FoRs, the spatial relation in $T$, and the relatum's orientation when applicable.

## 4 Models and Tasks

The FoREST benchmark supports multiple tasks, including FoR identification, Question Answering (QA) that requires FoR comprehension, and Text-to-Image (T2I). This paper focuses on QA and T2I for a deeper evaluation of spatial reasoning. FoR identification experiments are provided in Appendix E.

### 4.1 Question-Answering (QA)

**Task.** This QA task evaluates LLMs' ability to adapt contextual perspectives across different FoRs. Both A and C splits are used in this task. The input is the context, consisting of a spatial expression $T$ and relatum orientation, if available, and a question $Q$ that queries the spatial relation from either an observer or the relatum's perspective. The output is a spatial relation $S$, restricted to {left, right, front, back}.

**Zero-shot baseline.** We call the LLM with instructions, a spatial context, $T$, and a question, $Q$, expecting a spatial relation as the response. The prompt instructs the model to answer the question with one of the candidate spatial relations without any explanations.

**Few-shot baseline.** We create four spatial expressions, each assigned to a single FoR class to prevent bias. Following the steps in Section 3.4, we generate a corresponding question and answer for each. These serve as examples in our few-shot prompting. The input to the model is instruction, example, spatial context, and the question.

**Chain-of-Thought baseline (Wei et al., 2023).** To create Chain-of-Thought (CoT) examples, we modify the prompt to require reasoning before answering. We manually crafted reasoning explanations with the necessary information for each example we used in the few-shot setting. The input to the model is instruction, CoT example, spatial context, and the question.

### 4.2 Text-To-Image (T2I)

**Task.** This task aims to determine the diffusion models' ability to consider FoR by evaluating their generated images. The input is a spatial expression, $T$, and the output is a generated image ($I$). We use the context from both C and A splits for this task.

**Stable Diffusion Models.** We use the stable diffusion models as the baseline for the T2I task. This model only needs the scene description as input.

**Layout Diffusion Models.** We evaluate the Layout Diffusion model, a more advanced T2I model operating in two phases: text-to-layout and layout-to-image. Given that LLMs can generate the bounding box layout (Cho et al., 2023b), we provide them with instructions and $T$ to create the layout. The layout consists of bounding box coordinates for each object in the format of {object: $[x, y, w, h]$}, where $x$ and $y$ denote the starting point and $h$ and $w$ denote the height and width. The bounding box coordinates and $T$ are then passed to the layout-to-image model to produce the final image, $I$.

### 4.3 Spatial-Guide Prompting

We hypothesize that the spatial relation types and FoR classes explained in Section 2 can improve question-answering and layout generation. For instance, the *external intrinsic* FoR emphasizes that spatial relations originate from the relatum's perspective. To leverage this, we propose Spatial-Guided (SG) prompting, an additional step applied before QA or layout generation steps. This step extracts spatial information, including direction, topology, distance as well as the FoR from spatial expression $T$. The extracted information will serve as supplementary for guiding LLMs in QA and layout generation. We manually craft four examples covering these aspects. First, we specify the perspective for *directional relations*, e.g., *left* relative to the observer, to distinguish intrinsic from relative FoR. Next, we indicate whether the locatum is inside or outside the relatum for *topological relations* to differentiate internal from external FoR. Lastly, we provide an estimated quantitative distance to support topological and directional relation identification, such as *far*. These examples are then provided as a few-shot example for the model to extracted information automatically.

## 5 Experimental Results

### 5.1 Evaluation Metrics

**QA.** We report an accuracy measure defined as follows. Since the questions can have multiple correct answers, specifically in A-split, as explained in Section 3, the prediction is correct if it matches any valid answer. Additionally, we report the model's bias distribution when FoR ambiguity exists. $I\%$ is the percentage of correct answers when assuming an intrinsic FoR, while $R\%$ is this percentage with a relative FoR assumption. Note that cases where both FoR assumptions lead to the same answer are excluded from these calculations.

**T2I.** We adopt *spatialEval* (Cho et al., 2023b) approach for evaluating T2I spatial ability. However, we modify it to account for FoR. We convert all relations to a camera perspective before passing them to spatialEval, which assumes this viewpoint. Accuracy is determined by comparing the bounding box and depth map of the relatum and locatum. For FoR ambiguity, a generated image is correct if it aligns with at least one valid FoR interpretation. We report results using VISOR$_{cond}$ and VISOR$_{uncond}$ (Gokhale et al., 2023), metrics for assessing T2I spatial understanding. VISOR$_{cond}$ evaluates spatial relations only when both objects appear correctly, aligning with our focus on spatial reasoning rather than object creation. In contrast, VISOR$_{uncond}$ evaluates the overall performance, including object creation errors.

### 5.2 Experimental Setting

**QA.** We use Llama3-70B (Llama, 2024), Qwen2-72B (Qwen Team, 2024), and GPT-4o (*gpt-4o-2024-11-20*) (OpenAI, 2024) as the backbones for

| Model | Camera perspective | | | | | | | | | Relatum perspective | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cow | | | Car | | | Box | Pen | Avg. | Cow | | | Car | | | Avg. |
| | R% | I% | Acc. | R% | I% | Acc. | Acc. | Acc. | Acc. | R% | I% | Acc. | R% | I% | Acc. | Acc. |
| Llama3-70B (1) | 48.1 | **51.5** | 62.5 | **58.0** | 41.6 | 65.5 | 73.3 | 72.5 | 64.3 | **61.0** | 38.7 | 62.1 | **51.8** | 47.9 | 61.8 | 62.1 |
| Llama3-70B (2) | 49.1 | **50.5** | 62.2 | **52.2** | 47.4 | 64.7 | 85.8 | 85.5 | 65.8 | **59.6** | 40.1 | 57.1 | **55.5** | 44.2 | 61.8 | 57.7 |
| Llama3-70B (3) | 49.4 | **50.3** | 80.7 | 49.4 | **50.3** | 79.6 | 95.8 | 94.9 | 82.6 | **60.8** | 39.0 | 77.2 | **55.1** | 44.6 | 80.9 | 77.7 |
| Llama3-70B (4) | **59.4** | 40.2 | 73.6 | **57.9** | 41.7 | 74.8 | 100.0 | 100.0 | 77.5 | **60.6** | 39.1 | 65.7 | **56.0** | 43.7 | 67.7 | 66.0 |
| Qwen2-72B (1) | **96.6** | 2.9 | 95.6 | **95.9** | 3.6 | 95.0 | 100.0 | 100.0 | 96.1 | 8.8 | **90.6** | 79.3 | 7.8 | **91.7** | 83.6 | 79.9 |
| Qwen2-72B (2) | **89.0** | 10.5 | 84.4 | **85.6** | 13.9 | 85.5 | 100.0 | 100.0 | 86.8 | 17.7 | **81.8** | 78.3 | 10.4 | **89.1** | 86.3 | 79.4 |
| Qwen2-72B (3) | **67.2** | 32.4 | 88.6 | **62.0** | 37.6 | 83.4 | 100.0 | 100.0 | 89.6 | 21.3 | **78.3** | 85.5 | 22.7 | **76.9** | 83.6 | 85.2 |
| Qwen2-72B (4) | **93.0** | 6.5 | 90.1 | **94.6** | 4.9 | 93.3 | 100.0 | 98.6 | 91.7 | 8.2 | **91.2** | 86.0 | 10.5 | **89.0** | 87.4 | 86.2 |
| GPT-4o (1) | **84.3** | 15.3 | 94.5 | **88.5** | 11.0 | 97.3 | 99.2 | 99.8 | 95.6 | 21.6 | **78.0** | 91.6 | 16.1 | **83.5** | 90.5 | 91.4 |
| GPT-4o (2) | **69.0** | 30.6 | 76.6 | **80.3** | 19.2 | 89.5 | 100.0 | 100.0 | 81.5 | 29.0 | **70.5** | 74.7 | 30.9 | **68.7** | 77.5 | 75.1 |
| GPT-4o (3) | 41.5 | **58.3** | 92.3 | 38.2 | **61.6** | 91.0 | 100.0 | 99.8 | 93.2 | 33.9 | **65.8** | 93.9 | 32.0 | **67.6** | 93.9 | 93.9 |
| GPT-4o(4) | 26.0 | **73.9** | 79.2 | 27.7 | **72.1** | 79.4 | 96.7 | 94.3 | 81.4 | 16.2 | **83.4** | 95.5 | 19.2 | **80.4** | 94.8 | 95.4 |

