Localizing Events in Space: Comparing Humans and AI Models

Derrick Eui Gyu Kim, Kenneth Lai, James Pustejovsky
Department of Computer Science, Brandeis University
415 South Street, Waltham, MA 02453
egk265@brandeis.edu, klai12@brandeis.edu, jamesp@brandeis.edu

Abstract

Understanding how Large Language Models (LLMs) and Text-to-Image models (T2Is) acquire and apply implicit spatial knowledge remains an open challenge. In this paper, we present a novel dataset and evaluation framework designed to probe event localization capabilities in both humans, LLMs and T2Is. Our dataset includes 134 sentence pairs derived from Flickr30k captions, where explicit location information is systematically removed via Abstract Meaning Representation (AMR) parsing and manual refinement. Using this dataset, we analyze the effects of location ablation on spatial reasoning across human annotators, LLMs, and T2Is. Results show that while humans maintain robust location inferences after ablation, LLMs exhibit degraded performance, particularly for semantically polysemous verbs. T2Is demonstrate similar limitations, often generating visually inconsistent spatial contexts when locative cues are missing. Our findings highlight the gap between human and LLMs and T2Is in recovering implicit situational knowledge and suggest future directions for improving spatial reasoning in multimodal AI systems. This dataset contribution work serves as a proof-of-concept for systematic evaluation of implicit spatial reasoning and paves the way for larger-scale studies.

Keywords: event localization, spatial reasoning, multimodal evaluation, AMR, text-to-image generation, commonsense

1. Introduction and Motivation

This paper explores the abilities of humans and artificial intelligence (AI) models to identify the locations of events in text: this is the problem of *event localization* (Pustejovsky, 2013; Maienborn, 2001). Sometimes, the location of an event is explicitly mentioned in the text; for example, consider the following image caption, taken from the Flickr30k dataset (Young et al., 2014):

(1) Vendors sell products at a farmer's market.

For a human reading this caption and picturing the event in their head, it is clear that the location of the event, the selling of products, is in a market. Furthermore, assuming this human is familiar with the typical locations of farmer's markets, they are most likely imagining an outdoor scene. (The original image corresponding to this caption is shown in Figure 1 (a).)

Note that even when there is no specific location mentioned in a sentence, humans often still have some idea of where the event is taking place. For example, suppose we remove the phrase "at a farmer's market" from sentence (1):

(2) Vendors sell products.

Although the event is no longer explicitly mentioned as taking place in a market, there is still an association, derived from lived experience or world knowledge, between the act of selling and a market location. However, note that since the farmer's market was removed from the sentence, there is no longer a strong association with the outdoors.

(a) Flickr30k image

(b) DALL·E 3-generated image

Figure 1: Example of two images for the caption "Vendors sell products at a farmer's market".

Prior datasets addressing spatial reasoning—e.g., NLVR (Suhr et al., 2019), GQA (Hudson and Manning, 2019), and CLEVRER (Yi et al., 2020)—focus on structured or synthetic visual scenes with constrained linguistic diversity. However, there remains a lack of controlled benchmarks that specifically probe *implicit* spatial reasoning from text, where location cues are systematically manipulated. Our work introduces such a dataset and analysis pipeline, contrasting human, LLM, and text-to-image (T2I) model behavior under explicit versus ablated locative information.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11072–11084
11–16 May 2026.

We build on insights from linguistic theories of event localization (Pustejovsky, 2013) and situational presupposition (Stalnaker, 1974), linking them with recent embodied-AI research on affordance learning (Grauman et al., 2022). By combining AMR-based linguistic ablation with multimodal evaluation, we aim to expose where current AI systems diverge most sharply from human spatial reasoning.

Research Questions The dataset's precise control over location ablation enables experiments that isolate the effects of implicit spatial reasoning. This focused design facilitates clearer insights into model behavior under systematically controlled conditions, often obscured in larger and noisier datasets. We aim to address the following questions using our corpus:

- **LLM and Human Event Localization (Sec. 5.1):** Are LLMs capable of event localization in a manner similar to humans? In other words, given a sentence, do humans and LLMs assign the same location category to that sentence?
- **Text-to-Image Models' Localization Accuracy (Sec. 5.2):** Similarly, are T2Is capable of event localization? Given a sentence, a T2I generates an image; does the location category of the image match that of the sentence (as judged by humans), as shown in Figure 1(b)?
- **Effect of Removing Explicit Location Mentions (Sec. 5.3):** How do the location categories change, for both the sentences and the generated images, when the explicit location information is removed from the sentences?
- **Verb-Specific Resistance to Location Ablation (Sec. 5.4):** Are there certain events that resist the kind of location change mentioned above? Put another way, are certain verbs so strongly associated with particular locations, that those locations are recoverable even in the absence of a specific location mention?

Main Contributions The main contributions of this work are as follows:

- **A Novel Dataset for Event Localization:** We introduce a new benchmark dataset specifically designed to probe implicit spatial reasoning in both humans and AI systems. The dataset includes 134 sentence pairs with systematically controlled location ablations and over 800 corresponding images generated by state-of-the-art T2Is.
- **A Unified Evaluation Framework:** We propose a novel, and comprehensive framework for evaluating spatial reasoning across humans, LLMs, and T2Is, using hierarchical location annotations and Jensen-Shannon Divergence for distributional comparisons.
- **Empirical Insights into Model Limitations:** Through extensive experiments, we demonstrate that while humans maintain robust location inferences under explicit information ablation, LLMs and T2Is struggle, particularly with semantically polysemous verbs. Our findings reveal critical gaps in the implicit situational knowledge of modern LLMs and T2Is.
- **Recommendations for Future Research:** We identify key challenges in spatial reasoning for current AI systems and outline future research directions, including the need for richer multimodal grounding and improved integration of commonsense knowledge to bridge the gap between human and AI mental models.

