Where the Cat Sat: A Multilingual Framework for Spatial Language Understanding

Demian Inostroza[1], Ekaterina Vylomova[1], Charles Kemp[1], Mae Carroll[1], Wanchun Li, Meladel Mistica[1]
[1]University of Melbourne
{inostrozaad, vylomovae, c.kemp, mae.carroll, misticam}@unimelb.edu.au
wanclilee@gmail.com

Abstract

Spatial language understanding is fundamental to tasks from navigation and direction, robot control, to document understanding for a multitude of languages and tasks, yet current work exhibits biases toward English and prepositional marking. We present a multilingual framework and benchmark decomposing spatial relations into surface elements (figure, ground, predicate, markers) and semantic components (dynamicity, stasis). Evaluating frontier large language models (LLMs) on Spanish, Basque, and Chinese with text-only input, we find high performance on figure and ground identification but persistent gaps in two areas: semantic classification of topological and projective relations, and surface identification of morphological spatial markers—Basque case affixes proving most challenging with recognition of spatial elements as low as 15.3%. These results suggest that surface parsing does not entail spatial understanding, and that evaluation must include spatial marking strategies used across typologically diverse languages.

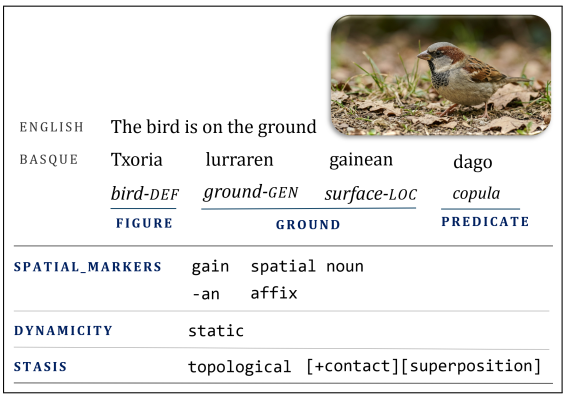

Figure 1: Example annotation for a Basque sentence describing a bird on the ground. Surface elements: *Txoria* ('the bird') is the `figure`; *lurraren gainean* ('on the ground's surface') is the `ground`; *dago* ('is') is the `spatial_predicate`, classified as copula; *gain* ('surface') and *-an* are `spatial_markers`, a spatial noun and affix respectively. Semantic components: `dynamicity` is static; `topological` is [+contact][superposition].

1 Introduction

Spatial language understanding remains a fundamental challenge in Natural Language Processing (NLP), as evidenced by ongoing workshop series dedicated to the topic (Kordjamshidi et al., 2018; Bhatia et al., 2019; Kordjamshidi et al., 2020; Alikhani et al., 2021; Kordjamshidi et al., 2024). Current approaches exhibit systematic biases toward English and well-resourced Indo-European languages (Ulinski et al., 2019; Olek and Piasecki, 2024), limiting our understanding of how well computational models genuinely comprehend spatial relations across typologically diverse languages.

Existing work exhibits four limitations: spatial markers are understood predominantly as prepositions, overlooking case marking and spatial nouns; spatial relations are labeled using English prepositions as categories rather than decomposed into semantic primitives (Liu et al., 2025; Beekhuizen, 2025), limiting analysis and comparison across languages; relations requiring frames of reference (Levinson and Wilkins, 2006) are ignored or conflated with topological ones; and finally spatial relations that involve motion have been rarely studied systematically.

This paper presents a multilingual framework and benchmark for spatial language understanding that addresses these limitations, covering three typologically diverse languages: Spanish (Indo-European), Basque (Isolated), and Chinese (Sino-Tibetan). Our evaluation framework identifies six components: four surface elements (`figure`, `ground`, `spatial_predicate`, `spatial_markers`) and two semantic components (`dynamicity`, `stasis`), with `stasis` divided into `topological` (topological primitives) and

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35344–35362

projective (frame-of-reference distinctions), the definitions of which are described in Section 2.2.

Our contributions include: a novel framework for spatial language understanding grounded in cross-linguistic typology; a multilingual benchmark covering Spanish, Basque, and Chinese with gold-standard annotations; recognition of spatial_markers beyond prepositions; primitive decomposition for topological relations and explicit frame-of-reference annotation; and systematic coverage of dynamicity × stasis combinations.

Our evaluation reveals systematic limitations in current LLMs. Basque spatial_markers prove particularly challenging across model scales, with small models showing severe difficulties and frontier models demonstrating variable performance—often the weakest compared to Spanish and Chinese markers. This systematic challenge appears to stem from these models' inability to reliably segment and classify case affixes as spatial markers. Across languages, models perform better on surface elements than semantic components.

2 A Multilingual Framework for Spatial Language Understanding

2.1 Framework Overview

Our framework decomposes spatial relations into six components organized along two dimensions. Figure 1 shows an example annotation for a Basque sentence describing a spatial scene. We first define the surface elements that must be extracted from text, then specify the semantic components that must be inferred.

Surface Elements. Linguistic forms present in the sentence that models must identify by parsing: figure, ground, spatial_predicate, and spatial_markers are detailed in Section 2.2.

Semantic Components. Conceptual categories not explicitly marked that models must infer: dynamicity and stasis (type of spatial relationship: topological or projective). These are detailed in Section 2.3.

2.2 Surface Elements

figure. The noun phrase denoting the entity whose location is being described (e.g., *Txoria* 'the bird'). We extract the complete noun phrase including all determiners, modifiers, and complements.

ground. The reference object relative to which the figure's location is specified (e.g., *lurraren gainean* 'the ground's surface'). We include spatial nouns like *gain* when they function as part of the ground phrase, essential for languages where spatial nouns are integral to spatial description (Levinson and Meira, 2003).

spatial_predicate. The predicate expressing the spatial relation linking figure and ground (e.g., *dago* 'exists/is', type: copula). We distinguish *verb* (i.e. content/lexical verbs) from *copula* (i.e. grammatical elements that link figure to ground).

spatial_markers. Elements that encode the spatial relationship between figure and ground. In the example (Figure 1), we identify two markers: *gain* (type: spatial noun) and *-an* (type: affix, the locative case marker). We recognize three types: adposition (prepositions, postpositions, or grammaticalized multi-word constructions), affix (case markers or morphological affixes encoding spatial meaning (Creissels, 2008)), and spatial noun (lexical nouns with spatial meaning functioning within the ground phrase). Only elements grammatically linked to the ground noun phrase are considered part of this component.

2.3 Semantic Components

dynamicity. The temporal aspect of the spatial relation. We distinguish three values: static (the figure's spatial relation to the ground remains constant), source (the figure moves away from the ground), and goal (the figure moves toward the ground). In our example, the value is static: the bird is not in motion. However, different visual stimuli can elicit sentences with different dynamicity values—"The bird lands on the branch" would be goal, while "The bird flies from the branch" would be source.

stasis. The overall type of spatial relationship between figure and ground. Following Levinson and Wilkins (2006), we distinguish two spatial domains: topological and projective relations. Table 4 (Appendix B) provides a complete mapping of all primitive values to representative English sentences for reference.

topological. These relations involve contiguity or close proximity between figure and ground, without requiring directional specification. Rather than labeling these with atomic cat-

egories, we characterize them through semantic primitives. In our example, the value is [+contact][superposition], indicating "the bird" is in physical contact with the `ground` and positioned above it. All `topological` values specify contact status ([+contact] or [-contact]). Beyond contact, we identify four topological types: containment (`figure` within boundaries of `ground`, specified as [open] or [closed]), attachment (`figure` mechanically fastened to `ground`), superposition (`figure` superior to `ground`, optionally [top]), and subposition (`figure` inferior to `ground`). This primitive-based approach aims to address the challenge of describing topological relations in a language-neutral way, as existing work relies on English prepositions as categorical labels, which may obscure distinctions that are expressed differently across languages. (Liu et al., 2025; Beekhuizen, 2025).

`projective`. These relations apply when `figure` and `ground` are separated in space and directional specification becomes necessary. Unlike `topological` relations, `projective` relations require external coordinate systems—frames of reference. Following Levinson and Wilkins (2006), we distinguish three frame types: relative (coordinates from the observer's bodily axes; values: left, right, front, back); intrinsic (coordinates from inherent facets of the `ground` object; values: front, back); and absolute (fixed environmental bearings; values: north, south, east, west). Appendix A.2 provides annotated examples of `projective` relations.

2.4 Limitations of Existing Approaches

Current spatial language evaluation frameworks were designed primarily for English and exhibit systematic biases that limit their applicability to typologically diverse languages. In terms of surface form, existing approaches extract only `figure` + preposition + `ground`, overlooking the broader range of adpositions and other encoding strategies such as case marking (Creissels, 2008), spatial nouns, or zero marking (Haspelmath, 2019; Stolz et al., 2014). Few authors have explicitly addressed the complexity of spatial case markers in cross-linguistic frameworks; Beekhuizen (2025) represents a recent exception, developing an extraction pipeline that handles affixal markers across languages, though not in the context of LLM evaluation. Spatial nouns tend to be ignored as contributing elements to spatial relations, despite some languages relying heavily or entirely on such elements (Levinson and Meira, 2003; Lakoff, 1987). Spatial predicates, which contribute to the overall spatial meaning, are also often excluded from evaluation.

Beyond surface elements, semantic evaluation presents additional challenges. `Topological` and `projective` relations are rarely distinguished—they are either treated as a single undifferentiated spatial category, or `projective` relations are ignored entirely. The semantic categories themselves are problematic: topological relations are labeled using English prepositions as theoretical categories (e.g., "on", "in", "above") rather than being decomposed into semantic primitives that could apply across languages. Finally, `dynamicity`—whether a spatial relation is static or involves motion—is rarely studied systematically, and the interaction between `dynamicity` and spatial `topological` is almost never evaluated jointly.

