Grounding Agent Reasoning in Image Schemas: A Neurosymbolic Approach to Embodied Cognition

Blue Sky Ideas Track

François Olivier
CRIL CNRS & Artois University
Lens, France
olivier@cril.fr

Zied Bouraoui
CRIL CNRS & Artois University
Lens, France
bouraoui@cril.fr

ABSTRACT

Despite advances in embodied AI, agent reasoning systems still struggle to capture the fundamental conceptual structures that humans naturally use to understand and interact with their environment. To address this, we propose a novel framework that bridges embodied cognition theory and agent systems by leveraging a formal characterization of image schemas, which are defined as recurring patterns of sensorimotor experience that structure human cognition. By customizing LLMs to translate natural language descriptions into formal representations based on these sensorimotor patterns, we will be able to create a neurosymbolic system that grounds the agent's understanding in fundamental conceptual structures. We argue that such an approach enhances both efficiency and interpretability while enabling more intuitive human-agent interactions through shared embodied understanding.

KEYWORDS

Embodied AI; embodied cognition; neurosymbolic AI; image schemas; natural language understanding; agent reasoning; mental simulation.

ACM Reference Format:
François Olivier and Zied Bouraoui. 2025. Grounding Agent Reasoning in Image Schemas: A Neurosymbolic Approach to Embodied Cognition: Blue Sky Ideas Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 5 pages.

1 INTRODUCTION

By the end of the 20th century, the classical paradigm of cognitive science was fundamentally challenged as evidence mounted that our minds do not operate like isolated symbol processing computers, but rather are inextricably linked to our bodily experiences in the world. This became particularly evident in how we understand and use language, as Lakoff and Johnson's groundbreaking work in '*Metaphors We Live By*' [19] demonstrated that we comprehend abstract concepts (the target domain) by relying on our physical experiences as a source domain – we understand time through location ("the future is ahead of us"), importance through size ("this is a big deal"), and emotional states through spatial orientation ("I'm feeling down").

To bridge the gap between bodily experience and thought, Johnson [16] introduced *image schemas* - recurring patterns abstracted from our sensorimotor interactions - and showed their pervasive role in structuring human thought across both concrete and abstract domains. Over the years, the theory has received robust experimental confirmation across multiple studies [24, 31] and has proven fruitful even in non-linguistic domains such as mathematics [20]. A common example of image schema is OBJECT_INTO_CONTAINER which arises from our early physical experiences of putting objects into containers (e.g., cups and buckets), and later serves as a source domain to understand literal sentences like "Bill is in the house", more abstract ones such as "Berlin is in Germany" or "to be in love", and mathematical expressions such as "$2 \in \mathbb{N}$". More recent work has explored how these image schemas can be decomposed into even more basic constituents called *conceptual primitives* [24]. For instance, our comprehension of the concept of SUPPORT requires one to have the conceptual primitives of UP/DOWN and CONTACT.

Just as cognitive science had to move beyond purely computational models to explain human cognition and linguistic ability, there is an ongoing debate about whether AI systems need similar grounding to achieve genuine language understanding and commonsense reasoning [4, 36]. While some recent work suggests that Large Language Models (LLMs) can grasp physical concepts through text alone [28], there are reasons to be skeptical about whether this statistical learning can capture the full depth of human conceptual understanding [23, 25]. For instance, [29] highlights that LLMs employing in-context learning face significant challenges with tasks that require extensive specification, particularly those where even human annotators must carefully review a complex set of annotation guidelines to perform the task correctly. Using a simulation task, [38] also demonstrates fundamental conceptual limitations of statistical methods - limitations that persist regardless of the scale of the data. Equipping artificial agents with such conceptual embodied structures therefore becomes a crucial goal, as it would not only enable more intuitive and explainable human-agent interactions through shared embodied understanding, but also possibly represent, as suggested by [7], the necessary step to move AI into its next major paradigm beyond current multimodal systems.