2. Related Work

Commonsense and Spatial Reasoning Understanding implicit spatial information in language requires models to perform sophisticated commonsense reasoning. Commonsense reasoning frameworks such as ATOMIC (Sap et al., 2019), COMET (Bosselut et al., 2019), and ATOMIC-10x (West et al., 2022) provide large-scale causal and social knowledge bases, but offer limited coverage of spatial grounding. Recent efforts in multimodal commonsense reasoning, such as VCR (Zellers et al., 2019), attempt to bridge textual and visual domains, yet seldom target event-location inference. Our work extends this line of research by focusing on how humans, LLMs and T2Is infer the *where* of events rather than the *why* or *how*.

Notably, our dataset builds upon the Event Localization Corpus (ELC) created by Ward (2016). They annotate a subset of images from Flickr30k (Young et al., 2014), a widely used dataset that provides rich, human-written natural language descriptions of images, classifying them in a three-level hierarchy according to their locations. By matching the image annotations with their corresponding captions, Ward (2016) create a resource for studying the associations between events mentioned in text with their locations displayed in images. We extend this work by collecting caption annotations from humans, LLMs, and T2Is, showing the differences in their event localization abilities, and by systematically ablating the locations, showing the effects of explicit location information on event localization.

Visual Commonsense and Implicit Spatial Knowledge The extraction of implicit spatial and functional relationships between objects has been explored in visual-commonsense tasks (Yatskar et al., 2016; Collell et al., 2018; Li et al., 2023). Recent large multimodal models such as Kosmos-2 (Peng et al., 2024), LLaVA-NeXT (Liu et al., 2024), and GPT-4V (OpenAI, 2023b) demonstrate partial success at reasoning about relative spatial relations. Nevertheless, they remain brittle under underspecification or ablation. While their approaches focused on modeling relationships between objects, our work extends this line of inquiry to the domain of event localization and language models, analyzing whether modern AI systems can infer similar implicit spatial knowledge purely from textual descriptions. Our evaluation directly probes this brittleness through controlled textual removal of locatives, assessing whether models can recover the implicit scene configuration.

Multimodal Grounding and Embodied Affordances Recent advances in multimodal models have demonstrated impressive capabilities in visual understanding and language grounding. Models such as Flamingo (Alayrac et al., 2022), Flamingo-2, and GIT2 (Wang et al., 2022) have pushed the boundaries of vision-language pretraining, achieving strong performance on a range of benchmarks. Embodied and egocentric datasets such as Ego4D (Grauman et al., 2022) have also recently enabled models to couple language, action, and spatial perception. Despite these advances, it remains unclear whether these models can perform fine-grained spatial reasoning under conditions of incomplete or underspecified locative information. The recent emergence of powerful multimodal LLMs, such as GPT-4V (OpenAI, 2023b), has further raised questions about the extent to which these models possess grounded spatial understanding. Although GPT-4V and newer models like GPT-4o (OpenAI, 2024) exhibit strong multimodal reasoning, their implicit spatial reasoning capabilities remain underexplored. Our dataset provides a structured evaluation framework for assessing these models' abilities to handle implicit spatial reasoning and event localization.

This line of research also connects with Gibson's theory of affordances (Gibson, 1977), which posits that objects inherently suggest their possible uses and interactions based on their physical properties. These affordances shape human expectations about how objects are typically situated and used in physical environments. However, most benchmarks focus on active agents rather than static event localization. Our dataset provides a complementary testbed emphasizing linguistic and inferential components of spatial reasoning.

Lexical Semantics and Presupposition Our work is also grounded in research on lexical semantics and presupposition theory. Concepts such as *situational presupposition* and *common ground* (Stalnaker, 1974) describe how humans rely on shared background knowledge to interpret underspecified language. From a linguistic standpoint, spatial inference draws upon verbs' selectional preferences (Resnik, 1996), lexical distributional constraints (Erk, 2012), and situational presuppositions (Stalnaker, 1974). Recent neural work revisits these ideas in the context of pretrained embeddings (Lenci and Sahlgren, 2023), suggesting that LLMs encode shallow but detectable biases linking actions and typical environments. Our experiments explicitly quantify how robustly these implicit priors survive ablation, bridging lexical semantics and multimodal grounding.

3. Data

We used the Event Localization Corpus (ELC) (Ward, 2016) as our original data source. It is built on top of the Flickr30k corpus (Young et al., 2014), which consists of over 30,000 images, each annotated with five human-written descriptive captions, describing the objects, actions, and interactions in the scene. As previously mentioned, Ward (2016) collected annotations for a 5,068-image subset of Flickr30k, classifying the image locations according to a three-level hierarchy, shown in Figure 2. Each image's location label was then associated with its five corresponding captions.

We first identified a subset of ELC captions that contain explicit location information. To find these captions, we parsed them into Abstract Meaning Representations (AMRs) (Banarescu et al., 2013), using the SPRING parser (Bevilacqua et al., 2021). AMRs are graph-based representations of a sentence's predicate-argument structure. An example of an AMR for sentence (1) is given in Figure 3. We defined captions with explicit location information to be those whose AMR graphs contain an edge with label `:location`, or where the root node has a numbered role with type `LOC`, as defined in PropBank (Palmer et al., 2005). For example, the predicate `sit-01` has an argument `ARG2` with type `LOC`, so if an AMR has `sit-01` as its root, with the `ARG2` role filled, we consider the corresponding sentence to contain explicit location information. In Figure 3, the location information is highlighted in yellow.