Our framework directly addresses each of these gaps. On the surface side, we extend extraction beyond the `figure` + preposition + `ground` triple to include `spatial_predicate` and `spatial_markers` of three types (adpositions, affixes, and spatial nouns), capturing a broader range of encoding strategies across typologically diverse languages. On the semantic side, we explicitly separate `topological` and `projective` relations, decompose topological relations into language-neutral primitives rather than English preposition labels, and systematically annotate `dynamicity` in joint interaction with `stasis`.

3 Multilingual Data Collection

We now demonstrate the applicability of this framework through data collection across three typologically diverse languages.

3.1 Language Selection

Spanish (Indo-European) encodes spatial relations primarily through prepositions, representing the marking pattern typical of major European languages. Basque (isolate) employs agglutinative morphology with spatial cases combined with spatial nouns, thus allowing us to evaluate LLMs' ability to recognize affixes as spatial markers. Chinese (Sino-Tibetan) distributes spatial meaning across prepositions and localizers.[1]

[1]Language-specific analytical decisions—regarding marker boundaries, contracted forms, and the grammatical status of elements such as Chinese localizers—are detailed in Appendix D.

3.2 Visual Stimuli Design

Following established methods for cross-linguistic spatial elicitation (Bowerman and Pederson, 1992), we created canonical images depicting spatial relations to elicit natural descriptions from native speakers.[2]

Our stimuli target specific primitive values for topological relations (e.g., contact vs. non-contact, open vs. closed containment), include projective relations covering all three frames of reference (relative, intrinsic, absolute), and systematically vary `dynamicity` (static, source, goal) for each spatial relation. This design yields systematic coverage of the `stasis` × `dynamicity` space: 63 unique combinations per language.

3.3 Elicitation Procedure

Data collection proceeded in two phases.

Pilot phase. For each language, we designed a pilot questionnaire with 12 images covering representative spatial relations: 1 `topological` type × 3 `dynamicity` values, plus 9 `projective` relations (3 frames × 3 `dynamicity` values). Native speaker experts produced sentences describing each image and segmented the grammatical components (`figure`, `ground`, `spatial_predicate`, `spatial_markers`). Informed by our investigation on spatial literature, these initial parses informed the development of language-specific parsing criteria.

Full phase. The procedure was expanded to cover all 63 `stasis` × `dynamicity` combinations. To reduce annotator burden, we simplified the task: native speakers produced only sentences, without grammatical analysis. Segmentation was then performed applying the criteria established in the pilot phase.

Verification. A second native speaker expert reviewed all analyses to ensure consistency with the operational definitions.

3.4 Resulting Gold Standard

The final parallel dataset comprises 189 sentences across the three languages, with gold-standard annotations for all six spatial relation components. This size reflects the diagnostic nature of the benchmark: each sentence targets a specific `stasis` × `dynamicity` combination, ensuring systematic coverage of the primitive space rather than broad naturalistic sampling.

[2]The dataset and evaluation code are publicly available in our GitHub repository.

4 Experimental Setup

4.1 Model Selection

We evaluate two groups of models to assess the effect of scale on spatial language understanding. All models are evaluated with text-only input (sentence alone); a supplementary multimodal condition is reported from Table 13 onwards.

Frontier models. We evaluate three large-scale multimodal models: Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o (OpenAI, 2024), and Qwen2.5-VL-72B-Instruct (Bai et al., 2025). These models represent current frontier capabilities and serve as the primary benchmark for our evaluation.

Small models. We additionally evaluate three smaller multimodal models: Claude-3.5-Haiku (Anthropic, 2024), GPT-4o-mini (OpenAI, 2024), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). All models are presented as multilingual, though language support details vary by provider.

Text-only models. To verify that observed patterns are not limited to using multimodal architectures on text-only input, we additionally evaluate three text-only models of comparable scale: Gemma-3-12B-IT (Gemma Team, 2025), Qwen2.5-14B-Instruct (Qwen Team, 2024), and DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025). Results (Table 12) confirm that the same performance patterns hold regardless of model architecture.

This dual-scale design allows us to distinguish limitations that reflect genuine difficulty in spatial understanding from limitations that reflect model capacity. Patterns of failure that appear at both scales suggest fundamental challenges; patterns that appear only at smaller scales suggest capacity-dependent limitations.

4.2 Prompting Strategy

To isolate the understanding of specific spatial components and mitigate error propagation, we employ a component-wise prompting strategy. Rather than requesting a single complex JSON object, the model was queried sequentially for each of the six components.

For each inference call, the prompt consisted of: (i) the operational definition for that com-

ponent (identical to the annotation guidelines in Appendix C), and (ii) the target sentence. This approach ensured that evaluation measured the model's ability to apply schema definitions rather than its ability to maintain long-context coherence.

4.3 Inference Configuration

Inference was performed using the OpenRouter API[3] for all models, with temperature 0.2 and maximum token limit of 8,192. To account for stochastic variance, we conducted three independent runs per model-language pair.

4.4 Evaluation Metrics

Our evaluation employs granular scoring that rewards partial understanding while penalizing hallucinations. All scores range from 0.0 to 1.0 and are reported and discussed as percentages throughout.

Figure and Ground. Evaluated using normalized Levenshtein similarity between gold and predicted text spans, yielding scores from 0.0 (no match) to 1.0 (exact match).

Spatial Predicate. Evaluated as the average of two components: (1) normalized Levenshtein similarity between gold and predicted text, and (2) binary match for predicate type (verb vs. copula).

Spatial Markers. A position-based metric that evaluates each marker independently, then averages across all positions (using the maximum of gold/predicted marker counts). For each marker, we compute the average of: (1) normalized Levenshtein similarity between marker texts (with length penalty for overprediction), and (2) binary match for marker type (adposition, affix, or spatial noun). Missing markers score 0.0.

Topological. Features extracted from bracket notation (e.g., [+contact], [superposition]) are evaluated using F1-score, treating each primitive as an independent feature.

Projective. Frame type and directional value are treated as independent features and evaluated using F1-score.

Dynamicity. Exact match accuracy over three categories: static, source, and goal.

[3] https://openrouter.ai

5 Results

5.1 Overall Performance

Table 1 summarizes performance across all components and languages for frontier models. These models achieve strong performance on surface elements across languages, with `figure` and `ground` identification generally high (83.0%–100%). `Spatial_predicate` identification remains strong for Spanish and Basque but lower for Chinese, primarily because models consistently predict *zài* as a copula predicate in constructions where no explicit predicate exists, rather than leaving the field empty; secondary errors include overpredicting by incorporating spatial markers into the predicate span (e.g., predicting *cóng* as part of the predicate) and truncating verb compounds (e.g., *shǐ lí* → *lí*). Semantic components show more variation: `topological` and `projective` scores are lower than surface element extraction.

Small models show lower and more variable performance compared to frontier models (see Table 6 in Appendix E for small model comparisons). `Figure` and `ground` identification remain relatively stable for Claude and Qwen-7B, though GPT drops to 40.8% on Basque `figure` extraction. The most pronounced difference between scales appears in semantic components and Basque `spatial_markers`.

Across both model scales, surface elements outperform semantic components, with this gap particularly pronounced for Chinese in frontier models.

5.2 Spatial Markers

Table 1 shows `spatial_markers` scores for frontier models. `Spatial_markers` extraction reveals cross-linguistic asymmetry in our evaluation. For frontier models, marker scores vary substantially: Claude maintains comparable performance across languages, while GPT and Qwen show notably lower scores for Basque markers despite strong performance on other surface elements.

Small models demonstrate larger gaps in performance (see Table 6 in Appendix E for comparison). Marker scores decrease across all languages, with Basque markers showing the largest decline. Basque marker extraction achieves the lowest scores among marker types across all small models (15.3%–29.6% vs. 49.1%–83.5% for Spanish), suggesting that segmenting and classifying case affixes poses particular difficulty at smaller scales.

Table 1: Overall frontier (%). Figure/Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score.

	Spanish			Basque			Chinese		
	Claude	GPT-4o	Qwen	Claude	GPT-4o	Qwen	Claude	GPT-4o	Qwen
Figure	100.0	100.0	100.0	99.8	96.5	83.0	99.0	96.3	97.8
Ground	85.0	83.5	86.3	99.3	97.3	83.7	99.3	100.0	99.1
Predicate	99.2	99.0	94.3	87.1	90.3	82.0	83.4	61.9	68.2
Markers	74.2	76.4	70.5	72.1	49.4	17.5	76.4	70.8	61.3
Dynamicity	98.9	97.4	92.1	97.9	96.3	89.4	95.8	97.9	88.9
Topological	73.7	66.9	49.4	64.6	68.9	42.8	57.0	50.1	50.2
Projective	87.2	92.2	60.6	79.7	82.8	40.6	74.0	54.6	65.0

Model	Spanish		Basque		Chinese	
	Surface	*Sem*	*Surface*	*Sem*	*Surface*	*Sem*
Claude-3.5-Sonnet	89.6	86.6	89.6	80.7	89.5	75.6
GPT-4o	89.7	85.5	83.4	82.7	82.3	67.5
Qwen2.5-VL-72B-Instruct	87.8	67.3	66.5	57.6	81.6	68.0

Table 2: Surface vs Semantic scores (%) for frontier models. Surface: avg. of Levenshtein-based metrics; Semantic: avg. of Dynamicity accuracy and Topological/Projective F1.