Table 1: QA accuracy in the A-Split across various LLMs. R% and I% represent the percentage the model assumes relative or intrinsic FoR for ambiguous expression explained in Section 5.1. Acc is the accuracy, and Avg is the micro-average of accuracy. (1): 0-shot, (2): 4-shot, (3): CoT, and (4): SG + CoT.

| Model | Camera perspective | | | | | Relatum perspective | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ER (CP) | EI (RP) | II (RP) | IR (CP) | Avg. | ER (CP) | EI (RP) | II (RP) | IR (CP) | Avg. |
| Llama3-70B (0-shot) | 44.8 | 38.4 | 39.7 | 54.4 | 42.6 | 42.2 | 47.1 | 62.5 | 34.4 | 45.1 |
| Llama3-70B (4-shot) | 43.0 | 40.0 | 39.1 | 47.3 | 41.9 | 41.8 | 60.9 | 77.7 | 35.2 | 52.0 |
| Llama3-70B (CoT) | 57.8 | 46.1 | 44.7 | 46.0 | 51.5 | **55.5** | 56.8 | 71.5 | 49.0 | 56.6 |
| Llama3-70B (SG + CoT) | 47.6 | 42.9 | 50.0 | 35.6 | 45.0 | 55.4 | 64.5 | 75.0 | 47.1 | 60.1 |
| Qwen2-72B (0-shot) | 94.5 | 35.2 | 31.8 | 93.2 | 66.9 | 28.7 | 89.3 | 93.6 | 23.8 | 59.0 |
| Qwen2-72B (4-shot) | 90.2 | 39.5 | 39.1 | 68.5 | 65.3 | 33.5 | 92.1 | 94.0 | 29.5 | 62.7 |
| Qwen2-72B (CoT) | 81.4 | 57.4 | 58.6 | 62.5 | 69.1 | 39.5 | 83.7 | 85.2 | 37.7 | 61.6 |
| Qwen2-72B (SG + CoT) | 97.6 | 42.5 | 31.3 | 93.8 | 71.4 | 42.8 | 86.6 | 92.0 | 34.0 | 64.5 |
| GPT-4o (0-shot) | 79.7 | 45.1 | 39.5 | 90.2 | 64.2 | 46.9 | 88.5 | 98.2 | 34.8 | 67.5 |
| GPT-4o (4-shot) | 68.0 | 52.6 | 60.7 | 74.1 | 61.8 | 44.9 | **98.2** | **100.0** | 37.5 | 71.2 |
| GPT-4o (CoT) | 81.7 | **76.1** | **82.4** | 71.5 | 78.8 | 53.0 | 91.1 | 90.6 | **50.8** | 71.9 |
| GPT-4o (SG + CoT) | **97.9** | 72.2 | 72.7 | **93.4** | **85.8** | 48.9 | 96.3 | 95.9 | 36.1 | **71.8** |

Table 2: QA accuracy in the C-Split across various LLMs. ER, EI, II, and IR denote external relative, external intrinsic, internal intrinsic, and internal relative FoRs. Avg represents the micro-average accuracy. CP refers to context with camera perspective, while RP denotes context with relatum perspective.

prompt engineering. To ensure reproducibility, we set the temperature of all models to 0. For all models, we apply *zero-shot*, *few-shot*, *CoT*, and our proposed prompting with CoT (SG+CoT).

**T2I.** We select Stable Diffusion SD-1.5 and SD-2.1 (Rombach et al., 2021) as our stable diffusion models and GLIGEN(Li et al., 2023) as the layout-to-image backbone. For translating spatial descriptions into textual bounding box information, we use Llama3-8B and Llama3-70B, as detailed in Section 4.2. The same LLMs are used to generate spatial information for SG prompting. We generate four images to compute the VISOR score following (Gokhale et al., 2023) Inference steps for all T2I models are set to 50. For the evaluation modules, we select grounding DINO (Liu et al., 2025) for object detection and DPT (Ranftl et al., 2021) for depth mapping, following VPEval (Cho et al., 2023b). The experiments were conducted on an $A6000$ GPU, totaling approximately 300 GPU hours.

### 5.3 Results

**RQ1. What is the bias of the LLMs for the ambiguous FoR?** Table 1 presents the QA results for the A-split. Ideally, a model that correctly extracts the spatial relation without considering perspective should achieve 100% accuracy, as the context lacks a fixed perspective. However, this ideal model is not the focus of our work. We aim to assess model bias by measuring how often LLMs adopt a specific perspective when answering. In the Cow and Pen case, relatum properties do not introduce FoR ambiguity in directional relations, making the task pure extraction rather than reasoning. Thus, we focus on the $I\%$ and $R\%$ of the Cow and Car cases, which best reflect LLMs' bias. Qwen2 achieves around 80% accuracy across all experiments by selecting spatial expressions directly from the context, suggesting it may disregard the question's perspective. GPT-4o shows similar bias in 0-shot and 4-shot settings but shifts toward intrinsic interpretation

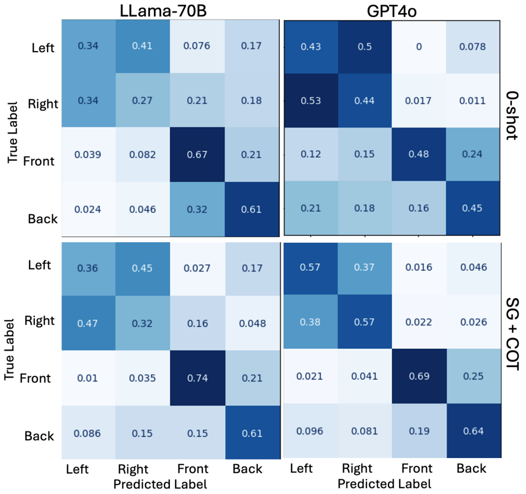



Figure 3: Confusion matrices of spatial relation answers when Llama3 and GPT-4o are required to adapt FoR in the 0-shot and (SG+CoT) settings.

with CoT. This bias reduces accuracy in camera-perspective questions from 93.2% to 81.4%, where FoR adaptation is more challenging than relation extraction. Llama3-70B lacks a strong preference, balancing assumptions but slightly favoring relative FoR. This uncertainty lowers performance, requiring more reasoning to reach the correct answer. In summary, Qwen2 achieves higher accuracy by focusing on relation extraction without considering FoR reasoning, while other models attempt reasoning but struggle to reach correct conclusions, leading to lower performance.

