“Let’s collide with the approaching car head-on” Introducing Synthes-IS: extending the Image Schema catalogue with synthetic-data.

Guendalina Righetti*,†[1], Stefano De Giorgis*,†[2]

[1]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Blindernveien 31 Georg Morgenstiernes hus 0313 Oslo

[2]Department of Artificial Intelligence, Vrije Universiteit Amsterdam, The Netherlands

Abstract

Image Schema (IS) research has been significantly constrained by the limited availability of annotated linguistic data, which has slowed empirical progress and evaluation in the field. One of the few existing resources — the Image Schema Catalogue — offers a collection of metaphorical sentences annotated with single image schemas. In prior work, we extended this catalogue in two key ways: (1) by allowing for the annotation of multiple image schemas per sentence, thereby reflecting the often-complex interplay of schemas in language use; and (2) by augmenting the dataset with new, non-metaphorical sentences grounded in practical scenarios, all enriched with image schematic content. Both extensions were generated and annotated with the assistance of a large language model (LLM), opening new possibilities for scalable IS research. However, that prior work lacked a rigorous evaluation of the quality of the LLM-generated sentences and their corresponding IS annotations. This paper addresses that limitation through a systematic expert-based evaluation. Independent domain experts were tasked with assessing both the relevance and accuracy of the image schemas assigned to each sentence, as well as the plausibility and linguistic quality of the LLM-generated content.

Keywords

image schemas, synthetic data, generative AI

1. Introduction

Image Schemas (IS) are foundational conceptual structures that are deemed central in explaining embodied cognition. IS encapsulate patterns of sensorimotor experience, and as such serve as the building blocks for higher-level cognitive processes such as commonsense reasoning and language understanding (see, e.g., Mandler and Pagàn Cànovas [1] and Talmy [2]). Specifically, they have been proven useful in analysing phenomena such as conceptual blending and metaphor understanding, especially in the context of computational approaches [3, 4, 5, 6]: they can be seen as sensorimotor packages enabling semantic mapping across different domains. Moreover, by enabling the encoding of recurrent patterns of sensorimotor experience, image schemas have played a growing role in cognitive robotics, particularly for supporting intuitive physics and anticipatory reasoning in autonomous agents [7, 8, 9, 10]. These schematic structures provide a bridge between embodied experience and formal knowledge representation, allowing robots to simulate and reason about spatial and causal relationships in human-like ways. To integrate such capabilities, researchers have developed formal representations of image schemas — ranging from logic-based approaches like Image Schema Logic (ISL) [11] to ontology-based frameworks such as ISL2OWL [12, 13] — to support semantic interpretation within robotic systems.

Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), September 8-9, 2025, Catania, Italy

*Corresponding authors.

†These authors contributed equally to this work.

guendalina.righetti@ifikk.uio.no (G. Righetti*,†); s.degiorgis@vu.nl (S. De Giorgis*,†)

0000-0002-4027-5434 (G. Righetti*,†); 0000-0003-4133-3445 (S. De Giorgis*,†)

For many of such works the Image Schema Catalogue (ISC) [14, 15][1] is the primary resource for empirical IS examples. More specifically, the dataset compiles linguistic examples drawn from a variety of established sources, including MetaNet [16], the works of Lakoff and Johnson [17, 18], and Dodge and Lakoff [19]. Most of the sentences used as examples involve a metaphorical use of image schema. For each sentence, the dataset collects the underlying conceptual metaphor along with its associated source and target domains. Additionally, it specifies the embodied grounding of the expression—referred to as the "sensorimotor" source domain—which may correspond to a basic spatial primitive or a more complex image schema. A dedicated column also indicates the specific image schema evoked by each sentence.

Notwithstanding its usefulness, the Catalogue presents certain limitations. First, each sentence in the catalogue is annotated with only one image schema, a simplification that fails to capture the layered and often co-occurring nature of image schematic structures in natural language. Second, the ISC offers limited data, and the examples it contains are predominantly metaphorical. While this suits metaphor theory and conceptual blending studies, it restricts applicability in areas that benefit from more concrete, literal examples—such as robotics, human-computer interaction, and embodied AI. Furthermore, annotations in metaphorical contexts are more susceptible to subjective interpretation, making them potentially less reliable. In contrast, literal examples tend to allow for more objective annotation and can serve as a stable reference point for validating the use of image schemas in corresponding metaphorical expressions.