However, the primary challenge in achieving such agents is to formalize these psychological theories and deeply embodied structures, and to intertwine the resulting symbolic language with neural recognition and metaphorical mapping techniques in a promising

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

way. In this work, we discuss the main challenges of such an endeavor and propose a promising approach that combines symbolic languages with neural architectures to create an integrated neurosymbolic framework. The main strengths of our approach compared to the existing work are the fully formal characterizations of conceptual structures, the use of existing symbolic solvers to reason with these characterizations, and the deep integration in a neural network in order to create a neurosymblic architecture.

The rest of the paper is organized as follows. Section 2 presents some related work from a symbolic and machine learning perspective. Section 3 discusses some of the main properties that the expected formalism should satisfy in order to, as shown in Section 4, effectively capture the different conceptual primitives that compose image-schematic structures. Section 5 presents how the formalism can be combined with neural networks in a meaningful way in order to enable fully embodied agents. Section 6 discusses the advantages gained in reasoning and natural language understanding with such embodied agents. Section 7 concludes the paper.

2 RELATED WORK

The formalization of image schemas is not a new endeavor - by the end of the twentieth century, Frank and Raubal [12] had already surveyed the existing formalisms. Among the formalisms that followed, notable approaches include bigraph-based representations [2], methods leveraging the WordNet lexical database [18], and approaches based on qualitative calculi [5, 13]. Qualitative calculi, which generally correspond to relation algebras [11], appeared particularly well-suited for this formalization task, as they abstract away from precise numerical measurements similarly to how human cognitive processing focuses on relative relationships. Significantly advancing the field, Hedblom's work made extensive use of the adequacy of qualitative calculi by combining the Region Connection Calculus, Qualitative Trajectory Calculus, Cardinal Directions, and Linear Temporal Logic in order to represent both spatial and temporal dimensions of image schemas [13]. Recently, Hedblom et al. proposed the Diagrammatic Image Schema Language (DISL) [14], a systematic diagrammatic representation language for image schemas that provides a structured visual framework.

Regarding the study of image schemas and embodied approaches in the machine learning community, the work of Wachowiak et al. explores how artificial agents capture implicit human intuitions underlying language [41] and introduces systematic methods for classifying natural language expressions into image schemas [39]. Recent advances in LLMs have also been leveraged to enhance performance in embodied learning tasks, particularly in Embodied Instruction Following [34], while standardized benchmarks to systematically evaluate these capabilities are emerging [21]. Finally, the framework developed in [38] closely aligns with our goal by approaching language understanding through mental simulation and metaphoric mappings.

3 FORMALISM PROPERTIES

As initiated in [24], image schemas can be decomposed into conceptual primitives. For instance, GOING_IN requires at least the notions of OBJECT, CONTAINER and PATH. To present our approach, we use the more recent classification from [14] reproduced in Table 1. As can be seen, some conceptual primitives are only spatial or spatiotemporal, whereas others are force dynamic primitives, which correspond to embodied feelings that cannot be represented in a spatiotemporal way (e.g., UMPH corresponds to the application of a *force*).

Table 1: Classification of conceptual primitives from [14].

	entity	relational	attributive
spatial	OBJECT CONTAINER PATH REGION DOWN (/UP)	LOCATION START_PATH END_PATH CONTACT CONTAINED SMALLER(/LARGER) PART_OF	OPEN CLOSED EMPTY OCCUPIED FULL
spatio-temporal		PERMANENCE	MOTION AT_REST ANIMATE_MOTION INANIMATE_MOTION
force dynamic		LINK	active-UMPH passive-UMPH