Model	Spanish		Basque		Chinese	
	Surface	*Sem*	*Surface*	*Sem*	*Surface*	*Sem*
Claude-3.5-Haiku	90.0	74.7	73.2	73.4	82.7	65.9
GPT-4o-Mini	77.4	53.2	56.0	45.4	74.1	47.6
Qwen-2.5-VL-7B-Instruct	80.6	46.4	64.8	35.4	55.8	31.8

Table 3: Surface vs Semantic scores (%) for small models. Surface: avg. of Levenshtein-based metrics; Semantic: avg. of Dynamicity accuracy and Topological/Projective F1.

5.3 Surface Elements vs Semantic Components

Table 2 and Table 3 summarize average scores for surface elements and semantic components separately, computed as the mean of the respective component-wise metrics for frontier and small models respectively. Given the heterogeneous nature of these metrics, comparisons across categories should be interpreted as qualitative tendencies rather than precise quantitative differences.

Surface element scores exceed semantic scores across nearly all conditions. This gap is partly expected by design, but notably `dynamicity`—also requiring inference—achieves 88.9%–98.9% accuracy, suggesting that the difficulty extends beyond the inherent challenge of semantic classification. The magnitude of this gap varies by language and model scale.

For frontier models, Spanish exhibits a narrow gap (89.6% vs. 86.6% for Claude), with surface and semantic scores nearly equivalent. Chinese shows substantial gaps: GPT-4o drops from 82.3% surface score to 67.5% semantic score, highlighting that correct parsing does not guarantee semantic comprehension. Basque demonstrates an intermediate pattern—surface element scores remain high except for `spatial_markers`, where models struggle to segment case suffixes from their stems (e.g., predicting *barruan* as a single token instead of separating *barru* + *-n*), and semantic scores fall below Spanish despite comparable parsing success on other surface elements, suggesting that morphological segmentation failures could propagate into semantic classification.

Small models show the same directional pattern but with lower overall performance and higher variance. The qualitative character of errors also differs between scales: frontier model errors cluster around specific components and languages, while small model errors are more uniformly distributed.

5.4 Stasis Breakdown

5.4.1 Topological

Frontier models reveal systematic patterns across semantic primitive combinations. Closed containment, superposition, and subposition achieve moderate-to-high scores, while open containment proves more challenging. The contact distinction and open versus closed geometry remain difficult across models. The [top] specification shows variable performance. (See Tables 7 and 8 in Appendix E for detailed performance on topological primitive combinations for frontier and small models, respectively).

Small models show moderate performance with systematic patterns emerging: [+contact] primitives consistently outperform [-contact] primitives, particularly for containment and superposition relations. However, overall scores remain lower than frontier models, with greater variance across primitive combinations and languages.

5.4.2 Projective

Frontier models achieve strong performance on absolute frame relations across Spanish and Basque, with Claude and GPT approaching perfect scores. Intrinsic frame relations also perform well for Claude and GPT, though Qwen shows notable difficulty. Relative frame relations, however, show mixed results. The [relative][left] and [relative][right] distinctions achieve high scores across models, while [relative][back] and [relative][front] prove more challenging. (See Tables 9 and 10 in Appendix E, which show performance on `projective` relations by frame of reference for frontier and small models, respectively).

Small models show strong differentiation by frame type: relative frame relations achieve high scores for Claude and GPT, while intrinsic frame relations prove most challenging across all models. Absolute frame performance varies by model, with Claude and GPT maintaining moderate-to-high scores while Qwen shows weakness.

6 Discussion

Parsing Does Not Entail Understanding. The most consistent finding across our evaluation is the dissociation between surface element identification and semantic classification. Across both frontier and small models, scores on surface elements—`figure`, `ground`, `spatial_predicate`, `spatial_markers`—exceed scores on semantic components—`topological` and `projective` relations. This gap persists regardless of model scale, though its magnitude and character differ.

Chinese provides a clear demonstration of this pattern for frontier models. Claude and GPT achieve strong `ground` extraction (99.3%–100%), correctly identifying spatial nouns as part of the `ground` phrase. Yet `topological` scores remain moderate (50.1%–57.0%), and `projective` scores vary substantially. Models identify *what* encodes spatial meaning without necessarily understanding *what spatial relation* is encoded.

Chinese `spatial_markers` illustrate a specific incoherence in frontier model behavior. Models include localizers as part of the `ground` phrase when prompted for `ground` extraction, yet classify these same elements as adposition rather than spatial noun when prompted for marker identification. This reveals that models apply task-specific heuristics rather than maintaining stable representations of grammatical status. An experiment with joint extraction of all components (Table 11, Appendix E) confirms this instability: localizers still appear in `ground` phrases but are not extracted as spatial nouns in markers, demonstrating the same contradiction. Overall performance degrades substantially—marker scores drop by up to 30 percentage points depending on the model.

The qualitative character of errors also differs between scales. Frontier model errors cluster around specific components and languages, which points to systematic challenges in particular aspects of spatial understanding. In contrast, small model errors spread more uniformly across different areas—a pattern consistent with general capacity constraints. This distinction is informative: where frontier models succeed but small models fail may reflect capacity-dependent learning; where both struggle may indicate more fundamental challenges in spatial reasoning that persist despite increased scale.

Dynamicity as Exception. Not all semantic components prove equally difficult. `Dynamicity` classification remains robust across languages and most model scales, with frontier models achieving 88.9%–98.9% accuracy. Frontier models achieve high scores, and most small models maintain reasonable performance on this component despite struggling with `topological` and `projective` relations. This resilience likely reflects the nature of the distinction: `dynamicity` involves a three-way categorical classification with strong correlations to verb semantics, making it more accessible than spatial primitives that require geometric reasoning.

Morphological Marking. Basque `spatial_markers` present a diagnostic case for morphological processing in spatial language understanding. Spatial relations in Basque are encoded through case suffixes that attach directly to nouns, requiring models to segment bound morphemes rather than identify free-standing tokens.

The contrast between model scales is pronounced. Small models show severe difficulties with Basque markers, with all three models achieving their lowest marker-type scores on this component. Frontier models improve substantially, though performance varies: Claude achieves comparable scores across languages (72.1%–76.4%), while GPT drops to 49.4% and Qwen to 17.5% for Basque markers despite strong performance on other surface elements.

This pattern suggests that segmenting and identifying case affixes as spatial encoding poses particular challenges, especially at smaller scales. This pattern could extend to other languages with spatial case marking, such as Finnish or Hungarian. Evaluation limited to prepositional languages would miss this systematic limitation entirely.

Topological. Primitives reveal asymmetric performance patterns. Closed containment, superposition, and subposition achieve moderate-to-high scores in frontier models, while open containment proves more challenging. The contact distinction remains difficult: [-contact] primitives generally underperform [+contact] across topological types, though the magnitude varies by model and language. The [open] versus [closed] distinction shows instability, with [open] containment achieving lower scores. The [top] specification shows variable performance, with scores dependent on lexical cues in the input. Small models exhibit systematic rather than random errors: [+contact] primitives consistently outperform [-contact] primitives, particularly for containment and superposition relations. However, overall scores remain lower than frontier models, with greater variance across primitive combinations and languages. A supplementary multimodal experiment shows no consistent improvement pattern overall (Table 14, Appendix E), but frontier models show selective gains for [+contact][containment][open] across all languages (+5.6 to +44.4 percentage points; Table 15), a pattern that does not extend to small models, where improvements are inconsistent (Table 16).

Projective. Projective relations present distinct evaluation challenges depending on frame type. Absolute frame relations—cardinal directions—rely on lexically explicit nouns, making them more tractable for text-only evaluation. Frontier models achieve consistently high scores on absolute frames in Spanish and Basque for Claude and GPT, though Qwen shows notable weakness even on these lexically explicit relations. Intrinsic and relative frame relations pose a more fundamental challenge: without visual context, these distinctions can be genuinely ambiguous in text alone—a sentence like "the fox is in front of the house" could describe either an intrinsic or a relative relation. Given this inherent ambiguity, we avoid strong claims about the relative difficulty of specific frame types, as observed patterns may reflect textual cues rather than spatial reasoning per se. Adding images does not resolve this—results show no consistent improvements across models or languages (Tables 17–18, Appendix E), consistent with recent findings on visual grounding for spatial distinctions (Tong et al., 2024).

7 Conclusion

We introduced a novel framework for spatial language understanding and a multilingual benchmark that addresses systematic gaps in existing evaluation approaches. By decomposing spatial relations into surface elements and semantic components, and by recognizing spatial markers beyond prepositions, our framework enables evaluation across typologically diverse languages.

Our evaluation of frontier and small-scale LLMs on Spanish, Basque, and Chinese reveals three key findings. First, surface element identification scores consistently exceed semantic classification scores, demonstrating that successful surface parsing does not entail spatial understanding. Second, morphological spatial marking poses persistent challenges: Basque case suffixes prove particularly difficult across model scales, with small models showing severe difficulties and frontier models demonstrating variable performance depending on architecture. Third, primitive decomposition using F1-score metrics exposes systematic patterns—contact asymmetries, containment geometry instability, frame-of-reference distinctions—that would be invisible under atomic labeling schemes.

These findings point to several research directions for advancing spatial language understanding in LLMs. First, expanding typological coverage to languages with diverse morphological marking systems is essential—current evaluation systematically excludes case-marking strategies, limiting our understanding of whether models comprehend spatial relations or succeed primarily on prepositional patterns. Second, component-wise evaluation frameworks provide diagnostic precision that atomic labels cannot, enabling separation of parsing failures from semantic failures and targeted analysis of systematic difficulties. Third, multimodal frameworks designed specifically for spatial reasoning merit dedicated investigation. Finally, projective relations require specialized methodologies for frame-of-reference distinctions, particularly for intrinsic and relative frames where textual ambiguity is inherent.