**RQ2. Can the model adapt FoR when answering the questions?** To address this research question, we analyze QA that required FoR comprehension results in C-Split from Table 2. Note that the context and question in these tasks explicitly indicate a perspective. The results indicate that LLMs struggle with FoR conversion, particularly when the question has relatum and the context has camera perspectives, achieving only up to 55.5% accuracy. We further demonstrate how Llama3 and GPT-4o adapt FoR using the confusion matrix in Figure 3. Our findings reveal that pure-text LLM (Llama3) has confusion between left and right. Humans typically reverse front and back while preserving left and right when describing the spatial relation from perspective. However, Llama3 incorrectly reverses left and right, leading to poor adaptation to the camera perspective. In contrast, very large multimodal-language models like GPT-4o follow the expected pattern, as observed by Zhang et al. 2025. While our GPT-4o results suggest some ability to convert the relatum's perspective into the camera's with in-context learning (72% accuracy), the reverse transformation in the textual domain remains challenging (53% accuracy). This difficulty persists when converting spatial relations from the camera perspective from images to the relatum's perspective as observed in Zhang et al. 2025.

**RQ3. How can an explicit FoR identification help spatial reasoning in QA?** We compare CoT and CoT+SG results to assess how explicit FoR identification affects LLMs' spatial reasoning in QA. Based on C-Split results (Table 3), incorporating SG encourages the model to identify the perspective in a given expression leading to improvement in ranging from 2.9% to 30% in cases where the context and question share the same perspective. These cases are easier as the models do not need FoR adaptation. The only exception is Llama3 for questions with the camera's perspective, where explicit FoR identification negatively impacts performance. Since Llama is performing poorly by a large margin compared to other LLMs (see 0-shot and 4-shot results), the SG prompting seems not helpful for this extreme case. We should note that among our selected LLMs, Llama3 is the only one not trained with visual information; we speculate this can be a factor in LLMs' understanding of FoR. Limitations in scenarios requiring perspective adaptation remain, even though SG helps them identify the correct perspective. SG identification results are reported in the Appendix E. This limitation is also evident in A-Split results (Table 1), where models only show significant improvement when SG aligns their preference toward the same perspective as the question, see Qwen2-72B and GPT-4o. Overall, incorporating FoR identification can help the model account perspective when performing spatial reasoning (see the Avg column for SG+CoT in Table 3).

**RQ4. How can explicit FoR identification help spatial reasoning in visualization?** We evaluate SG layout diffusion to assess the impact of incorporating FoR in image generation. We focus on VISOR$_{cond}$ metric, as it better reflects the model's spatial understanding than the overall performance measured by VISOR$_{uncond}$, which is reported in Appendix C due to space limitation. Table 3 shows that adding spatial information and FoR classes (SG+GLIGEN) improves performance across all splits compared to the baseline models (GLIGEN). SG improved the model's performance when expressions can be interpreted as relative FoR. These

| Model | VISOR(%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A-Split | | | | | C-Split | | |
| | cond (I) | | | cond (R) | cond (avg) | cond (I) | cond (R) | cond (avg) |
| | EI FoR | ER FoR | all | | | | | |
| SD-1.5 | 51.11 | 21.61 | 72.72 | 48.95 | 68.72 | 53.92 | 53.77 | 53.83 |
| SD-2.1 | 57.97 | 21.49 | 79.46 | 54.10 | 75.39 | **60.06** | 59.64 | 59.83 |
| Llama3-8B + GLIGEN | 53.67 | 25.78 | 79.45 | 66.08 | 77.38 | 57.51 | 65.98 | 62.12 |
| Llama3-70B + GLIGEN | 54.49 | 29.45 | 83.94 | 68.68 | 81.43 | 56.47 | 69.53 | 63.49 |
| Llama3-8B + SG + GLIGEN (Our) | 57.46 | 27.96 | 85.42 | **71.14** | 83.17 | 58.84 | **70.36** | **65.15** |
| Llama3-70B + SG + GLIGEN (Our) | 56.54 | 30.59 | **87.13** | 66.56 | **83.75** | 56.77 | 70.04 | 64.06 |

Table 3: VISOR$_{cond}$ score on the A and C splits 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. EI and ER FoR represent the generated image considered corrected by EI or ER FoR

results align with the QA results shown in Table 1 indicating that *Llama3 prefers relative FoR if dealing with the camera's perspective*. In contrast, baseline diffusion models (SD-1.5 and SD-2.1) perform better for intrinsic FoR even though GLIGEN is based on SD-2.1. This outcome might be due to GLIGEN's reliance on bounding boxes for generating spatial configurations, which makes it struggle with intrinsic FoR due to the absence of object properties and orientation. Nevertheless, incorporating FoR information via SG-prompting improves performance across all FoR classes despite this specific bias. We provide further analysis on SG for the layout generation in Appendix D.

## 6 Related Work

**Frame of Reference in Cognitive Studies.** The concept of the frame of reference in cognitive studies was introduced by Levinson 2003 and later expanded with more diverse spatial relations (Tenbrink, 2011). Subsequent research investigated the human preferences for specific FoR classes (Edmonds-Wathen, 2012; Vukovic and Williams, 2015; Shusterman and Li, 2016; Ruotolo et al., 2016). For instance, Ruotolo et al. 2016 examined how FoR influences scene memorization and description under time constraints. Their study found that participants performed better when spatial relations were based on their position rather than external objects, highlighting a distinction between relative and intrinsic FoR.

**Frame of Reference in AI.** Several benchmarks have been developed to evaluate the spatial understanding of AI models in multiple modalities; for instance, textual QA (Shi et al., 2022; Mirzaee and Kordjamshidi, 2022; Rizvi et al., 2024), and text-to-image (T2I) benchmarks (Gokhale et al., 2023; Huang et al., 2023; Cho et al., 2023a,b). However, most of these benchmarks overlook the frame of reference (FoR), assuming a single FoR for all instances despite its significance in cognitive studies. Recent works in vision-language research are beginning to address this problem. For instance, Liu et al. 2023 examines FoR's impact on visual question-answering but focuses only on limited FoR categories. Our work covers more diverse FoRs. Zhang et al. 2025 examine FoR ambiguity and understanding in vision-language models. Their study evaluates spatial relations in the visual input under different FoR questions, relying on images from a camera perspective. In our evaluation, we vary input perspectives to explore FoR's impact across various spatial relations.

## 7 Conclusion

Given the significance of spatial reasoning in AI applications, we introduce **F**rame **o**f **R**eference **E**valuation in **S**patial Reasoning **T**asks (FoREST) benchmark to evaluate Frame of Reference comprehension in textual spatial expressions via question-answering and grounding in visual modality by diffusion models. Based on this benchmark, our results reveal notable differences in FoR comprehension across LLMs and their struggle with questions that require adaptation between multiple FoRs. Moreover, the bias in FoR interpretations impacts the layout generation with LLMs for text-to-image models. To improve FoR comprehension, we propose Spatial-Guided prompting, which first generates spatial relation's topological, distal, and directional type information in addition to FoR and includes this information in downstream task prompting. Employing SG improves the overall performance in QA tasks requiring FoR understanding and text-to-image generation.