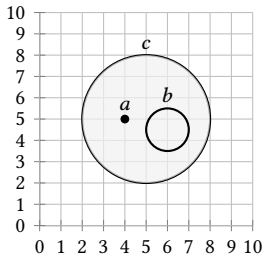
Property 1. Since image schemas can structure an infinite variety of physical configurations and scenarios, any formalism for representing them must be able to encode relationships qualitatively (e.g., being 'inside' something without knowing exact locations or shapes) [22]. This requirement has been widely recognized in previous formalization attempts.
Property 2. Objects of different types can be involved in an image schema, such as points for atomic OBJECTS or lines for PATHS. Additionally, an ordering over types may be useful for defining certain entities (e.g., a CONTAINER can be a circle, a square, etc). Therefore, the formalism should be order-sorted and support the definition of typed relations.
Property 3. Since image schemas can be understood as small narratives, the formalism should support the expression of time and the evolution of configurations over time.
Property 4. The formalism should support quantification to express general rules and assert the (non-)existence of objects (e.g., for the primitive EMPTY), as well as logical connectives to effectively express logical constraints.
Property 5. Finally, the formalism should support the use of a default operator to model default behaviors, such as gravity or the law of inertia (i.e., things remain the same unless an action caused them to change) [33]. Importantly, the inclusion of a default operator makes the formalism non-monotonic.

4 FORMALIZING IMAGE SCHEMAS

A promising candidate that meets these requirements, or allows for additional extensions to fulfill them, is to implement the Declarative Spatial Reasoning framework (DSR) [6] within the non-monotonic Quantified Equilibrium Logic with evaluable functions [8, 9]. Quantified equilibrium logic maintains the syntax of first-order logic while semantically interpreting negation as default negation (i.e., *negation as failure* [10]).

Evaluable functions enable the embedding of the DSR framework [6] since the latter fundamentally relies on parametric functions for

representing objects (see Figure 1, top right), and defines qualitative relations between objects through polynomial constraints on these parameters (bottom right) [30]. Contrary to the common practice of using the algebraic qualitative calculi mentioned in Section 2, the DSR framework allows the combination of heterogeneous objects and does not impose any conditions on the set of relations defined.



$x_a = 4 \quad x_b = 6 \quad x_c = 5$
$y_a = 5 \quad y_b = 4.5 \quad y_c = 5$
$r_b = 1 \quad r_c = 3$

$inside(a, c)$
$\leftrightarrow \ (x_a - x_c)^2 + (y_a - y_c)^2 < r_c^2$
$\leftrightarrow \ (-1)^2 + 0^2 < 3^2$
$\leftrightarrow \ 1 + 0 < 9$
$\leftrightarrow$ TRUE

Figure 1: In the DSR framework, the parameters of objects are used to define spatial relations between these objects.

Regarding temporal modeling, a first-order extension of Temporal Equilibrium Logic has been proposed in [1]. For the purpose of forthcoming examples, we consider the following temporal operators:

$\circ$	*next*	$\Box$	*always*	$\Diamond$	*eventually afterward*
$\mathbb{U}$	*until*	$\mathbb{F}$	*final*	$\blacklozenge$	*eventually before*

Finally, many-sorted formalisms have been developed for approaches closely related to equilibrium logic [3], while formal treatments of order-sorted logic can be found in [17].

In what follows, we explain how conceptual primitives are handled in our formalism and provide examples of some of their combinations. Our treatment shares similarities with [14] as we apply our formalism on their classification presented in Figure 1.