After we obtained the subset of location-containing captions, we grouped them according to their root predicate (excluding those captions whose root nodes are not PropBank predicates). For each predicate, we computed the distribution of captions with that predicate as the root, across each of the 16 possible location categories. We then

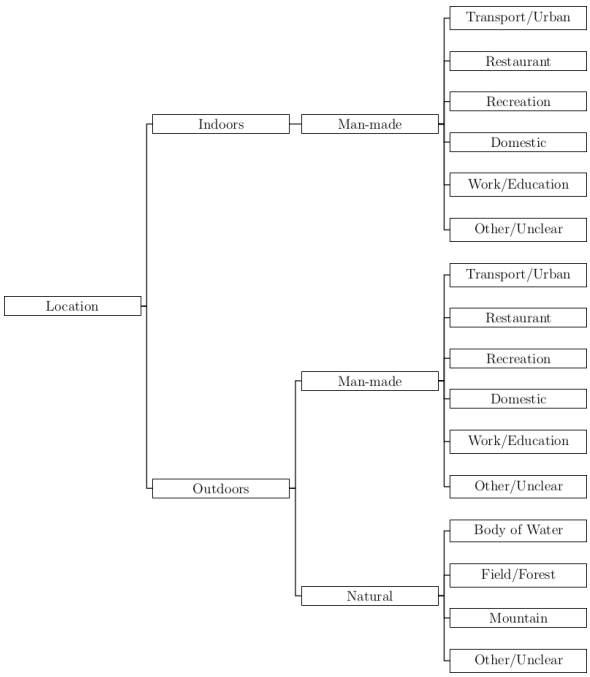

Figure 2: Location hierarchy. Locations are classified first as indoor or outdoor, then as man-made or natural, and finally in one of 16 specific categories.

```
(z1 / sell-01
  :ARG0 (z2 / person
    :ARG0-of (z3 / vend-01))
  :ARG1 (z4 / product)
  :location (z5 / market
    :mod (z6 / person
      :ARG0-of (z7 / farm-01)))
```

Figure 3: AMR for the caption "Vendors sell products at a farmer's market".

computed the Kullback-Leibler Divergence (KLD) of each of those distributions from the overall distribution of location categories in the entire ELC. In other words, for each predicate, if $P(x)$ is the predicate-specific distribution of locations $x \in \mathcal{X}$ and $Q(x)$ is the overall location distribution, we computed:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

We hypothesized that predicates with a higher KLD (i.e., those events that have highly specific location distributions) are more likely to be inherently associated with those locations in which they occur. To test this hypothesis, we ordered the predicates according to KLD, and split the list evenly into thirds. In order to get predicates associated with a wide variety of locations, within each third, we found the predicate with the highest value of $P(x)\log\left(\frac{P(x)}{Q(x)}\right)$ for each of the 16 location categories x.

For each such (predicate, location category) pair, we aimed to find 5 captions with that predicate as the root and labeled with that category. We found 27 such pairs,which will be shown in Table 4 in Appendix A. For each of the 134 captions[1], we manually removed the location information identified by the AMR parser, resulting in 134 original and 134 location-ablated captions in our dataset.

4. Annotation

4.1. Human Event Localization

To assess humans' event localization capabilities, we asked four human annotators to categorize each caption, both original and location-ablated, according to the location hierarchy. Annotators were graduate students at an American university and fluent speakers of English. Annotation was done through the Amazon Mechanical Turk Worker Sandbox, although annotators were recruited and paid outside of Mechanical Turk. Annotators were paid $15 per hour. An example of the caption annotation environment will be shown in Figure 5 in Appendix B.

At each level of the hierarchy, annotators were also asked to report their confidence in their selections on a 3-point Likert scale. Optional text boxes were provided for annotators to explain their reasoning, particularly for unclear or ambiguous annotations, i.e., if the annotator picked 1 (low confidence) on the Likert scale. This additional layer of qualitative data provides valuable insights into the difficulty and ambiguity of certain annotations, and serves as a rich source of information for future researchers and potential model training.

To mitigate potential order, memory, and condition biases, caption assignments were fully randomized, and annotators were blinded to whether each caption was original or location-ablated. Annotation of text and image data was conducted in separate phases, ensuring that annotators never saw both modalities for the same item. The guidelines explicitly instructed annotators to treat each caption as novel, without attempting to recall similar ones.

[1]Although there were originally $27 \times 5 = 135$ captions, one caption was inappropriate, and two of the T2Is refused to generate an associated image, so we excluded that caption.

4.2. LLM Event Localization

To probe the LLMs' event localization abilities, we sent the same 268 captions to ChatGPT-4o (OpenAI, 2024), Claude 3.7 Sonnet (Anthropic, 2025), DeepSeek-v3 (DeepSeek, 2024), and Llama 3.1 Sonar Large 128k Online (Perplexity, 2024). The LLMs were given similar instructions to the human annotators, except that they were not asked to give their confidence scores on the Likert scale. The prompt templates for each model will be given in Appendix B.