## 8 Limitations

While we analyze LLMs' shortcomings, our benchmark only highlights areas for improvement, not harming the model. The trustworthiness and reliability of the LLMs are still a research challenge. Our analysis is confined to the spatial reasoning domain and does not account for biases related to gender or race. However, we acknowledge that linguistic and cultural variations in spatial expression are not considered, as our study focuses solely on English. Extending this work to multiple languages could reveal important differences in FoR adaptation. Our analysis is still limited to the synthetic environment. Future research should consider the broader implications of the frame of reference of spatial reasoning in real-world applications. Additionally, our experiments require substantial GPU resources, limiting the selection of LLMs and constraining the feasibility of testing larger models. The computational demands also pose accessibility challenges for researchers with limited resources. We find no ethical concerns in our methodology or results, as our study does not involve human subjects or sensitive data.

## References


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)*, pages 14455–14465.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023a. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023b. Visual programming for step-by-step text-to-image generation and evaluation. In *Advances in Neural Information Processing Systems*, volume 36, pages 6048–6069. Curran Associates, Inc.

Anthony G. Cohn and Jochen Renz. 2008. Chapter 13 qualitative spatial representation and reasoning. In Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter, editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 551–596. Elsevier.

Cris Edmonds-Wathen. 2012. False friends in the multilingual mathematics classroom. In *:*, pages 5857–5866.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2023. Benchmarking spatial relationships in text-to-image generation. *Preprint*, arXiv:2212.10015.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 2024. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324.

Daniel Hernández, editor. 1994. *Reasoning with qualitative representations*, pages 55–103. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. 2020. Unity: A general platform for intelligent agents. *Preprint*, arXiv:1809.02627.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2010. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing*, 8:1–36.

Stephen C. Levinson. 2003. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. *CVPR*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

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. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55, Cham. Springer Nature Switzerland.

Llama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alibaba Group Qwen Team. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168.

Md Imbesat Rizvi, Xiaodan Zhu, and Iryna Gurevych. 2024. SpaRC and SpaRP: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4750–4767, Bangkok, Thailand. Association for Computational Linguistics.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Francesco Ruotolo, Tina Iachini, Gennaro Ruggiero, Ineke J. M. van der Ham, and Albert Postma. 2016. Frames of reference and categorical/coordinate spatial relations in a “what was where” task. *Experimental Brain Research*, 234(9):2687–2696.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11321–11329.

Anna Shusterman and Peggy Li. 2016. Frames of reference in spatial language acquisition. *Cognitive Psychology*, 88:115–161.

Thora Tenbrink. 2011. Reference frames of space and time in language. *Journal of Pragmatics*, 43(3):704–722. The Language of Space and Time.

Nikola Vukovic and John N. Williams. 2015. Individual differences in spatial cognition influence mental simulation of language. *Cognition*, 142:110–122.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*.

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

Yue Zhang and Parisa Kordjamshidi. 2022. LOViS: Learning orientation and visual signals for vision and language navigation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5745–5754, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*.

## A Dataset Statistics

The FoREST dataset statistic is provided in the Table 4.

## B Details Creation of FoREST dataset

We define the nine categories of objects selected in our dataset as indicated below in Table 5. 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 3.2. 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-spilt. 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 6. 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.

| Case | A-Split | A-Split with orientation | FoR class | C-Spilt | C-split with orientation |
|---|---|---|---|---|---|
| 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 4: Dataset Statistic of FoREST dataset.

| 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 5: All selected objects with two properties: intrinsic direction, affordance of being container

### 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 assert store[2]. All of them are free and available for download. All of the 3D models used are shown in Figure 4.

### B.3 Textual templates

All the templates used to create FoREST are given in Table 6.

[2] https://assetstore.unity.com

## C VISOR uncond Score

VISOR$_{uncond}$ provides the overall spatial relation score, including images with object generation errors. Since it is less focused on evaluating spatial interpretation than VISOR$_{cond}$, which assesses explicitly the text-to-image model's spatial reasoning, we report VISOR$_{uncond}$ results here in the Table 7 rather than in the main paper. The results are similar to the pattern observed in VISOR$_{uncond}$ that the based models(SD-1.5 and SD-2.1) perform better in the relative frame of reference, while the layout-to-image models, i.e., GLIGEN, are better in the intrinsic frame of reference.

## D Analyze the improvements in SG-prompting for T2I.

To further explain improvements of SG-prompting in T2I task, 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 is Table 8, our SG prompting improved 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 VISOR$_{cond}$ in Table 3. We report the layout$_{cond}$ score for a consistent evaluation in the same table. Layout$_{cond}$ shows that Llama3-8B improves within the same evaluation subset with VISOR$_{cond}$. Overall, by incorporating FoR information through SG layout diffusion, Llama3 generates better spatial configurations, enhancing

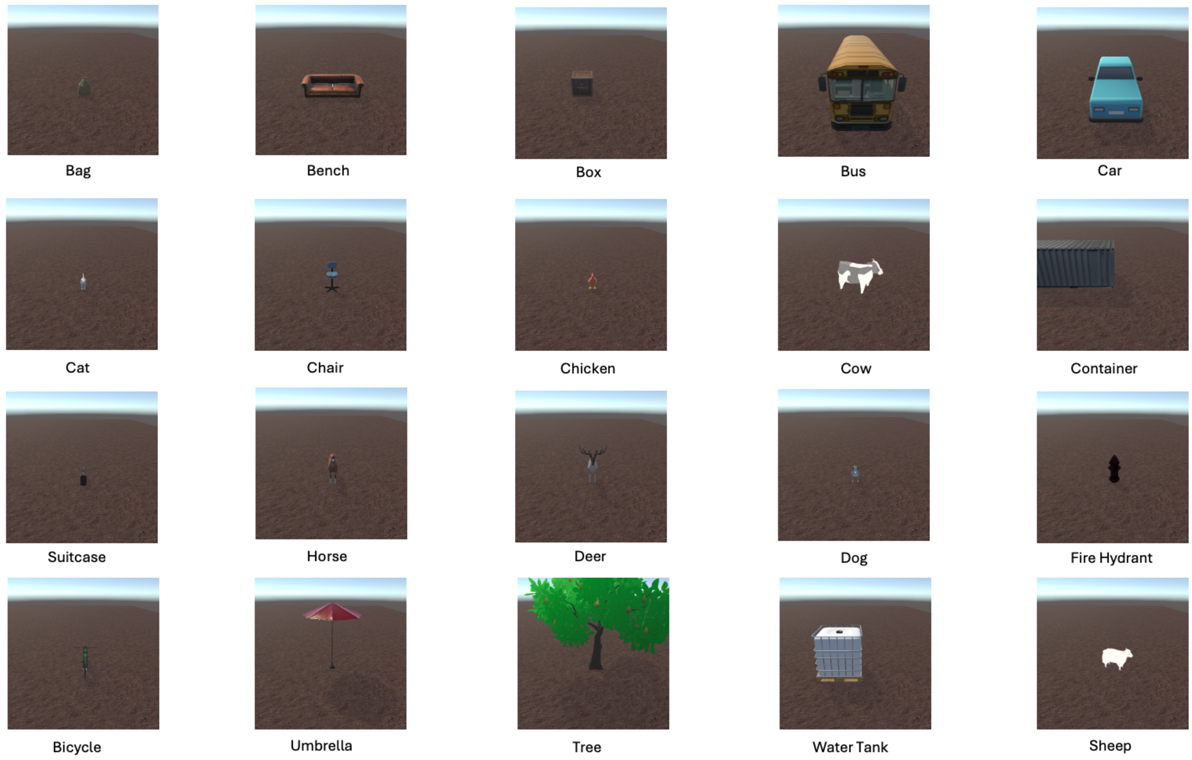



Figure 4: All 3d models used to generate visualizations for FoREST.

image generation performance.