Entities correspond to constants in the logic. The entity OBJECT simply corresponds to a point. The entity CONTAINER corresponds to any geometric objects that can be used in a relation of 'containment', such as *inside*, *properPart*, etc. Ordered sorts enable us to define this entity as a superclass, that is, any circle, rectangle, etc. is a CONTAINER entity. The PATH entity is modeled as a line with a starting and ending point. For instance, the image schema of SOURCE_PATH_GOAL which underlies our understanding of processes composed of consecutive steps (e.g., the progression of degrees in a student's academic journey, advancing through the bases in baseball, etc), can be represented by a series of location such as $l_1 \wedge \Diamond(l_i \wedge \Diamond(\ldots \wedge \Diamond l_n))$ where l_1 and l_n respectively stand for the START_PATH and END_PATH as specific locations, and each l_i represents an intermediate location. Forward movement is obtained by constraining the actual LOCATION with the previous ones by means of the $\blacklozenge$ operator. The entity REGION is modeled either by means of a distance function Δ or as a CONTAINER entity similar to the one above. Finally, the more abstract notion DOWN is either modeled as a line placed at the bottom of the scene, or is directly encoded within displacement actions. For instance, gravity can be modeled as $\Box(\forall x(\neg\exists y\ on(x, y) \rightarrow moveDown(x)))$, where x and y are any entities in the domain. Note the use of the default negation in the latter formula.

The relational primitives mainly correspond to binary (or higher-arity) relations. LOCATION can be expressed by means of positional or topological relations (e.g., *on*, *closeTo*, *inside*,...). START_PATH and END_PATH, as mentioned above, can be defined as points or geometric regions that delimitate a PATH entity. CONTACT, CONTAINED and PART_OF simply correspond to topological relations defined in the DSR framework, and similarly for SMALLER/LARGER as size relations. LINK can either be defined by means of a distance Δ that cannot exceed a certain threshold or as an actual segment that touches the objects linked. Finally, PERMANENCE can be expressed by means of a default negation, encoding the idea that if we cannot prove that a parametric function of an entity has changed, we conserve its value for the actual state.

Although attributive conceptual primitives first seem to correspond to unary predicates applied to entities, they will usually require complex formulas. For instance, EMPTY corresponds to a formula where we state that, for a CONTAINER, no entity is inside it. The force dynamic conceptual primitives active-UMPH and passive-UMPH are modeled with default negation. Basically, *unless* a contrary force is applied to an object, the latter is subject to an action at each state (possibly until a certain goal is achieved, using the $\mathbb{U}$ operator). Such a concept of force occurs in the characterization of gravity as presented above. Finally, MOTION, AT_REST and the (IN)ANIMATE primitives correspond to action predicates that modify/apply to the location of entities along states.

When combined, these conceptual primitives give rise to image schemas, each of which formally corresponds to a small theory Γ encapsulating its essential structure and enabling reasoning. Such fully formal characterizations may also contribute to clarifying and standardizing the definitions of image schemas within the field. From a model-theoretic perspective, each model of a theory Γ represents a possible instantiation of the structure under consideration, which aligns with the idea of a *schema* used as a template for generating infinitely many concrete *images* and scenarios.

5 NATURAL LANGUAGE PARSING VIA NEURAL IMAGE SCHEMA RECOGNITION

Having established a formal foundation for representing image schemas in the previous sections, we now turn to the challenge of automatically extracting these representations from natural language. Our goal is to develop a system that can take ordinary sentences and parse them into the non-monotonic quantified formalism presented above.

This task presents unique challenges compared to traditional semantic parsing. While conventional semantic parsers typically map language to classical logical systems [27, 42], our system must capture the embodied, spatiotemporal meaning inherent in language. For instance, when processing a sentence like "The monk climbs up the mountain" from the riddle presented in [14], the system must recognize not only the entities involved but also the complex interplay of image schemas such as SOURCE_PATH_GOAL and CONTACT, along with their temporal evolution.

To address this challenge, we propose leveraging recent advances in LLMs and neural architectures. Modern transformer-based models have demonstrated remarkable capabilities in understanding linguistic structure and generating complex outputs. We can build on

their strong language understanding and generation capabilities to translate natural language descriptions into our image schema formalism. A critical challenge in developing such a system is collecting sufficient high-quality mapping data between natural language sentences and their image schema representations. Fortunately, several existing resources can be leveraged:

- Structured databases from [39, 41] provide ready-to-use examples for training, validation and testing.
- Psychological experiments in the literature, for instance [31], offer empirically-grounded data on image schema evocation in human participants.
- LLMs can be strategically prompted to generate candidate image schema annotations for natural language sentences.
- Expert linguists and cognitive scientists can provide gold-standard annotations mapping linguistic constituents to schema roles and identifying active image schemas.