4.3. T2I Event Localization

Because T2Is do not return text output, we could not ask them to classify the captions directly. Instead, we generated images for each of the original and location-ablated captions using three T2Is, DALL·E 3 (OpenAI, 2023a), Flux.1 [dev] (Black Forest Labs, 2024), and Midjourney v6 (Midjourney, 2024). Each model was given only the caption as its prompt, with no other instructions. We generated 804 images in total, one image per model per caption.

Once these images were generated, in order to identify the locations represented in the images, we asked the human annotators (same group as above) to annotate the images according to the location hierarchy, also within the Amazon Mechanical Turk Worker Sandbox. An example of the image annotation interface is shown in Figure 6 in Appendix B. Again, annotators were asked to report their confidence in their annotations and given a text box for any comments.

To mitigate bias, Amazon Mechanical Turk gave the images to the annotators randomly, and the annotation guidelines asked them to treat the images as if they had not seen the sentences before so it would affect their annotations. All annotation interfaces were randomly ordered and condition-blinded to minimize presentation bias. Annotators were reminded to avoid guessing based on surface wording or visual stereotypes

After the images were annotated, we then adjudicated the results to create a gold standard label for each image. We took these labels to represent, for each image, the model's spatial localization of the caption used to generate it. Adjudication was done by majority vote of the annotators, with one of the authors serving as a tiebreaker when needed. To assess inter-annotator agreement, we computed Fleiss' kappa (Fleiss, 1971) at each level of the hierarchy. Agreement was very high, with kappa values of 0.900 at the indoor/outdoor level, 0.830 at the man-made/natural level, and 0.739 at the finest-grained category level.

5. Results

5.1. LLM and Human Event Localization

To understand how closely LLMs approximate human event localization, we compare the divergence in location category distributions between human annotations and LLM-assigned location categories. Jensen-Shannon Divergence (JSD) is used to quantify this difference due to its interpretability and symmetry. The goal is to assess the extent to which the probability distributions over location categories differ when generated by humans versus large language models. Specifically, the JSD between two probability distributions P and Q is defined as:

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}\left(P\|M\right) + \frac{1}{2}D_{\mathrm{KL}}\left(Q\|M\right) \quad (2)$$

where P is the distribution of location categories assigned by human annotators, Q is the distribution of location categories assigned by the LLMs, $M = \frac{1}{2}(P+Q)$ is the average distribution, and D_{KL} is the Kullback-Leibler divergence.

Higher JSD values indicate greater divergence between the human-annotated location distribution and the model's implicit assumptions. Conversely, lower JSD values reflect better alignment with the expected spatial contexts.

We set a significance threshold of $p < 0.05$ for all statistical tests.The mean JSD for original captions was 0.3336, while for location-ablated captions it was 0.4476. This difference was statistically significant according to Welch's t-test ($p = 4.07 \times 10^{-5}$), indicating a highly reliable difference between the two conditions.

The effect size, measured by Cohen's d, was -0.52, which represents a moderate negative effect. This suggests that location-ablated captions introduce a substantial increase in spatial uncertainty, reducing alignment with human-expected spatial contexts.

Together, these results indicate that LLM-generated captions, particularly after location ablation, diverge significantly from human mental models in event localization, both statistically and practically. Although LLMs can produce fluent and contextually appropriate sentences, their ability to capture implicit location cues remains limited compared to human annotators. This gap highlights the challenge for current models in acquiring and applying implicit spatial knowledge without explicit location mentions.

5.2. T2I Localization Accuracy

To assess the capability of T2Is in event localization, we analyzed whether the images generated from

captions correctly reflected the expected spatial settings. We used JSD again to quantify the alignment between the expected location distributions and those implied by the generated images.

We hypothesize that captions with inherently strong locational priors should yield lower JSD values, as the models more reliably generate appropriate spatial contexts. Captions with ambiguous or weak locational cues are expected to produce higher JSD values, reflecting uncertainty and inconsistency in generated scenes.

Captions with strong, stereotypical location expectations tended to yield more accurate spatial localization. For instance, the caption "A man is grilling meat on an outdoor grilling pit" (JSD $\approx 2.1 \times 10^{-6}$) consistently yielded backyard or patio scenes, suggesting strong location priors. In contrast, "A woman is drinking a beverage" (JSD $\approx$ 0.599) resulted in highly variable outputs, ranging from indoor bistros to open-air plazas, with several images lacking discernible cafe elements altogether. We include such examples of T2I generated images in Figure 4 .

We observed that captions with preserved location cues resulted in significantly better alignment between generated images and expected spatial contexts. Specifically, the average JSD for images from original captions was 0.2586, while the average JSD for images from location-ablated captions had a higher divergence of 0.3966. This increase in JSD after ablation indicates that event localization in T2I-generated images deteriorates when explicit locative cues are removed, reflecting a reliance on surface-level prompts and limited implicit spatial understanding.

These results highlight that T2Is perform well when captions have strong, prototypical location expectations but struggle under ambiguous conditions. Low JSD values correlate with accurate localization, indicating that models can reliably encode and reproduce well-established event-location priors. However, when faced with underspecified captions, higher JSD values reveal inconsistency and lack of spatial grounding.

While current T2Is demonstrate some capacity for event localization, their reliance on surface-level patterns and lack of deeper inferential spatial reasoning remains a key limitation. Improving spatial awareness in T2Is will likely require multimodal training incorporating richer environmental and situational knowledge.