## E Frame of Reference Identification

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

### E.1 Experimental Setting

**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.,

| | |
|---|---|
| Spatial Relation Templates | {locatum} is in front of {relatum}<br>{locatum} is on the left of {relatum}<br>{locatum} is to the left of {relatum}<br>{locatum} is behind of {relatum}<br>{locatum} is back of {relatum}<br>{locatum} is on the right of {relatum}<br>{locatum} is to the right of {relatum} |
| Topology Templates | within {relatum}<br>and inside {relatum}<br>and outside of {relatum} |
| Perspective Templates | from {relatum}'s view<br>relative to {relatum}<br>from {relatum}'s perspective<br>from my perspective<br>from my point of view<br>relative to observer |
| Orientation Templates | {relatum} facing toward that camera<br>{relatum}is facing away from the camera.<br>{relatum} facing left relative to the camera<br>{relatum} facing right relative to the camera |
| Question Templates | In the camera view, how is {locatum} positioned in relation to {relatum}?<br>Based on the camera perspective, where is the {locatum} from the {relatum}'s position?<br>From the camera perspective, what is the relation of the{locatum} to the {relatum}?<br>Looking through the camera perspective, how does {locatum} appear to be oriented relative to {relatum}'s position?<br>Based on the camera angle, where is {locatum} located with respect to {relatum}'s location? |

Table 6: All templates used to create FoREST dataset.

| Model | VISOR(%) | | | | | |
|---|---|---|---|---|---|---|
| | uncond (I) | uncond (R) | uncond (avg) | uncond (I) | uncond (R) | uncond (avg) |
| | A-Split | | | C-Split | | |
| SD-1.5 | 45.43 | 33.22 | 43.51 | 35.06 | 35.68 | 35.40 |
| SD-2.1 | **62.87** | 43.90 | **59.89** | **45.98** | 46.59 | **46.31** |
| 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 | **47.44** | 57.12 | 38.23 | **48.62** | 43.86 |

Table 7: VISOR$_{uncond}$ score on the A-Split and 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.

| Model | Layout | Layout$_{cond}$ |
|---|---|---|
| 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 8: 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 Layout$_{cond}$ uses the same testing examples as VISOR$_{cond}$.

| Model | inherently clear | | require template | |
|---|---|---|---|---|
| | CoT | SG | CoT | SG |
| Llama3-70B | 19.84 | 44.64 (↑ 24.80) | 76.72 | 87.39 (↑ 10.67) |
| Qwen2-72B | 58.20 | 84.22 (↑ 26.02) | 88.36 | 93.86 (↑ 10.67) |
| GPT-4o | 12.50 | 29.17 (↑ 16.67) | 87.73 | 90.74 (↑ 3.01) |

Table 9: The comparison between CoT and SG prompting in C-split separated by inherently clear / required template to be clear.

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 $F$.

## E.2 Evaluation Metrics

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.

## E.3 Results

### E.3.1 FoR Inherently Bias in LLMs

**C-spilt.** The *zero-shot* setting reflects the LLMs' inherent bias in identifying FoR. Table 10 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 prefer specific FoR classes. Notably, Gemme2-9B achieves a near-perfect accuracy on external relative FoR but performs poorly on other classes, especially external intrinsic, indicating a notable bias towards external relative. In contrast, GPT4o and Qwen2-72B perform exceptionally in both intrinsic FoR classes. However, they perform poorly in the relative FoRs.

**A-spilt.** We examine the FoR bias in the A-split.

| Model | A-split | C-Split | | | | |
|---|---|---|---|---|---|---|
| | | ER-C-Split | EC-Split | IC-Split | IR-C-Split | Avg. |
| Gemma2-9B (0-shot) | 94.17 | **94.24** | 35.98 | 53.91 | 57.66 | 60.45 |
| Gemma2-9B (4-shot) | 59.58 | 55.89(↓ 38.34) | 72.61(↑ 36.63) | 74.22(↑ 20.31) | 54.44(↓ 3.23) | 64.29(↑ 3.84) |
| Gemma2-9B (CoT) | 60.49 | 60.49(↓ 33.74) | 60.54(↑ 24.57) | 87.50(↑ 33.59) | 54.03(↓ 3.63) | 65.64(↑ 5.20) |
| Gemma2-9B (SG)(Our) | 72.67 | 65.87(↓ 28.37) | 65.54(↑ 29.57) | 53.12(↓ 0.78) | **95.97**(↑ 38.31) | 70.13(↑ 9.68) |
| llama3-8B (0-shot) | 60.21 | 32.20 | 90.11 | 75.78 | 0.00 | 49.52 |
| llama3-8B (4-shot) | 60.14 | 47.77(↑ 15.58) | 54.35(↓ 35.76) | 100.00(↑ 24.22) | 41.13(↑ 41.13) | 60.81(↑ 11.29) |
| llama3-8B (CoT) | 61.32 | 61.06(↑ 28.86) | 97.28(↑ 7.17) | 100.00(↑ 24.22) | 36.29(↑ 36.29) | 73.66(↑ 24.14) |
| llama3-8B (SG) (Our) | 62.95 | 63.29(↑ 31.09) | 94.57(↑ 4.46) | 100.00(↑ 24.22) | 43.55(↑ 43.55) | 75.35(↑ 25.83) |
| llama3-70B (0-shot) | 84.23 | 74.08 | 9.57 | 92.19 | 68.55 | 61.10 |
| llama3-70B (4-shot) | 78.47 | 81.81(↑ 7.72) | 64.89(↑ 55.33) | 100.00(↑ 7.81) | 75.81(↑ 7.26) | 80.63(↑ 19.53) |
| llama3-70B (CoT) | 69.11 | 72.05(↓ 2.03) | 97.07(↑ 87.50) | 100.00(↑ 7.81) | 79.44(↑ 10.89) | 87.14(↑ 26.04) |
| llama3-70B (SG) (Our) | 76.50 | 78.21(↑ 4.12) | 97.61(↑ 88.04) | 100.00(↑ 7.81) | 72.18(↑ 3.63) | 87.00(↑ 25.90) |
| Qwen2-7B (0-shot) | 83.64 | 79.97 | 59.24 | 77.34 | 40.73 | 64.32 |
| Qwen2-7B (4-shot) | 61.12 | 50.52(↓ 29.45) | 65.76(↑ 6.52) | 93.75(↑ 16.41) | 56.05(↑ 15.32) | 66.52(↑ 2.20) |
| Qwen2-7B (CoT) | 72.12 | 70.81(↓ 9.16) | 63.80(↑ 4.57) | 99.22(↑ 21.88) | 51.61(↑ 10.89) | 71.36(↑ 7.04) |
| Qwen2-7B (SG) | 70.61 | 68.00(↓ 11.98) | 71.20(↑ 11.96) | 88.28(↑ 10.94) | 57.26(↑ 16.53) | 71.18(↑ 6.86) |
| Qwen2-72B (0-shot) | 64.46 | 62.70 | 100.00 | 100.00 | 39.11 | 75.45 |
| Qwen2-72B (4-shot) | 79.12 | 78.73(↑ 16.03) | 99.35(↓ 0.65) | 87.50(↓ 12.50) | 87.10(↑ 47.98) | 88.17(↑ 12.72) |
| Qwen2-72B (CoT) | 88.54 | 88.87(↑ 26.18) | 89.57(↓ 10.43) | 93.75(↓ 6.25) | 83.47(↑ 44.35) | 88.91(↑ 13.46) |
| Qwen2-72B (SG) | 90.51 | 90.18(↑ 27.49) | 93.26(↓ 6.74) | 98.44(↓ 1.56) | 85.08(↑ 45.97) | 91.74(↑ 16.29) |
| GPT3.5 (0-shot) | 83.11 | 88.15 | 17.50 | 70.31 | 41.13 | 54.27 |
| GPT3.5 (4-shot) | 61.25 | 48.95(↓ 39.20) | 62.72(↑ 45.22) | 100.00(↑ 29.69) | 28.63(↓ 12.50) | 60.07(↑ 5.80) |
| GPT3.5 (CoT) | 66.55 | 66.62(↓ 21.53) | 96.85(↑ 79.35) | 100.00(↑ 29.69) | 50.81(↑ 9.68) | 78.57(↑ 24.30) |
| GPT3.5 (SG) (Our) | 70.61 | 73.30(↓ 14.86) | 92.93(↑ 75.43) | 99.22(↑ 28.91) | 49.19(↑ 8.06) | 78.66(↑ 24.39) |
| GPT4o (0-shot) | 73.82 | 71.27 | 98.80 | 100.00 | 70.56 | 85.16 |
| GPT4o (4-shot) | 66.23 | 67.87(↓ 3.40) | 98.70(↓ 0.11) | 100.00(↑ 0.00) | 78.63(↑ 8.06) | 86.30(↑ 1.14) |
| GPT4o (CoT) | 72.44 | 72.77(↑ 1.51) | 100.00(↑ 1.20) | 100.00(↑ 0.00) | 73.79(↑ 3.23) | 86.64(↑ 1.48) |
| GPT4o (SG) (Our) | 76.44 | 74.67(↑ 3.40) | 97.72(↓ 1.09) | 100.00(↑ 0.00) | 68.55(↓ 2.02) | 85.23(↑ 0.08) |