Regarding the formalization of the image schema representations, we propose a two-stage approach. First, we can leverage LLMs' strong reasoning capabilities to generate initial formal characterizations of identified schemas. Our formalism's adherence to first-order logic with temporal operators makes it particularly amenable to automated generation, as these logical structures are well-represented in LLMs' training data. Second, we can fine-tune a specialized translation model on our collected dataset of natural language sentences paired with their formal representations. This model would learn to directly map input text to well-formed expressions in our formalism. To ensure quality and consistency, we propose an iterative development process where model outputs are validated against expert annotations and refined based on error analysis.

Finally, evaluation of such a system requires going beyond simple accuracy metrics. While an exact match with gold-standard annotations provides one measure of success, we must also consider partial matching metrics that assess the system's ability to identify correct image schemas, assign appropriate roles, and maintain proper temporal structures. Additionally, the system's performance should be evaluated on downstream tasks that require genuine understanding of spatial relationships, motion events and force components.

6 NATURAL LANGUAGE UNDERSTANDING, REASONING AND ANALOGIES

The proposed model could serve as a crucial component in embodied AI systems, helping to bridge the gap between language understanding and physical interaction with the world. Image schemas, being grounded in bodily experience and spatial understanding, provide a natural intermediate representation between linguistic input and physical action. By capturing these embodied cognitive patterns in our formal notation, we enable AI systems to process language in a way that connects directly to spatial reasoning and motor planning. This creates a tighter coupling between natural language understanding and real-world interaction - rather than treating language as purely symbolic manipulations, the system can ground linguistic meanings in the same kind of spatial and motor primitives that humans use.

Reasoning would also be enhanced through closer alignment with human cognitive processes. By operating over the same kind of image-schematic representations that humans use, AI systems could better model and predict human understanding and misunderstanding. For example, an agent could identify when a human might struggle to grasp a concept by analyzing which image schemas are involved and whether they map naturally to familiar embodied experiences. Moreover, these agents could reason in ways that parallel human inference patterns. As Shimojima demonstrates in his analysis of diagrammatic reasoning [35], certain conclusions emerge naturally (or come "for free") from visual representations without explicit logical rules. Image schemas leverage this same principle, as the spatial constraints between entities capture the logical constraints in the target domain [26]. To realize these inferences in a computational framework, we can harness answer set programming via Clingo, as partly explored in work on related areas [32, 37, 40]. Clingo's ability to handle non-monotonic reasoning and incorporate custom theories such as the ones described to characterize image schemas makes it particularly suitable for implementing our formalism.

Finally, our formalism might turn out particularly useful in capturing analogical relationships, where a conceptual structure can be mapped to multiple target domains. Consider the classic analogy between the solar and (Rutherford–Bohr) atomic systems, exemplified in the sentences "electrons circle the nucleus" and "planets circle the sun" [15]. Both can be formalized using the same image-schematic structure where a distance $\Delta(x, y)$, between x as electrons/planets and y as the nucleus/sun, is constrained within certain bounds, and $\theta(x, y) < \circ\theta(x, y)$ ensures that the angular position of x relative to y continuously increases, capturing the circular orbital motion. The structural similarity revealed in these formalizations explains the cognitive power of the analogy - both scenarios share the same underlying image-schematic structure.