(a) T2I(DALL·E3) generated image for caption "A man is grilling meat on an outdoor grilling pit"

(b) T2I(Flux) generated image for caption "A woman is drinking a beverage"

(c) T2I(Midjourney) generated image for caption "A woman is drinking a beverage"

Figure 4: Example of three images where one shows strong location priors while other shows high variability.

5.3. Effect of Removing Explicit Location Mentions

We analyzed how the annotation labels changed before and after explicit location information was ablated from captions. This allowed us to examine the underlying mental models—both human and AI—that are activated when interpreting language with or without situational context. Specifically, we compared the number of annotations that matched the location classification of the original (unmodified) caption versus the location-ablated version.

We observed a clear dichotomy. Some verbs (Table 1) demonstrated high resilience to location removal, including `ski`, `row`, `grill`, `drink`, `pass-by`, and `practice`. For these verbs, both human and AI mental models consistently inferred the same location information, indicating that these actions are tightly bound to prototypical environments (e.g., skiing $\rightarrow$ mountain, rowing $\rightarrow$ body of

Verb	Original %	Location-Ablated %
ski-01	1.00	1.00
row-01	1.00	1.00
grill-01	0.70	0.70
drink-01	0.95	0.80
pass-by-17	1.00	0.75
practice-01	0.85	0.75

Table 1: Verbs with High Location Annotation Consistency After Location Ablation

Verb	Original %	Location-Ablated %
sleep-01	0.75	0.00
sit-01	0.45	0.00
catch-01	0.35	0.10
contrast-01	0.10	0.00

Table 2: Verbs with Large Drops in Location Annotation Consistency After Location Ablation

water), and that both types of models encode these associations robustly.

In contrast, other verbs (Table 2) revealed major drops in location annotation consistency after location ablation. These included sleep, sit, catch, and contrast, which are more context-dependent and less tied to a canonical location. Here, both human and AI mental models showed degraded inference, although in some cases humans were still better able to use implicit context or background assumptions.

These findings reveal a fundamental difference between human and AI mental models. Verbs with high KLD (e.g., ski, row) maintain strong location priors, making them robust to the location ablation of explicit cues. Both humans and LLMs seem to have internalized these associations, though likely via different mechanisms—embodied experience for humans and data-driven statistical generalization for AIs. In contrast, verbs with low KLD (e.g., sleep, sit) expose the limits of inference, especially for AI systems, when surface-level cues are ablated. This points to a core difference in flexibility and context sensitivity between humans, LLMs and T2Is.

5.4. Verb-Specific Resistance to Location Ablation

We define resistance to ablation as the ability to retain correct location category classification after explicit locative cues are ablated.

In this section, we investigate whether certain verbs exhibit higher resistance to location removal, as measured by their impact on classification accuracy for humans, LLMs and T2Is.

As shown in Table 3, some verbs exhibit poor performance by LLMs due to their semantic flexibility and weak locative priors. Verbs such as drink, jump, and catch occur in a wide range of environments, making them highly context-sensitive.

Classification Level	Verb	Human Drop %	LLM Drop %	T2I Drop %
Fine-Grained	drink-01	0.15	0.60	0.40
	catch-01	0.25	0.40	0.33
	jump-03	0.00	0.3	0.33
Indoor/Outdoor	lie-07	0.10	0.35	0.06
	lean-01	0.05	0.15	0.27
	speak-01	0.00	0.10	0.14
Man-made/Natural	catch-01	0.10	0.30	0.33
	lie-07	0.10	0.30	0.07
	swing-01	0.00	0.05	0.13

Table 3: Post-ablation absolute accuracy drop (%) across classification levels and verbs. Higher drop indicates lower robustness to spatial ablation.

Conversely, verbs like lie, catch and speak exhibit large performance drops in LLMs and T2Is, particularly T2Is, indicating difficulty in inferring situated context from minimal cues. These verbs occur in diverse settings and lack strong locative priors, making spatial inference harder when explicit cues are removed.

We can distill the following patterns: 1) For high KLD verbs, they are more location bound and better preserved 2) For lower KLD verbs, they are more context sensitive, and harder to infer when stripped.

We can clearly see that humans are better at context recovery, and LLMs rely on statistical priors, while T2Is frequently produce visually and spatially incoherent outputs in the absence of explicit locative cues.

These findings suggest that while AI systems have internalized some event-location associations, they lack the flexible inferential mechanisms humans use to fill in missing contextual information.

6. Discussion and Conclusion

This work presents a systematic evaluation of implicit spatial reasoning in LLMs and T2Is. When explicit locative cues are removed, LLMs struggle to infer fine-grained location categories, relying heavily on surface-level statistical associations rather than situational reasoning. This limitation is particularly evident for semantically flexible verbs that do not strongly imply stereotypical locations.

T2Is face similar challenges, often generating visually incoherent scenes when deprived of explicit spatial cues. These models depend largely on textual prompts and learned co-occurrence patterns, lacking the embodied experience and pragmatic reasoning that enable humans to resolve spatial ambiguity. As a result, verbs like *lean*, *drink*, and *speak* frequently lead to ambiguous or inaccurate visual generations.

T2Is exhibit the largest performance drops when classifying scenes along the indoor-outdoor and manmade-natural dimensions, especially for context-dependent actions like *sit*, *jump*, and *catch*. Humans resolve such ambiguities by drawing on

rich, embodied knowledge and ecological affordances, while current models tend to default to prototypical or dataset-biased representations, often misaligning with real-world contexts.