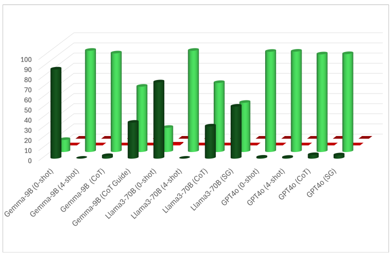
Table 10: Accuracy results report from FoR Identification with LLMs. The correct prediction is one of the valid FoR classes for the given spatial expression. All FoR classes are external relative (ER), external intrinsic (EI), internal intrinsic (II), and internal relative (IR).

Based on the results in Table 10, we plotted the top-3 models' results (Gemma2-9B, Llama3-70B, and GPT4o) for a more precise analysis in Figures 5. 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.

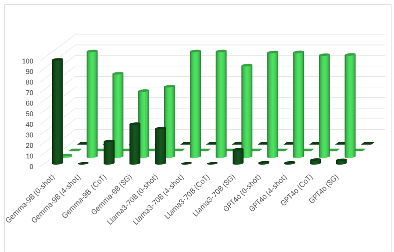
### E.3.2 Behavior with ICL variations

**C-spilt.** We evaluate the models' behavior under various in-context learning (ICL) methods. As observed in Table 10, the *few-shot* method improves the performance of the *zero-shot* method across multiple LLMs by reducing their original bias toward specific classes. Reducing the bias, however, lowers the performance in some cases, such as the performance of Gemma 2 in ER class. One noteworthy observation is that while the *CoT* prompting generally improves performance in larger LLMs, it is counterproductive in smaller models for some FoR classes. This suggests that the smaller models have difficulty inferring FoR from the longer context. This negative effect also appears in SG prompting, which uses longer explanations. Despite performance degradation in particular classes of small models, SG prompting performs exceptionally well across various models and achieves outstanding performance with Qwen2-72B. We further investigate the performance of CoT and SG prompting. As shown in Table 9, CoT exhibits a substantial difference in performance between contexts with inherently clear FoR and contexts requiring the template to clarify FoR ambiguity. This implies that CoT heavily relies on the specific template to identify FoR classes. In contrast, SG prompting demonstrates a smaller gap between these two scenarios and significantly enhances performance over CoT in inherently clear FoR contexts. Therefore, guiding the model to provide characteristics regarding topological, distance, and directional types of relations improves FoR comprehension.

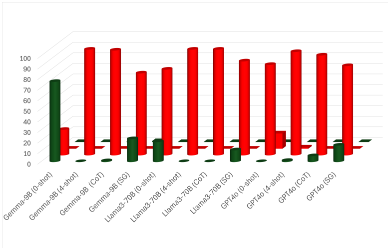
**A-spilt.** We use the same Figure 5 to observe the behavior when applying ICL. The A-split shows minimal improvement with ICL variations, though some notable changes are observed. With *few-shot*, all models show a strong bias toward external intrinsic FoR, even when the relatum lacks intrinsic directions, i.e., Box and Pen cases. This bias appears even in Gemma2-9B, which usually behaves



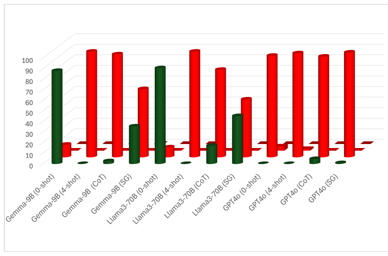
(a) Results of Cow Case in A-Split.



(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 5: 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**.

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 10. The Llama3-70B behaviors are also seen in LLama3-8B and GPT3.5. The plots for LLama3-8B and GPT3.5 are in Figure 6.

#### E.3.3 Experiment with different temperatures

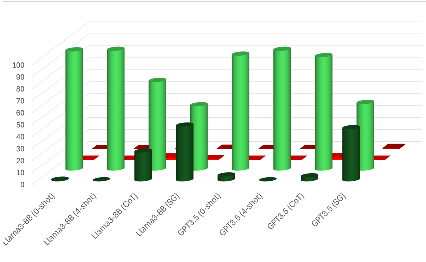
We conducted additional experiments to further investigate the impact of temperature on the biased interpretation of the model in the A-split of our dataset. As presented in Table 11, comparing distinct temperatures (0 and 1) revealed a shift in the distribution. The frequencies of the classes experienced a change of up to 10%. However, the magnitude of this change is relatively minor, and the relative preferences for most categories remained unchanged. Specifically, the model exhibited the highest frequency responses for the cow, car, and pen cases, even with higher frequencies in certain settings. Consequently, a high temperature does not substantially alter the diversity of LLMs' responses to this task, which is an intriguing finding.