In our previous work [20], we addressed these two issues — limited annotation and data scarcity — by developing a pipeline for the automated extension and enrichment of the ISC using large language models (LLMs). The process included four key stages: selecting a competent LLM (Claude 3.5 Sonnet), performing multi-label IS classification of existing catalogue sentences, generating literal counterparts to metaphorical expressions, and quantitatively evaluating annotation accuracy. The LLM was used both to annotate existing sentences with multiple image schemas and to produce literal reformulations of the metaphorical sentences that preserved the original image schematic structure. The new database, synthes-IS, is available at: https://github.com/StenDoipanni/ISAAC/tree/main/ISD9. While we evaluated the classification task quantitatively, the generated literal sentences were not yet subjected to qualitative review.

The present study addresses this gap. We perform a structured evaluation of the literal sentences generated by the LLM in our previous work. Specifically, we assess whether these sentences (i) constitute plausible and grounded literal reformulations of their metaphorical counterparts, and (ii) remain consistent with the original image schema annotations. Our evaluation is conducted by IS experts using a 7-point Likert scale (7 = completely appropriate, 1 = completely inappropriate) to allow annotators express their judgement with enough granularity.

The analysis offers a critical perspective on the reliability of such automated outputs for extending schema-based resources. In the sections that follow, we outline the evaluation methodology, present the results, and discuss their implications for the development of IS-based applications and resources.

2. Methodology

To assess the quality of the literal sentences generated from the original metaphorical expressions, we conducted a targeted qualitative evaluation. While our previous work [20] focused on the quantitative accuracy of image schema annotations, this new evaluation investigates the appropriateness and coherence of the literal sentence generation process.

Before the evaluation, we performed a preprocessing step to exclude all the sentences in German from the catalogue, to avoid possible biases or ambiguity due to translation.

We then randomly sampled 30% of the generated literal sentences from the full dataset (evenly distributed across image schemas). The evaluation was carried out independently by the two authors

[1]For this work we consider the refinement of the original Image Schema Catalogue as in github.com/dgromann/ImageSchemaRepository

of the study, both of whom are domain experts in image schema theory. Each evaluator assessed a different subset of the data to maximise coverage. Two dimensions were used for evaluation:

1. Metaphor-to-Literal Appropriateness (M2L): This dimension examined how effectively the generated sentence served as a concrete and plausible literal reformulation of the original metaphorical expression. We aimed to determine whether the model successfully grounded the metaphor in a tangible scenario, while preserving the core conceptual mapping. For instance, given the metaphorical sentence "Our agenda is packed with events," we would consider "The bag is packed with clothes" a strong literal counterpart.
2. Original Annotation Appropriateness (OA): This dimension evaluated whether the generated literal sentence remained consistent with the original image schemas annotated for the metaphorical sentence.[2] In other words, we assessed whether the literal reformulation reflected the same underlying image schematic structure. In the above example, for instance, we would evaluate the literal sentence to preserve the CONTAINMENT image schema.

Each sentence was rated on a 7-point Likert scale (1 = completely inappropriate, 7 = highly appropriate) for both dimensions.

3. Analysis and Discussion

Out of a total of 2055 English sentences in the Catalogue, we manually evaluated a representative sample of 741 generated literal sentences, selected to ensure balanced coverage across different image schemas. As described above, the evaluation focused on two main criteria: Metaphor-to-Literal Appropriateness (M2L) and Original Annotation Appropriateness (OA), each rated on a 7-point Likert scale.

Overall, the results indicate strong performance by the LLM in generating coherent, grounded literal counterparts to metaphorical expressions. The mean score for M2L was 6.38, while the mean score for OA was 5.97, suggesting that the literal sentences were not only contextually plausible but also generally consistent with the original image schema annotations.

A breakdown of mean scores across the different image schemas is shown below:

Image Schema	Metaphor-to-Literal Appropriateness	Original Annotation Appropriateness
Center-Periphery	6.66	6.38
Contact	7	7
Containment	6.20	5.92
Covering	5.57	5.57
Force	6.29	5.23
Link	6	6
Object	6.43	6.08
Part-Whole	6.76	6.57
Scale	6.72	6.84
Source-Path-Goal	6.24	5.46
Splitting	6.10	6.70
Substance	6.71	6.57
Support	7.00	7.00
Verticality	6.48	6.32
OVERALL	6.38	5.97

These results highlight that certain schemas — such as Contact, Part-Whole, Support, Scale and Substance — achieved particularly high scores in both dimensions, indicating that the LLM was especially effective at generating literal expressions for these conceptual structures. On the other hand, slightly

[2]Please note that the Original Annotations come from the initial construction of the IS Catalogue and were assigned manually by humans.

lower scores for schemas like Covering, Force, and Source-Path-Goal suggest areas where interpretation and grounding remain more challenging.