6.1. Gibsonian Affordances and Model Limitations

Our findings echo Gibson's theory that affordances arise from agent-environment relationships, not object properties alone. Humans intuitively associate actions like *skiing* with mountains through direct perception of affordances, while LLMs and T2Is lack this embodied understanding. Current T2Is treat locations as mere co-occurrence patterns in text and images rather than dynamic affordance spaces. Incorporating embodied datasets such as Ego4D (Grauman et al., 2022) may help models develop more grounded spatial reasoning.

Our findings provide empirical support for Gibsonian affordances in the context of language grounding, reinforcing the conclusion that current AI systems lack equivalent embodied priors necessary for flexible spatial reasoning.

While LLMs and T2Is have acquired some event-location associations, they lack the flexible, context-sensitive reasoning required for robust spatial understanding. Enhancing spatial reasoning in AI systems will require richer multimodal training, integration of commonsense and world knowledge, and potentially, exposure to embodied interaction data. Our dataset provides a starting point for such research and highlights the need for models that reason beyond surface-level patterns.

6.2. Future Work

Future efforts should focus on expanding the dataset, diversifying annotator backgrounds, and developing automated, scalable methods for linguistic ablation. Furthermore, a promising direction for future work is to fine-tune language and vision-language models on datasets specifically curated for spatial reasoning. Even modest instruction tuning, using prompts that explicitly highlight spatial context (e.g., "Where is this event most likely happening?"), could help models better align with human expectations in ambiguous settings. Alternatively, targeted fine-tuning using contrastive examples—paired sentences with and without locative cues—may help reduce reliance on shallow co-occurrence patterns and improve robustness to location ablation. Given the structured nature of our dataset and annotations, this kind of intervention is not only feasible but also a natural next step toward building models with stronger situational grounding. Despite its limitations, this work establishes a structured framework for evaluating implicit spatial reasoning and provides a foundation for future large-scale studies. We will release the dataset and code under a permissive license upon publication to encourage further research.

Limitations

This study is a proof-of-concept exploration rather than a comprehensive benchmark. The dataset size (134 sentence pairs and 804 images) limits the generalizability of our findings and additionally, while AMR parsing enabled systematic location ablation, reliance on automatic parsing and manual refinement may have introduced inconsistencies.

Furthermore, in terms of the annotations, because all annotators were graduate students at a U.S. university and native English speakers, their spatial reasoning and world knowledge may reflect Western cultural and linguistic biases. Future work should include annotators from more diverse linguistic and cultural backgrounds to ensure broader generalizability.

Ethics Statement

This study involves human annotation and model analysis for spatial reasoning. All annotators were graduate students recruited directly and compensated at ($15/hour), consistent with academic wage standards. No personal identifying information was collected. Annotators were fully informed and gave written consent. Annotation was conducted via the Mechanical Turk sandbox, but compensation occurred outside the platform.

The study was determined to be exempt from IRB review, as it posed minimal risk and involved only de-identified data. All image captions were sourced from the publicly available Flickr30k and Event Localization Corpus datasets, and images generated by T2Is do not contain identifiable individuals.

We acknowledge that datasets involving spatial scenes could theoretically be misused for surveillance or simulation. To mitigate this, the dataset will be released under a permissive, research-only license, with documentation discouraging inappropriate use.

We used LLMs (ChatGPT-4o, Claude 3.7, DeepSeek-V3, and LLaMA 3.1 Sonar) for annotation and comparison. Their use is documented in Appendix B. We also used ChatGPT to polish the abstract and the introduction.

7. Bibliographical References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm

Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: A Visual Language Model for Few-Shot Learning. In *Proceedings of NeurIPS*.

Anthropic. 2025. Claude 3.7 Sonnet. URL https://claude.ai/.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 178–186.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One Spring to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without Pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12564–12573.

Black Forest Labs. 2024. Flux.1 [dev]. URL https://huggingface.co/lllyasviel/flux1-dev-bnb-nf4/blob/main/flux1-dev-bnb-nf4-v2.safetensors.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 6765–6772.

DeepSeek. 2024. DeepSeek-V3. URL https://chat.deepseek.com/.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

James J Gibson. 1977. *The Theory of Affordances*. Lawrence Erlbaum Associates.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of CVPR*, pages 18995–19012.

Drew A Hudson and Christopher D Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of CVPR*.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional Semantics*. Studies in Natural Language Processing. Cambridge University Press.

Menghao Li, Chunlei Wang, Wenquan Feng, Shuchang Lyu, Guangliang Cheng, Xiangtai Li, Binghao Liu, and Qi Zhao. 2023. Iterative Robust Visual Grounding with Masked Reference based Centerpoint Supervision . In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4653–4658, Los Alamitos, CA, USA. IEEE Computer Society.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Claudia Maienborn. 2001. On the position and interpretation of locative modifiers. *Natural language semantics*, 9(2):191–240.

Midjourney. 2024. Midjourney V6. URL https://www.midjourney.com/imagine.

OpenAI. 2023a. DALL·E-3. URL https://chatgpt.com/.

OpenAI. 2023b. GPT-4V(ision) System Card. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

OpenAI. 2024. ChatGPT-4o. URL https://chatgpt.com/.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding Multimodal Large Language Models to the World. In *Proceedings of ICLR*.

Perplexity. 2024. Llama 3.1 Sonar Large 128k Online. URL https://www.perplexity.ai/.

James Pustejovsky. 2013. Where Things Happen: On the Semantics of Event Localization. In *Proceedings of the IWCS 2013 Workshop on Computational Models of Spatial Language Interpretation and Generation (CoSLI-3)*, pages 29–39.