Advancing spatial language understanding in LLMs requires expanding evaluation beyond prepositional languages to determine whether models genuinely comprehend spatial meaning or succeed primarily on surface patterns. Our benchmark provides initial evidence for this necessity, and our component-wise framework offers a methodology that can be extended to additional languages and spatial phenomena in future research.

8 Limitations and Future Work

Limitations inspire future research directions. Our language sample, while typologically diverse, covers only three languages. The patterns we observe—particularly the difficulty with morphological marking—should be tested on additional languages with spatial case systems (e.g., Finnish, Hungarian) to confirm their generality.

Our multimodal evaluation provides initial evidence that visual grounding does not resolve semantic challenges, but was limited in scope. The experiment used canonical static images with text-only prompts, without systematic manipulation of visual properties, attention mechanisms, or multimodal integration strategies. A comprehensive understanding of why visual input fails to help would require controlled studies examining these factors —which is beyond the scope of this benchmark-focused study.

Our typology of spatial markers—adpositions, affixes, and spatial nouns—does not exhaust the crosslinguistic inventory. Less common strategies such as tonal marking could encode spatial meaning in some languages, but fall outside our current annotation scheme.

The primitive decomposition, while more fine-grained than existing schemes, does not exhaust spatial semantics. The `topological` component could be extended to capture relations like "between" or "among" and configurations involving encirclement or multiple grounds, which are not reducible to our current feature set. Similarly, the `projective` component could incorporate finer angular distinctions and language-specific absolute systems beyond cardinal directions. Our current treatment of relative frame relations also does not account for more fine-grained phenomena such as the variation in coordinate projection strategies across languages (Levinson, 2003; Zhang et al., 2025). Scaling the dataset to include Indigenous Languages of Australia (e.g., Pitjantjatjara, Murrinhpatha, Bininj Kunwok) would be especially interesting, as they are widely known for abstract cardinal and geomorphic terms for spatial reference, with egocentric terms not being attested much (Palmer et al., 2022).

Finally, the framework may serve as a basis for future studies of spatial systems, their components and their relationship to the topographic environments, an aspect that still remains under-studied (Palmer et al., 2019).

9 Ethical Considerations

This research was approved by the University of Melbourne Human Research Ethics Committee. Native speaker participants were recruited through personal networks and volunteered without financial compensation for this brief linguistic elicitation task. All participants provided informed consent and were informed that their anonymized sentence productions would be used for NLP research and made publicly available as part of a dataset. The dataset contains only anonymized linguistic annotations with no personally identifiable information.

Acknowledgements

We thank the native speaker experts who contributed sentence elicitations and verification for the Spanish, Basque, and Chinese portions of the dataset. We thank David Martinez Iraola for checking the Basque examples. EV acknowledges the Australian Research Council Discovery Early Career Research Award (Grant No. DE260100695). We gratefully acknowledge the Melbourne Data Analytics Platform (MDAP) at the University of Melbourne for supporting the conference registration, and the Faculty of Engineering and Information Technology at the University of Melbourne for supporting conference travel through the 2026 ENG&IT Conference Travel Grant. This research was supported by The University of Melbourne's Research Computing Services and the Spartan High Performance Computing system (Lafayette et al., 2016). We also thank the anonymous reviewers for their constructive feedback.

References

Malihe Alikhani, Valts Blukis, Parisa Kordjamshidi, Aishwarya Padmakumar, and Hao Tan. 2021. Proceedings of second international combined workshop on spatial language understanding and

grounded communication for robotics. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, Online. Association for Computational Linguistics.

Anthropic. 2024. The Claude model family: Claude 3.5 Haiku and Claude 3.5 Sonnet – Model Card Addendum. Technical report, Anthropic.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.

Barend Beekhuizen. 2025. Spatial relation marking across languages: Extraction, evaluation, analysis. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 571–585, Vienna, Austria. Association for Computational Linguistics.

Archna Bhatia, Yonatan Bisk, Parisa Kordjamshidi, and Jesse Thomason. 2019. Proceedings of the combined workshop on spatial language understanding (splu) and grounded communication for robotics (robonlp). In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Melissa Bowerman and Erik Pederson. 1992. Topological relations picture series. In Stephen C. Levinson, editor, *Space Stimuli Kit 1.2*, page 51. Max Planck Institute for Psycholinguistics, Nijmegen.

Denis Creissels. 2008. Spatial cases. In Andrej L. Malchukov and Andrew Spencer, editors, *The Oxford Handbook of Case*. Oxford University Press.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Redouane Djamouri, Waltraud Paul, and John Whitman. 2013. Postpositions vs prepositions in Mandarin Chinese: The articulation of disharmony. In Theresa Biberauer and Michelle Sheehan, editors, *Theoretical Approaches to Disharmonic Word Order*. Oxford University Press.

Ricardo Etxepare. 2013. Basque primary adpositions from a clausal perspective. *Catalan Journal of Linguistics*, 12:41–82.

Ricardo Etxepare and Bernard Oyharçabal. 2012. Datives and adpositions in Northeastern Basque. In Beatriz Fernández and Ricardo Etxepare, editors, *Variation in Datives: A Microcomparative Perspective*. Oxford University Press.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Martin Haspelmath. 2019. Differential place marking and differential object marking. *STUF – Language Typology and Universals*, 72(3):313–334.

José Ignacio Hualde and Jon Ortiz de Urbina. 2011. *A Grammar of Basque*. De Gruyter Mouton, Berlin.

C.-T. James Huang, Y.-H. Audrey Li, and Yafei Li. 2009. *The Syntax of Chinese*. Cambridge Syntax Guides. Cambridge University Press, Cambridge.

Parisa Kordjamshidi, Archna Bhatia, Malihe Alikhani, Jason Baldridge, Mohit Bansal, and Marie-Francine Moens. 2020. Proceedings of the third international workshop on spatial language understanding. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, Online. Association for Computational Linguistics.

Parisa Kordjamshidi, Archna Bhatia, James Pustejovsky, and Marie-Francine Moens. 2018. Proceedings of the first international workshop on spatial language understanding. In *Proceedings of the First International Workshop on Spatial Language Understanding*, New Orleans. Association for Computational Linguistics.

Parisa Kordjamshidi, Xin Eric Wang, Yue Zhang, Ziqiao Ma, and Mert Inan. 2024. Proceedings of the 4th workshop on spatial language understanding and grounded communication for robotics (splu-robonlp 2024). In *Proceedings of the 4th Workshop on Spatial Language Understanding and Grounded Communication for Robotics (SpLU-RoboNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Lev Lafayette, Greg Sauter, Linh Vu, and Bernard Meade. 2016. Spartan performance and flexibility: An HPC-Cloud chimera. In *OpenStack Summit, Barcelona*.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago.

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

Stephen C. Levinson and Sergio Meira. 2003. 'Natural concepts' in the spatial topological domain—adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3):485–516.

Stephen C. Levinson and David P. Wilkins. 2006. *Grammars of Space: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press, Cambridge.

Jingxia Lin. 2011. A figure's final location must be identifiable: Localizer distribution in Chinese motion expressions. In *Annual Meeting of the Berkeley Linguistics Society*, pages 242–256.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632, Vienna, Austria. Association for Computational Linguistics.

Michał Olek and Maciej Piasecki. 2024. Three-stage extraction of spatial relationships using markers. In *Advances in Computational Collective Intelligence*, pages 159–172, Cham. Springer Nature Switzerland.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Bill Palmer, Joe Blythe, Alice Gaby, Dorothea Hoffmann, and Maïa Ponsonnet. 2019. Geospatial natural language in indigenous Australia: Research priorities. In *Proceedings of the Workshop on Speaking of Location 2019: Communicating about Space*, volume 2455 of *CEUR Workshop Proceedings*, pages 17–27, Regensburg, Germany. CEUR-WS.org.

Bill Palmer, Dorothea Hoffmann, Joe Blythe, Alice Gaby, Bill Pascoe, and Maïa Ponsonnet. 2022. Frames of spatial reference in five Australian languages. *Spatial Cognition & Computation*, 22(3–4):225–263.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Thomas Stolz, Sander Lestrade, and Christel Stolz. 2014. *The Crosslinguistics of Zero-Marking of Spatial Relations*. Number 15 in Studia Typologica. De Gruyter Mouton, Berlin.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. *Preprint*, arXiv:2401.06209.

Morgan Ulinski, Bob Coyne, and Julia Hirschberg. 2019. SpatialNet: A declarative resource for spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 61–70, Minneapolis, Minnesota. Association for Computational Linguistics.

I.-hao Woo. 2021. On preverbal zai in Mandarin Chinese: Its progressive and prepositional functions. *Linguistics*, 59(3):513–539.

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. *Preprint*, arXiv:2410.17385.

A Full Annotation Examples

A.1 Topological Example: Bird on Ground

This section provides the complete annotations for Spanish and Chinese descriptions of the spatial scene shown in Figure 1 (the bird on the ground). The Basque annotation appears in the main text (Figure 1).

```
{‘‘sentence’’: ‘‘El pájaro está en el suelo.’’,
 ‘‘language’’: ‘‘es’’,
 ‘‘figure’’: ‘‘El pájaro’’,
 ‘‘ground’’: ‘‘el suelo’’,
 ‘‘spatial_predicate’’:
    {‘‘text’’: ‘‘está’’, ‘‘type’’: ‘‘copula’’},
 ‘‘spatial_markers’’: [
    {‘‘text’’: ‘‘en’’, ‘‘type’’: ‘‘adposition’’}],
 ‘‘dynamicity’’: ‘‘static’’,
 ‘‘stasis’’:
    {‘‘topological’’:
        ‘‘[+contact][superposition]’’,
     ‘‘projective’’: {‘‘type’’: ‘‘’’, ‘‘value’’:
        ‘‘’’}}}
```

Figure 2: Spanish annotation for “El pájaro está en el suelo” (The bird is on the ground). Spanish encodes the spatial relation using the copula *está* and a single adposition *en*.