## F In-context learning

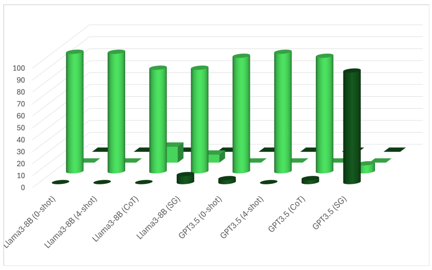
### F.1 FoR Identification

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.}"

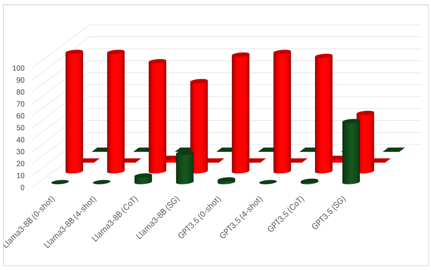
For the Chain of Thought (CoT), we only modified the instruction answer to "Answer only the category with an explanation. The answer should be in the form of {Explanation: Explanation Answer: Category.}" Similarly to CoT, we only modified the instruction answer to "Answer only the category with an explanation regarding topological, distance, and direction aspects. The answer should be in the form of {Explanation: Explanation Answer: Category.}", respectively. The example responses are



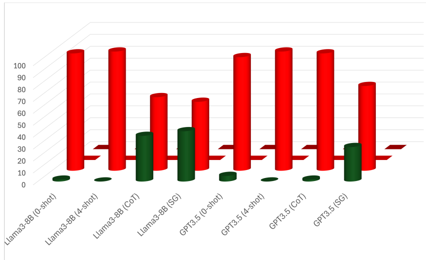

(a) Results of Cow Case in A-Split.




(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.

provided in Listing 4 for Spatial Guided prompting.

```
# Instruction to find frame of reference
     class of given context
"""
Instruction:
You specialize in language and spatial
    relations, specifically in the frame
     of context (multiple perspectives
    in the spatial relation). Identify
    the frame of reference category
    given the following context. There
    are four classes of the frame of
    reference (external intrinsic,
    internal intrinsic, external
    relative, internal relative). Note
    that the intrinsic direction refers
    to whether the model has the front/
    back by itself. (Example: a bird,
    human. Counter Example: a ball, a
    box). "

External intrinsic. The spatial
    description of an entity A relative
    to another entity B, where (1) A is
    not contained by B, (2) the spatial
    relation is based on B's facing
    orientation (intrinsic direction) if
     B has one.

Internal intrinsic. The spatial
    description of an entity A relative
    to another entity B, where (1) A is
    contained by B, (2) the spatial
    relation is based on B's facing
    orientation (intrinsic direction) if
     B has one.

External relative. The spatial
    description of an entity A relative
    to another entity B, where (1) A is
    not contained by B, (2) The spatial
    relation is relative to entity C,
    where C is not involved in the
    spatial relation and only provides
    the perspective. An example of C is
    the camera.

Internal relative. The spatial
    description of an entity A relative
    to another entity B, where (1) A is
    contained by B, (2) The spatial
    relation is relative to entity C,
    where C is not involved in the
    spatial relation and only provides
    the perspective. An example of C is
    the camera.

{Instruction answer}
# Normal Instruction answer: Find the
    frame of reference category without
    any explanations. The answer should
    be in the form of Answer: Category.
# COT Instruction answer: Find the frame
     of reference category with an
    explanation. The answer should be in
     the form of Explanation:
    Explanation. Answer: Category.
# SG Instruction answer: Find the frame
    of reference category with an
    explanation. Additionally, include
    three more explanations relating to
    topology, distance, and direction
    from the given context. The answer
    should be in the form of \"
    Explanation: Explanation. Topology:
```

| Model | ER | | EI | | II | | IR | |
|---|---|---|---|---|---|---|---|---|
| | temp-0 | temp-1 | temp-0 | temp-1 | temp-0 | temp-1 | temp-0 | temp-1 |
| Cow Case | | | | | | | | |
| 0-shot | 75.38 | 87.12 | 23.86 | 12.50 | 0.76 | 0.13 | 0.00 | 0.25 |
| 4-shot | 0.00 | 15.66 | 100.00 | 84.34 | 0.00 | 0.00 | 0.00 | 0.00 |
| CoT | 31.82 | 49.87 | 68.18 | 49.87 | 0.00 | 0.13 | 0.00 | 0.13 |
| SG | 51.39 | 70.45 | 48.61 | 29.42 | 0.00 | 0.00 | 0.00 | 0.13 |
| Box Case | | | | | | | | |
| 0-shot | 22.50 | 41.67 | 77.50 | 58.33 | 0.00 | 0.13 | 0.00 | 0.25 |
| 4-shot | 0.00 | 0.00 | 100.00 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CoT | 0.00 | 5.83 | 100.00 | 94.17 | 0.00 | 0.00 | 0.00 | 0.00 |
| SG | 11.67 | 33.33 | 88.33 | 66.67 | 0.00 | 0.00 | 0.00 | 0.00 |
| Car Case | | | | | | | | |
| 0-shot | 55.20 | 68.24 | 49.01 | 31.15 | 0.79 | 0.61 | 0.00 | 0.00 |
| 4-shot | 0.60 | 5.94 | 99.40 | 94.06 | 0.00 | 0.00 | 0.00 | 0.00 |
| CoT | 19.64 | 38.52 | 80.16 | 61.27 | 0.20 | 0.20 | 0.00 | 0.00 |
| SG | 44.25 | 56.97 | 55.75 | 43.03 | 0.00 | 0.00 | 0.00 | 0.00 |
| Pen Case | | | | | | | | |
| 0-shot | 90.62 | 96.88 | 9.38 | 3.12 | 0.00 | 0.61 | 0.00 | 0.00 |
| 4-shot | 0.00 | 7.03 | 100.00 | 92.97 | 0.00 | 0.00 | 0.00 | 0.00 |
| CoT | 17.19 | 28.91 | 82.81 | 71.09 | 0.20 | 0.20 | 0.00 | 0.00 |
| SG | 48.31 | 57.81 | 54.69 | 42.19 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 11: The results between two different temperatures of Llam3-70B on the A-spilt of FoREST. The number shows the percentage frequency of responses from the model.

```
    Explanation. Distance: Explanation.
    Direction: Explanation. Answer:
    Category.

Context: {spatial exprssion}
"""
```