7 CONCLUSION AND CHALLENGES

This paper has presented a comprehensive approach to bridging the gap between natural language understanding and embodied cognition. Building on cognitive theories of image schemas and recent advances in large language models, we have outlined a formalism that captures the essential spatial, temporal and force dynamic primitives underlying human conceptual understanding. While the complete formalization remains to be fully developed, we have demonstrated how the key components can be systematically combined to represent complex conceptual structures. The integration of this formalism with modern transformer architectures opens new possibilities for grounding language understanding in embodied experiences. By capturing image schemas in a computationally tractable form, we enable systems to process language in ways that mirror human cognitive patterns. The resulting representations support natural forms of reasoning and analogical mapping, as demonstrated through examples ranging from basic containment relationships to complex analogies. Our work provides a foundation for developing AI systems that can understand and reason with language in more human-like ways.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (ANR) under grant ANR-22-CE23-0002 ERIANA.

REFERENCES

[1] Felicidad Aguado, Pedro Cabalar, Gilberto Pérez, Concepción Vidal, and Martin Dieguez. 2017. Temporal logic programs with variables. *Theory and Practice of Logic Programming* 17, 2 (2017), 226–243.

[2] Robert St Amant, Clayton T Morrison, Yu-Han Chang, Paul R Cohen, and Carole Beal. 2006. An image schema language. In *Submitted to The 7th International Conference on Cognitive Modelling (ICCM 2006)*. Citeseer.

[3] Evgenii Balai, Michael Gelfond, and Yuanlin Zhang. 2013. Towards Answer Set Programming with Sorts. In *Proceedings of the 12th International Conference on Logic Programming and Nonmonotonic Reasoning - Volume 8148* (Corunna, Spain) *(LPNMR 2013)*. Springer-Verlag, Berlin, Heidelberg, 135–147. https://doi.org/10.1007/978-3-642-40564-8_14

[4] Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 5185–5198.

[5] Brandon Bennett and Claudia Cialone. 2014. Corpus Guided Sense Cluster Analysis: a methodology for ontology development (with examples from the spatial domain). In *Formal Ontology in Information Systems*. IOS Press, 213–226.

[6] Mehul Bhatt, Jae Hee Lee, and Carl Schultz. 2011. CLP(QS): A Declarative Spatial Reasoning Framework. In *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings (Lecture Notes in Computer Science, Vol. 6899)*, Max J. Egenhofer, Nicholas A. Giudice, Reinhard Moratz, and Michael F. Worboys (Eds.). Springer, 210–230. https://doi.org/10.1007/978-3-642-23196-4_12

[7] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

[8] Pedro Cabalar. 2011. Functional answer set programming. *Theory and Practice of Logic Programming* 11, 2-3 (2011), 203–233.

[9] Pedro Cabalar, Jorge Fandinno, Luis Farinas Del Cerro, and David Pearce. 2018. Functional ASP with intensional sets: Application to Gelfond-Zhang aggregates. *Theory and Practice of Logic Programming* 18, 3-4 (2018), 390–405.

[10] Keith L Clark. 1977. Negation as failure. In *Logic and data bases*. Springer, 293–322.

[11] Frank Dylla, Jae Hee Lee, Till Mossakowski, Thomas Schneider, André Van Delden, Jasper Van De Ven, and Diedrich Wolter. 2017. A survey of qualitative spatial and temporal calculi: Algebraic and computational properties. *ACM Computing Surveys (CSUR)* 50, 1 (2017), 1–39.

[12] Andrew U Frank and Martin Raubal. 1999. Formal specification of image schemata–a step towards interoperability in geographic information systems. *Spatial Cognition and Computation* 1 (1999), 67–101.

[13] Maria M Hedblom. 2020. *Image schemas and concept invention: cognitive, logical, and linguistic investigations*. Springer Nature.

[14] Maria M Hedblom, Fabian Neuhaus, and Till Mossakowski. 2024. The Diagrammatic Image Schema Language (DISL). *Spatial Cognition & Computation* (2024), 1–38.

[15] Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding. *arXiv preprint arXiv:2310.12874* (2023).

[16] Mark Johnson. 1987. *The body in the mind: The bodily basis of reason and imagination*. Chicago: University of Chicago Press.