A.1.1 Chinese

Cross-linguistic observations. The three languages encode the same topological relation ([+contact][superposition]) using different surface strategies:

- **Spanish** relies on a copula (*está*) and a single general-purpose preposition (*en*) that does not specify the precise topological relationship.
- **Basque** uses a lexical verb (*dago*) and two markers: a spatial noun (*gain* ‘top/surface’) that specifies the topological relationship, plus a locative case affix (*-an*) marking the `ground`.
- **Chinese** has no explicit `spatial_predicate` but employs two markers: a preposition (*zài*) marking general spatial location, and a spatial noun (*shàng* ‘top/on’) specifying the topological relationship.

```
{‘‘sentence’’: ‘‘Niao zai di shang.’’,
 ‘‘translation’’: ‘‘The bird is on the ground’’,
 ‘‘language’’: ‘‘zh’’,
 ‘‘figure’’: ‘‘niao’’,
 ‘‘ground’’: ‘‘di shang’’,
 ‘‘spatial_predicate’’:
    {‘‘text’’: ‘‘’’, ‘‘type’’: ‘‘’’},
 ‘‘spatial_markers’’: [
    {‘‘text’’: ‘‘zai’’, ‘‘type’’: ‘‘adposition’’},
    {‘‘text’’: ‘‘shang’’, ‘‘type’’:
        ‘‘spatial_noun’’}],
 ‘‘dynamicity’’: ‘‘static’’,
 ‘‘stasis’’:
    {‘‘topological’’:
        ‘‘[+contact][superposition]’’,
     ‘‘projective’’: {‘‘type’’: ‘‘’’, ‘‘value’’:
        ‘‘’’}}}
```


Figure 3: Chinese annotation for “鸟在地上” (Niǎo zài dì shàng / The bird is on the ground). Chinese uses no explicit `spatial_predicate`, but employs two `spatial_markers`: the preposition *zài* and the spatial noun *shàng* (‘top/on’).

This variation demonstrates that evaluating only prepositional markers would miss crucial spatial encoding strategies. The semantic annotation (`topological` and `dynamicity`), however, remains consistent across languages, showing that our primitive-based approach captures meaning independently of surface form.

A.2 Projective Examples: Frame-of-Reference Relations

This section provides examples of `projective` relations requiring different frames of reference. Unlike the bird example, these relations involve directional specification.

A.2.1 Intrinsic Frame: Fox in Front of House

Why this is projective. Unlike the bird-on-ground example, which can be described purely through topological primitives, the fox-house relation requires specifying a direction. The house has an inherent orientation—a canonical front side (typically where the door is located)—and the fox’s position is described relative to this intrinsic property of the `ground` object. The observer’s position is irrelevant; the relation holds regardless of where one views the scene from.

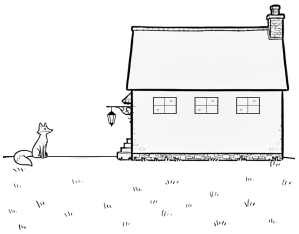

Figure 4: Visual stimulus for intrinsic frame relation: the fox’s position is described relative to the house’s inherent front facet.


```
{‘‘sentence’’: ‘‘Azeria etxearen aurrean dago.’’,
 ‘‘translation’’: ‘‘The fox is in front of the
     house’’,
 ‘‘language’’: ‘‘eu’’,
 ‘‘figure’’: ‘‘Azeria’’,
 ‘‘ground’’: ‘‘etxearen aurrean’’,
 ‘‘spatial_predicate’’:
    {‘‘text’’: ‘‘dago’’, ‘‘type’’: ‘‘copula’’},
 ‘‘spatial_markers’’: [
    {‘‘text’’: ‘‘aurre’’, ‘‘type’’:
        ‘‘spatial_noun’’},
    {‘‘text’’: ‘‘–an’’, ‘‘type’’: ‘‘affix’’}],
 ‘‘dynamicity’’: ‘‘static’’,
 ‘‘stasis’’:
    {‘‘topological’’: ‘‘’’,
     ‘‘projective’’: {‘‘type’’: ‘‘intrinsic’’,
        ‘‘value’’: ‘‘front’’}}}
```


Figure 5: Basque annotation for “Azeria etxearen aurrean dago” (The fox is in front of the house). This is an intrinsic frame relation: the spatial description relies on identifying the house’s inherent front facet. The spatial noun *aurre* (‘front’) combined with the locative case *-an* encodes this directional relationship.

A.2.2 Relative Frame Example

Relative frame relations depend on the observer's viewpoint. For example, "The ball is to the left of the tree" describes the ball's position using coordinates derived from the observer's bodily axes. If the observer moves to the opposite side of the tree, the same physical configuration might be described as "to the right of the tree."

Our dataset includes relative frame examples with all four directional values: left, right, front, and back.

A.2.3 Absolute Frame Example

Absolute frame relations use fixed environmental bearings. For example, "The village is south of the mountain" employs cardinal directions that remain constant regardless of observer position or object orientation.

Our dataset includes absolute frame examples with all four cardinal values: north, south, east, and west.

A.3 Dynamicity Variation

The same visual configuration can be described with different `dynamicity` values depending on whether motion is involved and, if so, in which direction.

Static: "The bird is on the ground" describes a stable spatial relation without motion.

Goal: "The bird lands on the ground" or "The bird flies onto the ground" describes motion toward the `ground` as destination. The `topological` value ([+contact][superposition]) describes the **final** spatial state after the motion is complete.

Source: "The bird takes off from the ground" or "The bird flies from the ground" describes motion away from the `ground` as departure point. The `topological` value describes the **initial** spatial state before the motion begins.

This systematic variation allows us to evaluate whether models understand that `dynamicity` and `topological` are independent but interacting dimensions: the same topological relationship can occur in static contexts or as the source/goal of motion events.

B Primitive Values and Examples

C Annotation Guidelines

Figure

The noun phrase denoting the entity whose location or motion is being described. Extract the complete NP in its base form: include all determiners, modifiers, and complements that belong to the phrase itself, but exclude independent adpositions. For contractions (e.g., 'al' = 'a' + 'el'), extract only the NP component ('el').

Ground

The reference object, site, or region relative to which the Figure's location or motion is specified. Extract the complete noun phrase that serves as the landmark, including all determiners, modifiers, and complements. This includes any lexical elements with spatial meaning (like 'top', 'front', 'right', 'side', 'cima', 'lado', '上面') when they function as the head noun of the phrase. The entire noun phrase built around such spatial nouns (e.g., 'la cima del edificio', 'the top of the mountain', '大楼的上面') must be extracted as the complete ground.

Spatial Predicate

The predicate that expresses the static relation or motion event linking Figure and Ground. Extract the complete predicate as it appears in the sentence, including all its grammatical elements and any complements that are part of the verbal unit itself. Do NOT include spatial markers that are connected to the Ground. Includes both 'text' (the surface form) and 'type' (grammatical category). If the construction does not have a spatial predicate, the entire field should be an empty string "".

- verb: Lexical verbs that denote actions, states, or processes (e.g., lies, sits, runs, goes, remains).
- copula: Grammatical elements that link the figure to the ground, typically expressing attribution rather than action (e.g., is, was).

Spatial Markers

A list of elements that encode the spatial relationship between a figure (the entity whose location or motion is being described) and a ground (the reference object). Only include elements that are grammatically linked to a ground NP. Each marker is

Primitives	Example
Topological	
[+contact][containment][open]	The cat is in the box
[-contact][containment][open]	The fish is in the fish tank
[+contact][containment][closed]	The cat is in the house
[-contact][containment][closed]	The fly is flying inside the house
[+contact][attachment]	The magnet is stuck to the refrigerator
[+contact][superposition]	The bird is on the ground
[-contact][superposition]	The bird is flying above the house
[+contact][superposition][top]	The cat is at the top of the mountain
[-contact][superposition][top]	The helicopter is hovering above the roof of the building
[+contact][subposition]	The cat is under the blanket
[-contact][subposition]	The cat is under the table
Projective (angular)	
Absolute	
[absolute][east]	The car is to the east of the church
[absolute][north]	The car is to the north of the church
[absolute][south]	The car is to the south of the church
[absolute][west]	The car is to the west of the church
Intrinsic	
[intrinsic][back]	The fox is behind the house
[intrinsic][front]	The fox is in front of the house
Relative	
[relative][back]	The cat is behind the box
[relative][front]	The cat is in front of the box
[relative][left]	The cat is to the left of the box
[relative][right]	The cat is to the right of the box

Table 4: Primitive components for topological and angular (projective) spatial relations with representative examples.

recorded separately with 'text' and 'type'. Markers must be listed in their order of appearance in the sentence. If the construction has no spatial markers, the entire field should be an empty list [].

- adposition: Prepositions, postpositions, or fixed multi-word constructions that have grammaticalized as inseparable units (e.g., 'on', 'in', 'at', 'on top of', 'encima de', '在'). When contractions or fused forms incorporate elements that belong to the ground NP (e.g., Spanish 'del' = 'de' + 'el'), report only the spatial marker portion (e.g., 'de'), while the ground NP element remains part of the ground. For multi-word adpositions, determine if the expression is grammaticalized as a fixed unit: if it does not accept internal modification (e.g., 'on top of', 'encima de'), mark it as a single adposition; if the spatial noun retains nominal behavior and accepts modification (e.g., 'on the top of', 'en la cima de'), separate the adposition from the spatial noun.