Philip Resnik. 1996. Selectional preferences and sense disambiguation. *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 52–57.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *Proceedings of AAAI*.

Robert C. Stalnaker. 1974. Pragmatic presuppositions. In Milton K. Munitz and Peter K. Unger, editors, *Semantics and Philosophy*, pages 197–213. New York University Press, New York.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*.

Chris Ward. 2016. A corpus for event localization. Master's thesis, Brandeis University, Waltham, Massachusetts. Available at https://scholarworks.brandeis.edu/view/pdfCoverPage?download=true&filePid=13419027310001921&instCode=01BRAND_INST.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the Obvious: Extracting Visual Common Sense Knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California. Association for Computational Linguistics.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2020. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 479–488.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A. List of Selected Predicates

Table 4 shows the PropBank predicates we used, grouped by location category and KL divergence level (high, medium, or low).

Location	High	Medium	Low
indoors/man-made/domestic	n/a	lie-07	n/a
indoors/man-made/other_unclear	n/a	speak-01	n/a
indoors/man-made/recreation	sing-01	practice-01	n/a
indoors/man-made/restaurant	drink-01	n/a	enjoy-01
indoors/man-made/transportation_urban	sleep-01	wait-01	sit-01
indoors/man-made/work_education	present-01	examine-01	use-01
outdoors/man-made/domestic	grill-01	n/a	contrast-01
outdoors/man-made/other_unclear	n/a	n/a	n/a
outdoors/man-made/recreation	swing-01	race-02	play-01
outdoors/man-made/restaurant	n/a	n/a	n/a
outdoors/man-made/transportation_urban	pass-by-17	drive-01	lean-01
outdoors/man-made/work_education	sell-01	n/a	n/a
outdoors/natural/body_of_water	row-01	n/a	kneel-01
outdoors/natural/field_forest	catch-01	run-02	jump-03
outdoors/natural/mountain	ski-01	n/a	n/a
outdoors/natural/other_unclear	n/a	n/a	n/a

Table 4: Selected PropBank predicates, grouped according to location category and KL divergence level.

B. Annotation Interfaces and LLM Prompts

Figure 5 shows the Amazon Mechanical Turk Worker Sandbox user interface for caption annotation, and Figure 6 shows that for image annotation. The prompt templates for each LLM's caption annotations are below.

ChatGPT-4o and Claude 3.7 Sonnet

Given the following caption, classify it according to the hierarchy below. Return only the matching tuple from the hierarchy, nothing else.
Caption: "{caption}"
Hierarchy: {hierarchy}
Classification:

DeepSeek-V3 and Llama 3.1 Sonar Large 128k Online

Classify the following caption according to the given hierarchy.
Caption: '{caption}'
Hierarchy: {hierarchy}
Instructions:
1. Choose exactly one classification from the hierarchy for each of the three levels:
- First, classify where the caption is happening: indoors or outdoors.
- Second, classify whether it involves man-made or natural elements.
- Third, select the most fitting category based on the activity or context in the caption (e.g., work, recreation, etc.)
2. Provide a full classification in the format: ('location', 'nature', 'activity')
3. Only include the classification tuple and no other text.
4. Do not omit any category, even if it is less obvious.
Classification:

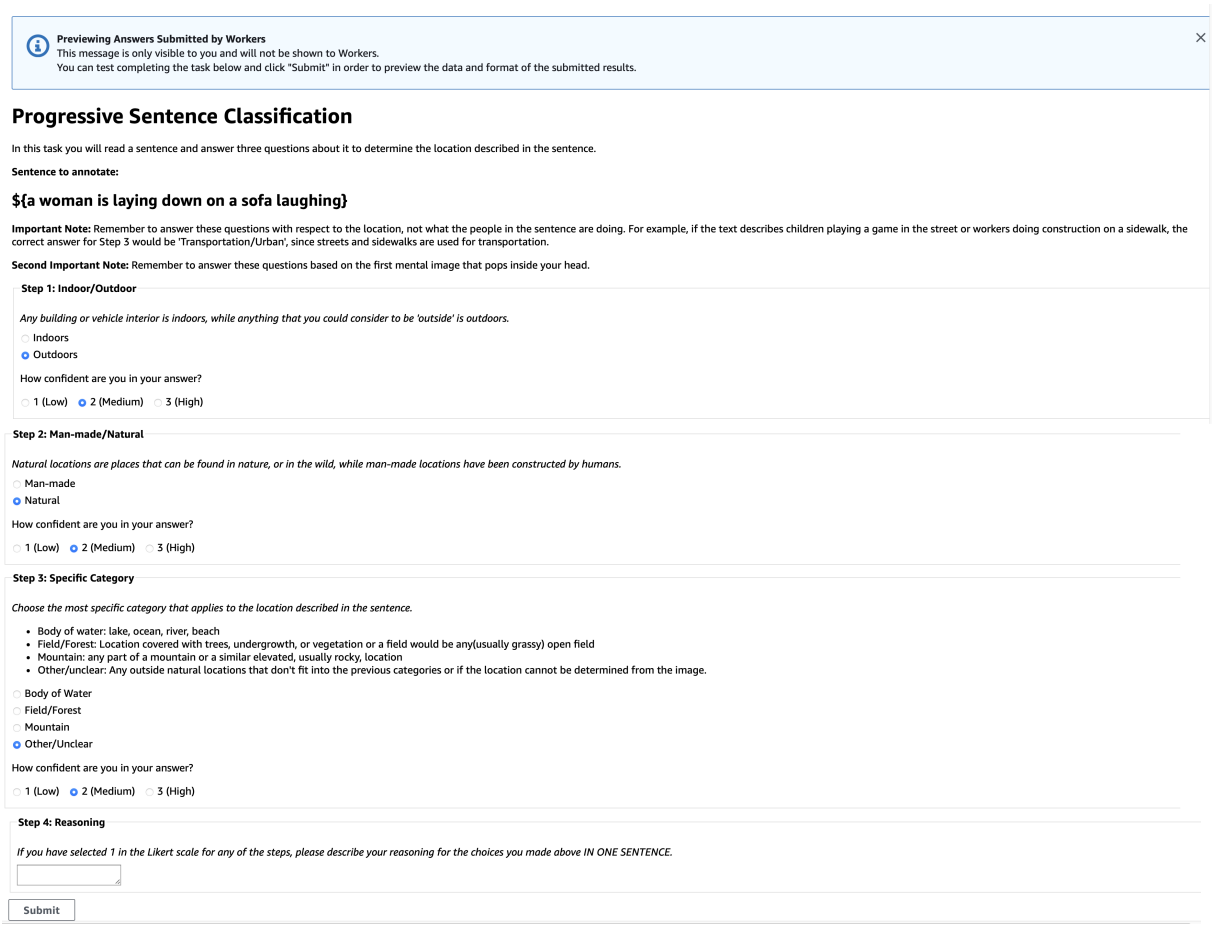

Figure 5: Sentence annotation user interface.

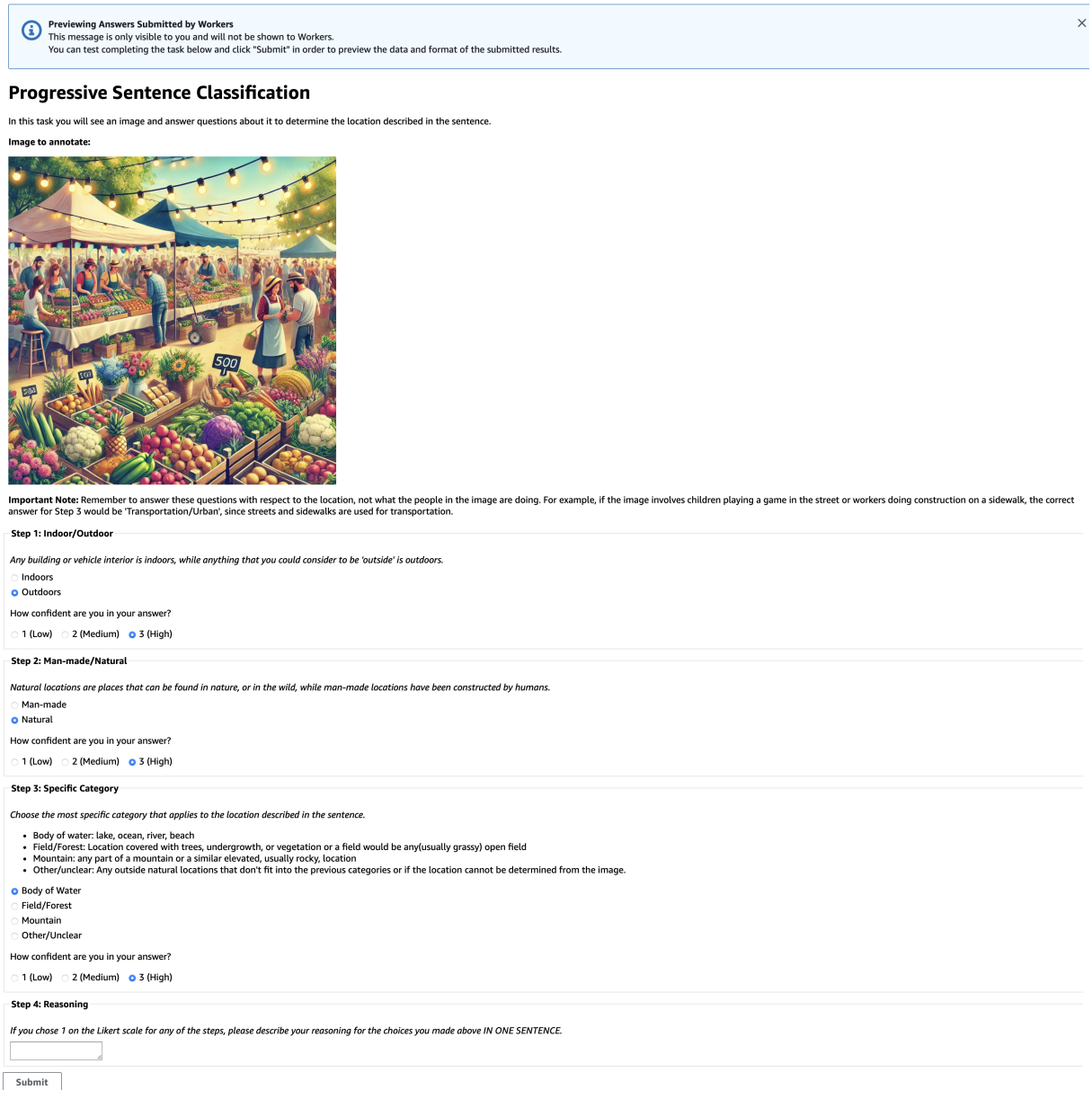

Figure 6: Image annotation user interface.