[17] Ken Kaneiwa. 2004. Order-sorted logic programming with predicate hierarchy. *Artificial Intelligence* 158, 2 (2004), 155–188.

[18] Werner Kuhn. 2007. An image-schematic account of spatial categories. In *International Conference on Spatial Information Theory*. Springer, 152–168.

[19] George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

[20] George Lakoff and Rafael Núñez. 2000. *Where mathematics comes from*. Vol. 6. New York: Basic Books.

[21] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. 2025. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* 37 (2025), 100428–100534.

[22] Gérard Ligozat. 2013. *Qualitative spatial and temporal reasoning*. John Wiley & Sons.

[23] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences* 28, 6 (2024), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

[24] Jean M Mandler and Cristóbal Pagán Cánovas. 2014. On defining image schemas. *Language and Cognition* 6, 4 (2014), 510–532.

[25] R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638* (2023).

[26] François Olivier. 2022. *Spatial relations in reasoning : a computational model*. Theses. Université Paris sciences et lettres. https://theses.hal.science/tel-03984759

[27] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295* (2023).

[28] Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

[29] Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does In-context Learning Fall Short and Why? A Study on Specification-Heavy Tasks. *CoRR* abs/2311.08993 (2023). https://doi.org/10.48550/ARXIV.2311.08993 arXiv:2311.08993

[30] Franco P Preparata and Michael I Shamos. 2012. *Computational geometry: an introduction*. Springer Science & Business Media.

[31] Daniel C Richardson, Michael J Spivey, Shimon Edelman, and Adam J Naples. 2001. " Language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 23.

[32] Carl Schultz, Mehul Bhatt, Jakob Suchan, and Przemysław Andrzej Wałęga. 2018. Answer Set Programming Modulo 'Space-Time'. In *International Joint Conference on Rules and Reasoning*. Springer, 318–326.

[33] Murray Shanahan. 1997. *Solving the frame problem: a mathematical investigation of the common sense law of inertia*. MIT press.

[34] Haochen Shi, Zhiyuan Sun, Xingdi Yuan, Marc-Alexandre Côté, and Bang Liu. 2024. OPEx: A Component-Wise Analysis of LLM-Centric Agents in Embodied Instruction Following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 622–636. https://doi.org/10.18653/v1/2024.acl-long.37

[35] Atsushi Shimojima. 2015. *Semantic Properties of Diagrams and Their Cognitive Potentials*. CSLI Publications, Stanford, California.

[36] Anders Søgaard. 2023. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines* 33, 1 (2023), 33–54.

[37] Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. 2015. Talking about the Moving Image: A Declarative Model for Image Schema Based Embodied Perception Grounding and Language Generation. *CoRR* abs/1508.03276 (2015). arXiv:1508.03276 http://arxiv.org/abs/1508.03276

[38] Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. 2020. Language (Re)modelling: Towards Embodied Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6268–6281. https://doi.org/10.18653/v1/2020.acl-main.559

[39] Lennart Wachowiak and Dagmar Gromann. 2022. Systematic Analysis of Image Schemas in Natural Language through Explainable Multilingual Neural Language Processing. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 5571–5581. https://aclanthology.org/2022.coling-1.493/

[40] Przemysław Andrzej Wałęga, Carl Schultz, and Mehul Bhatt. 2017. Non-monotonic spatial reasoning with answer set programming modulo theories. *Theory and Practice of Logic Programming* 17, 2 (2017), 205–225. https://doi.org/10.1017/S1471068416000193

[41] Philipp Wicke and Lennart Wachowiak. 2024. Exploring Spatial Schema Intuitions in Large Language and Vision Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6102–6117. https://doi.org/10.18653/v1/2024.findings-acl.365

[42] Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024. Harnessing the Power of Large Language Models for Natural Language to First-Order Logic Translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6942–6959. https://doi.org/10.18653/v1/2024.acl-long.375