- affix: Case markers or other morphological affixes that encode spatial meaning (e.g., Latin ablative '-o', accusative '-um', Basque inessive '-n').

- spatial_noun: Lexical nouns with spatial meaning that function as the head of the ground NP. Extract only the stem or head noun itself (e.g., 'top', 'cima', '上面', 'gain'), excluding determiners, modifiers, and inflectional morphology. Examples: 'cima' (from 'la cima del edificio'), 'gain' (from 'mahaiaren gainean').

Dynamicity

The temporal aspect of the spatial relation, indicating whether the relation is static or involves motion.

- static: The Figure's spatial relation to the Ground is stable; no translational motion occurs.

- source: The Figure moves away from the Ground, which serves as the point of departure of the motion.

- goal: The Figure moves toward the Ground, which serves as the final location of the motion.

Stasis

The spatial relationship between Figure and Ground. ALWAYS include both 'topological' and

'projective' keys.

dynamic_rule: For motion: if Figure moves AWAY FROM Ground, describe INITIAL state; if TOWARD Ground, describe FINAL state.

topological: Topological relationship as concatenated primitives in brackets (e.g., '[+contact][containment][open]'). Contact ([+contact] or [-contact]) must ALWAYS be first. Use empty string if projective is used.

- contact: [+contact] or [-contact] - physical contact between Figure and Ground. ALWAYS first.
- containment: [containment][open] (partial enclosure) or [containment][closed] (full enclosure)
- attachment: [attachment] - mechanically fastened
- superposition: [superposition] - Figure above Ground. Add [top] if on uppermost surface.
- subposition: [subposition] - Figure below Ground

projective: Angular relationships via frames of reference. If topological is used, include this key with empty strings.

type: 'absolute' (fixed bearings), 'relative' (viewer-dependent), or 'intrinsic' (object's inherent parts)

value: north | south | east | west | left | right | front | back

D Language-Specific Considerations

Each language presented distinct analytical challenges.

Spanish. Contracted forms combining prepositions with articles (e.g., *del* from *de* + *el*) required splitting to isolate the `spatial_marker`. Multiword adpositions such as *encima de* ('on top of') were treated as single markers when grammaticalized as fixed units.

Basque. Basque employs an agglutinative system where spatial relations are encoded through case suffixes—inessive *-n*, allative *-ra*, ablative *-tik*—combined with spatial nouns (Hualde and Ortiz de Urbina, 2011). Spatial nouns like *gain* ('top') frequently combine with these case markers and are treated as nouns following Etxepare (2013). Directional postpositions such as *behera* ('down') were analyzed as adpositions following Etxepare and Oyharçabal (2012). Genitive cases functioning as linkers between reference objects and spatial nouns were not annotated as `spatial_markers`, as they establish possessive relationships rather than contributing spatial information per se.

Chinese. Chinese presents particular challenges due to the grammaticalization status of elements like *lǐ* ('inside'), *shàng* ('top/on'), and *xià* ('under'). These forms—termed 'localizers' in the literature—derive historically from nouns, but their synchronic category remains debated, having been analyzed variously as NP enclitics, locative particles, postpositions, or as a subclass of nouns (Huang et al., 2009; Lin, 2011; Djamouri et al., 2013). We annotate these elements as `spatial_noun` markers, consistent with their nominal etymology and syntactic behavior. Regarding *zài*, we follow Woo (2021) in treating it consistently as a spatiotemporal preposition that introduces the Ground. Separately, directional complements appearing in verb-directional constructions (e.g., *chūlái* 'exit-come' in *zuān le chūlái* 'burrowed out') were annotated as part of the `spatial_predicate`, as these elements modify the motion event rather than marking the `ground`.

E Tables

Table 5: Overall frontier results for Claude-3.5-Sonnet, GPT-4o, and Qwen2.5-VL-72B (%). Figure/Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score.

	Spanish			**Basque**			**Chinese**		
	Claude	GPT-4o	Qwen	Claude	GPT-4o	Qwen	Claude	GPT-4o	Qwen
Figure	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.8 ± 0.4	96.5 ± 0.4	83.0 ± 1.1	99.0 ± 0.8	96.3 ± 0.8	97.8 ± 0.6
Ground	85.0 ± 0.5	83.5 ± 0.5	86.3 ± 0.0	99.3 ± 0.4	97.3 ± 0.1	83.7 ± 0.3	99.3 ± 0.6	100.0 ± 0.0	99.1 ± 0.3
Predicate	99.2 ± 0.0	99.0 ± 0.2	94.3 ± 0.7	87.1 ± 0.8	90.3 ± 2.6	82.0 ± 0.5	83.4 ± 3.6	61.9 ± 1.5	68.2 ± 0.3
Markers	74.2 ± 0.5	76.4 ± 1.6	70.5 ± 0.6	72.1 ± 1.6	49.4 ± 0.9	17.5 ± 0.5	76.4 ± 1.3	70.8 ± 4.2	61.3 ± 0.7
Dynamicity	98.9 ± 0.9	97.4 ± 0.9	92.1 ± 0.0	97.9 ± 0.9	96.3 ± 0.9	89.4 ± 0.9	95.8 ± 0.9	97.9 ± 0.9	88.9 ± 0.0
Topological	73.7 ± 2.2	66.9 ± 2.4	49.4 ± 1.9	64.6 ± 1.5	68.9 ± 2.0	42.8 ± 0.9	57.0 ± 1.9	50.1 ± 3.0	50.2 ± 0.3
Projective	87.2 ± 1.0	92.2 ± 1.0	60.6 ± 2.5	79.7 ± 0.0	82.8 ± 3.7	40.6 ± 2.5	74.0 ± 0.9	54.6 ± 1.6	65.0 ± 0.0

Table 6: Overall small model results for Claude-3.5-Haiku, GPT-4o-mini, and Qwen2.5-VL-7B (%). Figure/-Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score (with standard deviation).

	Spanish			**Basque**			**Chinese**		
	Haiku	GPT-mini	Qwen-7B	Haiku	GPT-mini	Qwen-7B	Haiku	GPT-mini	Qwen-7B
Figure	97.3 ± 0.0	96.4 ± 0.0	94.8 ± 0.5	86.0 ± 0.5	40.8 ± 2.1	85.2 ± 1.2	88.2 ± 0.0	65.8 ± 1.0	91.0 ± 1.5
Ground	86.5 ± 0.1	89.5 ± 1.2	88.3 ± 0.9	93.0 ± 0.3	78.8 ± 1.0	67.2 ± 2.6	99.2 ± 0.5	96.4 ± 0.4	43.6 ± 0.5
Predicate	92.7 ± 0.2	74.7 ± 0.4	75.8 ± 0.5	84.1 ± 0.0	89.2 ± 1.2	88.3 ± 2.4	69.8 ± 0.0	66.9 ± 0.8	59.2 ± 0.5
Markers	83.5 ± 0.7	49.1 ± 0.5	63.4 ± 1.4	29.6 ± 0.6	15.3 ± 0.7	18.5 ± 2.5	73.7 ± 1.3	67.1 ± 0.7	29.3 ± 2.1
Dynamicity	95.2 ± 0.0	89.4 ± 2.4	75.7 ± 0.9	94.7 ± 0.9	77.8 ± 2.7	58.2 ± 0.9	93.7 ± 0.0	76.7 ± 0.9	43.9 ± 0.9
Topological	50.3 ± 1.6	30.4 ± 0.6	30.6 ± 2.3	42.6 ± 2.1	26.4 ± 1.3	21.9 ± 0.5	31.8 ± 0.3	24.9 ± 0.2	22.6 ± 3.8
Projective	78.7 ± 1.3	39.9 ± 0.9	32.8 ± 3.2	83.0 ± 1.3	31.9 ± 1.1	26.0 ± 2.7	72.3 ± 2.8	41.0 ± 1.2	28.8 ± 5.1

Table 7: Topological Breakdown by Primitive Combination frontier (with standard deviation)

	Claude-3.5-Sonnet			GPT-4o			Qwen2.5-72B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[+contact][containment][open]	70.4 ± 12.8	55.6 ± 0.0	77.8 ± 0.0	63.0 ± 6.4	54.1 ± 4.6	77.8 ± 0.0	88.9 ± 0.0	96.3 ± 6.4	88.9 ± 0.0
[-contact][containment][open]	37.0 ± 6.4	44.4 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	44.4 ± 0.0	33.3 ± 0.0	44.4 ± 0.0	55.6 ± 0.0	44.4 ± 0.0
[+contact][containment][closed]	77.8 ± 0.0	81.5 ± 6.4	80.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	68.5 ± 17.0	77.8 ± 0.0	74.1 ± 3.2	77.8 ± 0.0
[-contact][containment][closed]	77.8 ± 0.0	51.9 ± 6.4	55.6 ± 0.0	66.7 ± 0.0	70.4 ± 6.4	66.7 ± 0.0	37.0 ± 6.4	44.4 ± 0.0	44.4 ± 0.0
[+contact][attachment]	100.0 ± 0.0	80.0 ± 0.0	60.0 ± 0.0	77.8 ± 9.6	80.0 ± 0.0	40.0 ± 0.0	83.0 ± 5.1	45.9 ± 5.1	48.9 ± 1.9
[+contact][superposition]	72.2 ± 9.6	83.3 ± 0.0	95.6 ± 3.8	74.1 ± 12.8	86.7 ± 3.8	63.0 ± 12.8	53.7 ± 3.2	74.4 ± 9.6	96.3 ± 6.4
[-contact][superposition]	80.0 ± 0.0	80.0 ± 0.0	73.3 ± 11.5	43.3 ± 0.0	44.4 ± 1.9	41.1 ± 1.9	55.6 ± 0.0	37.8 ± 7.7	55.6 ± 0.0
[+contact][superposition][top]	95.6 ± 3.8	53.3 ± 11.5	46.7 ± 0.0	88.9 ± 7.7	100.0 ± 0.0	47.8 ± 13.5	80.0 ± 0.0	56.3 ± 12.8	66.7 ± 0.0
[-contact][superposition][top]	74.1 ± 2.6	40.0 ± 7.7	44.4 ± 0.0	66.7 ± 0.0	35.6 ± 0.0	44.4 ± 0.0	26.7 ± 0.0	26.7 ± 0.0	57.8 ± 0.0
[+contact][subposition]	83.3 ± 0.0	83.3 ± 0.0	80.0 ± 0.0	48.1 ± 30.6	63.0 ± 6.4	36.7 ± 17.3	100.0 ± 0.0	63.3 ± 0.0	100.0 ± 0.0
[-contact][subposition]	66.7 ± 0.0	100.0 ± 0.0	50.0 ± 0.0	74.1 ± 6.4	85.2 ± 3.2	70.0 ± 17.3	44.4 ± 9.6	38.9 ± 9.6	50.0 ± 0.0