Listing 1: Prompt for finding the frame of reference class of given context.

```
## These are four examples used in few-
    shot, CoT, and SG
# Example 1: The bird is outside the car
     and in front of the car relative to
     the car.
# Few-shot response: Answer: external
    intrinsic
# CoT response: Explanation: The car
    does not contain the bird, and the
    spatial relation is relative to the
    car's facing orientation. The
    intrinsic direction of the car is
    relevant in this context, as the
    description \"in front of\" implies
    a direction relative to the sheep's
    front or back. Answer: external
    intrinsic.
# SG response: Explanation: The car does
     not contain the bird, and the
    spatial relation is relative to the
    car's facing orientation. The
    intrinsic direction of the car is
    relevant in this context, as the
    description \"in front of\" implies
    a direction relative to the sheep's
    front or back. Topology: The bird is
     outside the car. Distance: There is
     little distance between the bird
    and the car. Direction: The bird is
    in the front direction of the car (
    entity B). Answer: external
    intrinsic.

# Example 2: The bird is inside of the
    car and left of the car from the car
    's perspective.
# Few-shot response: Answer: internal
    intrinsic
# CoT response: Explanation: In this
    scenario, the bird is contained
    within the car, meaning it is an
    internal relationship. Additionally,
     the description specifies the bird'
    s position relative to the car,
    indicating that the spatial relation
     is relative to the car itself,
    considering its intrinsic direction.
     Answer: internal intrinsic.
# SG response:  Explanation: In this
    scenario, the bird is contained
    within the car, meaning it is an
    internal relationship. Additionally,
     the description specifies the bird'
    s position relative to the car,
    indicating that the spatial relation
     is relative to the car itself,
    considering its intrinsic direction.
     Topological: The bird is inside the
     car. Distance: The bird is in the
    same location as the car. Direction:
     The bird is in the left direction
    of the car (entity B). Answer:
    internal intrinsic.

# Example 3: The box is inside and at
    the back of the room.
# Few-shot response: Answer: internal
    relative
# CoT response: Explanation: In this
    context, the box (entity A) is
    described as inside the room (entity
     B), fulfilling the first condition
    for an internal frame of reference.
    However, the spatial relation given
    (\" at the back\") cannot originate
    from the room since the room does
    not have intrinsic direction,
    implying it is relative to an
    external observer or reference point
     (C), such as a camera or a person
    viewing the scene. Answer: internal
    relative
# SG response:  Explanation: In this
    context, the box (entity A) is
    described as inside the room (entity
     B), fulfilling the first condition
    for an internal frame of reference.
```

```
    However, the spatial relation given
    (\" at the back\") cannot originate
    from the room since the room does
    not have intrinsic direction,
    implying it is relative to an
    external observer or reference point
     (C), such as a camera or a person
    viewing the scene. Topological: The
    box is inside the room. Distance:
    The box is in the same location as
    the room. Direction: From the
    observer's perspective, the box is
    located at the back of the room.
    Answer: internal relative

# Example 4: A phone is on the left of a
     tablet from my perspective.
# Few-shot response: Answer: external
     relative
# CoT response: Explanation: In this
    context, the spatial relation is
    described from the perspective of an
     observer (C) who is not involved in
     the spatial relation. The phone (A)
     is not contained by the tablet (B),
     and the spatial relation \"left of
    \" is relative to the observer's
    perspective, not the tablet's facing
     orientation. Answer: external
    relative.
# SG response:  Explanation: In this
    context, the spatial relation is
    described from the perspective of an
     observer (C) who is not involved in
     the spatial relation. The phone (A)
     is not contained by the tablet (B),
     and the spatial relation \"left of
    \" is relative to the observer's
    perspective, not the tablet's facing
     orientation. Topological: The phone
     is not contained by the tablet.
    Distance: There is some distance
    between the phone and the tablet.
    Direction: From the observer's
    perspective, the phone is located to
     the left of the tablet. Answer:
    external relative.
```

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

### F.2 Question Answering Example

```
'''
#Context: The bird is outside the car
    and in front of the car relative to
    the car. The car is facing toward
    the camera.
#Question: Based on the camera's
    perspective, where is the bird from
    the car position in the scene?
#SG information (used in SG + CoT):
    Frame of reference Explanation: The
    car does not contain the bird, and
    the spatial relation is relative to
    the car's facing orientation. The
    intrinsic direction of the car is
    relevant in this context, as the
    description \"in front of\" implies
    a direction relative to the sheep's
    front or back. Topology: The bird is
     outside the car. Distance: There is
     little distance between the bird
    and the car. Direction: The bird is
    in the front direction of the car (
    entity B). Frame of reference:
    external intrinsic.
#Normal response: Answer: front
#CoT / (SG + COT) response:Explanation:
    Based on the context, the bird's
    position is in the front direction
    of the car. The car is facing toward
     the camera. Then, the car's front
    direction is the camera's front
    direction. Therefore, the bird's
    position is in front of the car's
    position from the camera's
    perspective. Answer: front

#Context: The bird is inside the car and
     left of the car from the car's
    perspective. The car is facing to
    the right relative to the camera.  #
    Question: Based on the camera's
    perspective, where is the bird from
    the car's position?
#SG information (used in SG + CoT):
    Frame of reference Explanation: In
    this scenario, the bird is contained
     within the car, meaning it is an
    internal relationship. Additionally,
     the description specifies the bird'
    s position relative to the car,
    indicating that the spatial relation
     is relative to the car itself,
    considering its intrinsic direction.
     Topological: The bird is inside the
     car. Distance: The bird is in the
    same location as the car. Direction:
     The bird is in the left direction
    of the car (entity B). Frame of
    reference: internal intrinsic.
#Normal response: Answer: back
#CoT / (SG + COT) response: Explanation:
     Based on the context, the bird's
    position is in the left direction of
     the car. The car is facing to the
    right. Then, the car's left
    direction is the camera's back
    direction. Therefore, the bird's
    position is to the back of the car's
     position from the camera's
    perspective. Answer: back

Context: The box is inside and to the
    right of the room from the observer'
    s perspective.
Question: From the observer's
    perspective, what is the spatial
    relation of the box to the room?
#SG information (used in SG + CoT):
    Frame of reference Explanation: In
    this context, the box (entity A) is
    described as inside the room (entity
     B), fulfilling the first condition
    for an internal frame of reference.
    However, the spatial relation given
    (\" at the back\") cannot originate
```

```
    from the room since the room does
    not have intrinsic direction,
    implying it is relative to an
    external observer or reference point
     (C), such as a camera or a person
    viewing the scene. Topological: The
    box is inside the room. Distance:
    The box is in the same location as
    the room. Direction: From the
    observer's perspective, the box is
    located at the back of the room.
    Frame of reference: internal
    relative.
#Normal response: Answer: right
#CoT / (SG + COT) response: Explanation:
     Based on the context, the box is to
     the right of the room from the
    camera's direction. Therefore, the
    box's position is to the right of
    the room's position from the
    observer's perspective. Answer:
    right

Context: A phone is to the left of a
    tablet from my perspective. The
    tablet is facing to the right.
    Question: From my perspective, what
    is the spatial relation of the phone
     to the tablet?
#SG information (used in SG + CoT):
    Frame of Reference Explanation: In
    this context, the spatial relation
    is described from the perspective of
     an observer (C) who is not involved
     in the spatial relation. The phone
    (A) is not contained by the tablet (
    B), and the spatial relation \"left
    of\" is relative to the observer's
    perspective, not the tablet's facing
     orientation. Topological: The phone
     is not contained by the tablet.
    Distance: There is some distance
    between the phone and the tablet.
    Direction: From the observer's
    perspective, the phone is located to
     the left of the tablet. Frame of
    Reference: external relative.
#Normal response: Answer: left
#CoT / (SG + COT) response: Explanation:
     Based on the context, the phone is
    to the left of the tablet from my
    perspective. The direction of the
    tablet is not relevant in this
    context since the left relation is
    from my perspective. Therefore, from
     my perspective, the phone is to the
     left of the tablet. Answer: left
'''
```

Listing 3: Spatial expression examples using for few-shots, Chain-of-Thought (CoT), and Spatial Guide (SG) prompting for question-answering.

### F.3 Text to Layout

```
    # Instruction for generating
        bounding box
"""
Your task is to generate the bounding
    boxes of objects mentioned in the
    caption.
The image is size 512x512. The bounding
    box should be in the format of (x, y
    , width, height). Please considering
     the frame of reference of caption
    and direction of reference object if
     possible. If needed, you can make
    the reasonable guess.
"""
```

Listing 4: Prompt for generating bounding coordinates to use as the layout for layout-to-image models.