Table 8: Topological Breakdown by Primitive Combination small (with standard deviation)

	Claude-3.5-Haiku			GPT-4o-mini			Qwen2.5-7B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[+contact][containment][open]	88.9 ± 0.0	82.2 ± 11.5	88.9 ± 0.0	92.6 ± 6.4	100.0 ± 0.0	88.9 ± 0.0	49.6 ± 12.6	42.2 ± 14.6	46.7 ± 13.3
[-contact][containment][open]	40.7 ± 6.4	51.9 ± 12.8	40.7 ± 6.4	66.7 ± 0.0	59.3 ± 6.4	55.6 ± 0.0	38.5 ± 20.5	28.1 ± 6.8	22.2 ± 0.0
[+contact][containment][closed]	77.8 ± 0.0	74.1 ± 6.4	68.9 ± 0.0	63.0 ± 6.4	74.1 ± 6.4	66.7 ± 0.0	65.9 ± 11.4	43.0 ± 14.3	46.7 ± 14.7
[-contact][containment][closed]	40.7 ± 6.4	48.1 ± 6.4	43.5 ± 0.0	33.3 ± 0.0	44.4 ± 0.0	33.3 ± 0.0	34.1 ± 19.2	27.7 ± 13.4	3.7 ± 6.4
[+contact][attachment]	100.0 ± 0.0	80.0 ± 0.0	40.0 ± 0.0	60.0 ± 0.0	40.0 ± 0.0	38.7 ± 2.2	57.8 ± 4.8	11.1 ± 9.6	36.7 ± 5.8
[+contact][superposition]	66.7 ± 0.0	63.3 ± 8.8	84.4 ± 3.8	40.0 ± 0.0	53.3 ± 0.0	50.0 ± 5.8	27.8 ± 14.7	36.3 ± 15.1	48.5 ± 16.2
[-contact][superposition]	27.8 ± 1.9	51.1 ± 10.2	40.0 ± 0.0	40.0 ± 0.0	35.6 ± 7.7	26.7 ± 0.0	47.8 ± 1.9	32.2 ± 16.8	38.9 ± 9.6
[+contact][superposition][top]	100.0 ± 0.0	74.3 ± 4.7	83.3 ± 0.0	63.0 ± 12.8	63.0 ± 25.7	57.0 ± 1.3	62.2 ± 7.7	68.9 ± 16.8	71.1 ± 15.6
[-contact][superposition][top]	65.1 ± 18.0	54.8 ± 5.1	44.4 ± 0.0	44.4 ± 0.0	22.2 ± 0.0	66.7 ± 0.0	70.7 ± 8.0	55.9 ± 13.0	51.1 ± 21.4
[+contact][subposition]	63.3 ± 4.8	76.7 ± 8.8	100.0 ± 0.0	57.8 ± 0.0	40.0 ± 0.0	40.0 ± 0.0	63.0 ± 6.4	29.6 ± 12.8	48.1 ± 11.6
[-contact][subposition]	50.0 ± 16.7	66.7 ± 0.0	33.3 ± 0.0	27.4 ± 9.0	18.5 ± 3.2	0.0 ± 0.0	64.8 ± 3.2	37.0 ± 12.8	42.6 ± 22.5

Table 9: Projective Breakdown by Primitive Combination frontier (with standard deviation)

	Claude-3.5-Sonnet			GPT-4o			Qwen2.5-72B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[absolute][east]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	83.3 ± 16.7	50.0 ± 0.0	77.8 ± 9.6	11.1 ± 9.6	66.7 ± 0.0
[absolute][north]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	55.6 ± 19.2	94.4 ± 9.6	66.7 ± 0.0	100.0 ± 0.0
[absolute][south]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	77.8 ± 19.2	55.6 ± 9.6	83.3 ± 0.0	33.3 ± 0.0	100.0 ± 0.0
[absolute][west]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 0.0	66.7 ± 0.0	27.8 ± 9.6	66.7 ± 0.0
[intrinsic][back]	66.7 ± 0.0	100.0 ± 0.0	66.7 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	77.8 ± 19.2	0.0 ± 0.0	16.7 ± 0.0	16.7 ± 0.0
[intrinsic][front]	100.0 ± 0.0	66.7 ± 0.0	66.7 ± 0.0	94.4 ± 9.6	100.0 ± 0.0	100.0 ± 0.0	38.9 ± 9.6	16.7 ± 16.7	33.3 ± 0.0
[relative][back]	50.0 ± 0.0	50.0 ± 0.0	38.9 ± 9.6	77.8 ± 9.6	50.0 ± 0.0	38.9 ± 9.6	11.1 ± 19.2	0.0 ± 0.0	33.3 ± 0.0
[relative][front]	55.6 ± 9.6	33.3 ± 0.0	33.3 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	44.4 ± 9.6	33.3 ± 0.0	33.3 ± 0.0	66.7 ± 0.0
[relative][left]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	83.3 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
[relative][right]	100.0 ± 0.0	100.0 ± 0.0	83.3 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	38.9 ± 9.6	100.0 ± 0.0	100.0 ± 0.0	66.7 ± 0.0

Table 10: Projective Breakdown by Primitive Combination small (with standard deviation)

	Claude-3.5-Haiku			GPT-4o-mini			Qwen2.5-7B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[absolute][east]	100.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	83.3 ± 0.0	61.1 ± 9.6	100.0 ± 0.0	50.0 ± 0.0	50.0 ± 16.7	46.3 ± 17.9
[absolute][north]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	88.9 ± 9.6	66.7 ± 0.0	94.4 ± 9.6	66.7 ± 0.0	38.9 ± 9.6	44.4 ± 9.6
[absolute][south]	100.0 ± 0.0	94.4 ± 9.6	100.0 ± 0.0	83.3 ± 0.0	33.3 ± 0.0	88.9 ± 9.6	38.9 ± 9.6	44.4 ± 9.6	55.6 ± 19.2
[absolute][west]	100.0 ± 0.0	100.0 ± 0.0	77.8 ± 19.2	77.8 ± 9.6	5.6 ± 9.6	88.9 ± 9.6	38.9 ± 9.6	38.9 ± 9.6	27.8 ± 9.6
[intrinsic][back]	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	11.1 ± 9.6	11.1 ± 9.6	0.0 ± 0.0
[intrinsic][front]	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	22.2 ± 25.5	35.2 ± 14.0	5.6 ± 9.6
[relative][back]	100.0 ± 0.0	100.0 ± 0.0	66.7 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	50.0 ± 16.7	55.6 ± 19.2	33.3 ± 16.7
[relative][front]	100.0 ± 0.0	100.0 ± 0.0	88.9 ± 19.2	100.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	77.8 ± 19.2	50.0 ± 28.9	55.6 ± 19.2
[relative][left]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	61.1 ± 34.7	66.7 ± 33.3	77.8 ± 38.5
[relative][right]	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	61.1 ± 25.5	11.1 ± 9.6	22.2 ± 9.6

Table 11: Chinese results with unified prompt. Figure/Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score.

	Chinese		
	Claude	GPT-4o	Qwen
Figure	100.0	99.4	99.4
Ground	85.3	89.9	92.9
Predicate	86.7	69.6	65.4
Markers	52.6	41.5	44.4
Dynamicity	98.4	100.0	90.5
Topological	51.3	28.3	36.3
Projective	66.7	50.0	40.3

Table 12: Overall text-only models (%). Figure/Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score (with standard deviation).

	Spanish			**Basque**			**Chinese**		
	Gemma-12B	Qwen-14B	DeepSeek-14B	Gemma-12B	Qwen-14B	DeepSeek-14B	Gemma-12B	Qwen-14B	DeepSeek-14B
Figure	96.4 ± 0.0	99.5 ± 0.0	66.0 ± 3.2	68.4 ± 0.0	62.6 ± 1.2	47.5 ± 2.9	74.2 ± 0.8	78.0 ± 0.0	33.5 ± 7.2
Ground	86.4 ± 0.1	82.2 ± 0.6	82.6 ± 1.0	76.1 ± 0.1	80.0 ± 0.3	76.0 ± 3.5	100.0 ± 0.0	99.1 ± 0.4	92.1 ± 0.5
Predicate	89.6 ± 0.1	73.8 ± 0.4	79.9 ± 1.3	98.4 ± 0.0	64.5 ± 1.0	86.8 ± 1.2	64.2 ± 0.3	52.5 ± 2.1	56.8 ± 1.2
Markers	72.3 ± 0.3	69.2 ± 1.5	77.6 ± 3.4	13.3 ± 0.2	21.1 ± 0.5	14.5 ± 1.7	70.5 ± 0.7	67.3 ± 0.5	47.4 ± 6.4
Dynamicity	93.7 ± 0.0	90.5 ± 0.0	97.9 ± 1.8	90.5 ± 0.0	79.4 ± 0.0	69.3 ± 3.3	88.9 ± 0.0	84.1 ± 0.0	87.3 ± 1.6
Topological	41.0 ± 1.0	36.4 ± 0.1	38.8 ± 2.7	38.7 ± 4.3	24.3 ± 5.2	21.5 ± 1.9	23.1 ± 6.5	32.6 ± 0.6	32.4 ± 6.6
Projective	48.3 ± 1.6	65.4 ± 5.5	54.1 ± 8.3	44.1 ± 3.0	30.2 ± 5.6	18.2 ± 3.0	31.4 ± 2.7	52.8 ± 5.6	39.8 ± 9.1

Table 13: Evaluation Performance by Component (multimodal) (%). Figure/Ground: Levenshtein similarity; Predicate/Markers: avg. Levenshtein/type match; Dynamicity: accuracy; Topological/Projective: F1-score.

	Claude			**Haiku**			**GPT-4o**			**GPT-mini**			**Qwen**			**Qwen-7B**		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
Figure	97.3	100.0	100.0	96.4	86.5	93.9	100.0	96.1	100.0	88.6	41.2	60.1	99.5	89.2	100.0	78.2	69.8	74.0
Ground	79.0	77.3	91.4	86.0	60.5	97.7	85.4	98.7	100.0	84.3	95.7	95.0	83.6	92.4	91.6	80.1	51.5	66.9
Predicate	99.2	91.0	67.7	86.2	88.7	69.3	98.4	96.6	64.2	48.0	89.2	59.9	88.2	85.7	62.2	59.9	35.2	37.5
Markers	73.0	50.7	83.2	74.3	17.8	66.0	78.9	42.0	67.8	39.8	15.0	50.3	71.6	15.6	55.0	54.7	16.3	28.0
Dynamicity	85.7	100.0	95.2	93.7	95.2	88.9	93.7	96.8	96.8	52.4	85.7	82.5	90.5	79.4	92.1	69.8	60.3	55.6
Topological	62.9	70.1	58.9	33.1	29.9	33.7	63.9	58.1	50.4	27.8	11.8	28.0	47.1	45.6	47.8	33.0	26.0	25.5
Projective	78.6	81.8	70.8	65.0	70.0	47.7	90.0	61.8	54.4	39.6	55.0	43.3	63.3	50.0	66.7	38.4	33.9	39.1

Table 14: Performance Difference: Multimodal vs Text-only (Multimodal − Text-only, percentage points)

	Claude-3.5-Haiku			Claude-3.5-Sonnet			GPT-4o			GPT-4o-mini			Qwen2.5-72B			Qwen2.5-7B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
Topological	-18.7	-12.7	+1.9	-8.1	+6.3	+3.7	-5.6	-9.6	+3.4	-1.9	-13.1	+3.0	-4.2	+3.0	-2.1	+4.7	+3.4	-0.7
Projective	-14.4	-12.8	-25.9	-8.1	+2.1	-2.6	-3.3	-25.3	-0.1	-0.9	+24.0	+1.3	+0.0	+6.7	+1.7	+2.0	+4.9	+12.8

Table 15: Topological Breakdown Difference: Multimodal vs Text-only (big) (percentage points)

	Claude-3.5-Sonnet			GPT-4o			Qwen2.5-72B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[+contact][containment][open]	+22.2	+44.4	+11.1	+38.9	+27.8	+22.2	+11.1	+5.6	+11.1
[-contact][containment][open]	+5.6	+0.0	-11.1	+22.2	+11.1	+22.2	+22.2	+11.1	+22.2
[+contact][containment][closed]	+0.0	-5.6	-11.1	-16.7	-13.9	-27.8	-11.1	+5.6	-11.1
[-contact][containment][closed]	+11.1	+5.6	-11.1	-11.1	-11.1	-22.2	-5.6	+0.0	-11.1
[+contact][attachment]	+0.0	+20.0	+0.0	-13.9	-2.2	-13.3	-23.3	+6.7	+1.7
[+contact][superposition]	-16.7	-11.1	-10.0	-8.9	-14.4	-0.0	+25.0	+17.2	-22.2
[-contact][superposition]	-46.7	-13.3	-36.7	+13.3	+10.6	+10.6	-5.6	+10.0	-12.2
[+contact][superposition][top]	-30.0	+23.3	+46.7	-23.3	-66.7	+20.0	+0.0	+4.4	+13.3
[-contact][superposition][top]	-24.4	+2.2	+15.6	+5.6	-13.3	-8.9	+22.2	+22.2	+0.0
[+contact][subposition]	-23.3	-3.3	-20.0	-16.1	-1.1	-6.7	-60.0	-23.3	-40.0
[-contact][subposition]	-33.3	-50.0	-16.7	-25.0	-33.3	-15.0	+8.3	-25.0	+0.0

Table 16: Topological Breakdown Difference: Multimodal vs Text-only (small) (percentage points)

	Claude-3.5-Haiku			GPT-4o-mini			Qwen2.5-7B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[+contact][containment][open]	+0.0	-12.2	+11.1	+5.6	-66.7	+11.1	+41.1	-8.6	+37.8
[-contact][containment][open]	+0.0	+0.0	+16.7	+0.0	-61.1	+11.1	+13.3	+37.8	+44.4
[+contact][containment][closed]	-11.1	-22.2	-2.2	+11.1	-38.9	+0.0	-5.6	+26.7	+6.1
[-contact][containment][closed]	-5.6	+0.0	-10.2	+5.6	-33.3	+0.0	-11.1	+14.1	+22.2
[+contact][attachment]	+0.0	-13.3	+0.0	-8.3	-40.0	+1.9	+0.6	+36.1	+26.7
[+contact][superposition]	-33.3	-40.0	-61.1	+13.3	+0.0	-8.3	+19.4	+5.0	+23.9
[-contact][superposition]	+26.7	+3.3	+33.3	-10.0	+15.6	-2.2	-3.3	+0.0	-16.7
[+contact][superposition][top]	-100.0	-59.2	+16.7	-11.1	-55.6	+42.2	-16.7	-25.6	-30.0
[-contact][superposition][top]	-53.2	-31.1	+31.1	-26.7	+0.0	+0.0	-20.0	-32.2	-43.3
[+contact][subposition]	-24.4	-15.0	-60.0	-17.8	-40.0	+0.0	+18.9	+22.2	+38.3
[-contact][subposition]	+5.0	-16.7	+11.1	-30.0	+13.9	+0.0	-2.8	-11.1	-30.6

Table 17: Projective Breakdown Difference: Multimodal vs Text-only (big) (percentage points)

	Claude-3.5-Sonnet			GPT-4o			Qwen2.5-72B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[absolute][east]	+0.0	+0.0	+0.0	+0.0	-33.3	+0.0	+16.7	-8.3	+33.3
[absolute][north]	+0.0	+0.0	+0.0	+0.0	-16.7	-16.7	-16.7	+33.3	+0.0
[absolute][south]	+0.0	+0.0	+0.0	+0.0	+0.0	+8.3	+16.7	+66.7	+0.0
[absolute][west]	+0.0	+0.0	+0.0	+0.0	-33.3	+16.7	+33.3	+50.0	+33.3
[intrinsic][back]	+33.3	+0.0	+33.3	+0.0	+0.0	+16.7	+50.0	-16.7	+50.0
[intrinsic][front]	+0.0	+33.3	+33.3	+0.0	+0.0	+0.0	-33.3	+0.0	+0.0
[relative][back]	+0.0	+0.0	+16.7	-8.3	+0.0	+8.3	-16.7	+0.0	-33.3
[relative][front]	+8.3	+16.7	+16.7	+0.0	+0.0	+8.3	-33.3	-33.3	-66.7
[relative][left]	+0.0	+0.0	+0.0	+0.0	-50.0	-33.3	+0.0	+0.0	+0.0
[relative][right]	+0.0	+0.0	-33.3	-16.7	-33.3	+16.7	+0.0	+0.0	+0.0

Table 18: Projective Breakdown Difference: Multimodal vs Text-only (small) (percentage points)

	Claude-3.5-Haiku			GPT-4o-mini			Qwen2.5-7B		
	ES	EU	ZH	ES	EU	ZH	ES	EU	ZH
[absolute][east]	+0.0	+0.0	+0.0	+8.3	+25.0	+0.0	+16.7	+8.3	+33.3
[absolute][north]	+0.0	+0.0	+0.0	-8.3	+33.3	+8.3	+33.3	+58.3	+25.0
[absolute][south]	+0.0	+8.3	+0.0	+8.3	+16.7	+8.3	+58.3	+25.0	+0.0
[absolute][west]	+0.0	+0.0	+16.7	+16.7	+50.0	+16.7	+8.3	+16.7	+16.7
[intrinsic][back]	+0.0	+0.0	-50.0	-8.3	-16.7	+0.0	-8.3	-8.3	+0.0
[intrinsic][front]	+0.0	+0.0	-33.3	-8.3	-16.7	+0.0	+8.3	+8.3	+50.0
[relative][back]	+0.0	-33.3	-66.7	-8.3	-16.7	-16.7	+8.3	+0.0	-8.3
[relative][front]	-33.3	-50.0	-33.3	+0.0	+0.0	+8.3	+16.7	+33.3	+33.3
[relative][left]	+0.0	+0.0	+0.0	-16.7	+0.0	+0.0	+33.3	+50.0	+16.7
[relative][right]	+0.0	+0.0	+0.0	-33.3	+0.0	-16.7	+50.0	+33.3	+41.7