Overall, this evaluation suggests that LLMs, when appropriately prompted, are capable of producing high-quality, image-schematically faithful literal sentence counterparts to metaphorical expressions.

We hypothesised that the cases of low performance could correlate with instances in which the LLM selected a different image schema than the one annotated by the human expert as most appropriate. To investigate this, we further analysed the data by isolating entries that received below-average evaluation scores. Out of the 741 evaluated sentences, 212 were rated below the mean in at least one of the two assessed dimensions. Then we compared these cases with results from our previous study, which quantitatively assessed the accuracy of image schema (IS) annotations produced by the LLM. Notably, 63% of these lower-rated entries (153 out of 212) also corresponded to incorrect IS annotations—specifically, cases where the LLM assigned a different image schema than the one identified by human annotators. This overlap suggests a strong correlation between lower evaluation scores and annotation mismatches.

Another question we explored was whether the cases in which the LLM assigned a different image schema than the one identified by human annotators—and which received below-average evaluation scores—were genuine errors, or if, in some instances, the LLM's interpretation might be justifiable. In other words, to what extent do we actually agree with the LLM over the original human annotation? To address this, we focused on entries where our two evaluation dimensions—Metaphor-to-Literal Appropriateness (M2L) and Original Annotation Appropriateness (OA)—showed a mismatch. Specifically, we examined cases where the M2L score was relatively positive (≥ 4), but the OA score was low (≤ 3), suggesting that the literal sentence was coherent and plausible, even if it did not align with the original annotation. Out of the 212 entries that received below-average scores, 153 involved disagreements between the LLM and human annotation. Of these, 60 cases (roughly 1/3) displayed the pattern described above. This suggests that in a significant portion of instances, we may side with the LLM's interpretation over the original annotation, indicating possible ambiguity or subjectivity in the human-labeled data.

Some examples are reported in the following table. Full data is available at: https://github.com/StenDoipanni/ISAAC/blob/main/ISD9/interesting_cases.csv

Metaphorical Sentence	Literal Sentence	Original Annotation	LLM Annotation
You can't get a single joke out of him.	You can't remove a single object from this container.	SUBSTANCE	CONTAINMENT
I couldn't do much sprinting until the end.	I couldn't run fast until I reached the finish line.	SUBSTANCE	SOURCE_PATH_GOAL
Let's spread the conference over two weeks.	Let's distribute the papers over two tables.	SOURCE_PATH_GOAL	SPLITTING
The test was at the forefront of my attention.	The book was at the front of the shelf.	SOURCE_PATH_GOAL	CENTER-PERIPHERY
It's difficult to put my ideas into words.	It's difficult to put water into a bottle.	OBJECT	CONTAINMENT
FDR's leadership brought the country out of the depression.	The guide led the hikers out of the cave.	FORCE	SOURCE_PATH_GOAL

4. Conclusions and Future Work

This study complements our prior work on enriching the Image Schema Catalogue (ISC) through large language models (LLMs), by introducing a structured qualitative evaluation of the literal sentences generated by Claude 3.5 Sonnet. Building on a pipeline that performed multi-label IS annotation and literal reformulation of metaphorical expressions, we assessed the resulting data with expert evaluations across two dimensions: Metaphor-to-Literal Appropriateness (M2L) and Original Annotation Appropriateness (OA).

Out of 741 evaluated entries, both dimensions received overall high scores (M2L = 6.38; OA = 5.97), indicating that LLM-generated sentences generally preserve both semantic plausibility and image schematic alignment.

To understand the nature of lower-scoring outputs, we isolated 212 entries rated below average. Of these, 63% (153 cases) also featured a mismatch between the LLM-assigned and human-assigned image schema, suggesting a strong correlation between annotation disagreement and reduced output quality. However, a finer-grained analysis revealed that in approximately one-third of those cases (53/153), annotators rated the literal sentence positively despite disagreeing with the original annotation. This indicates that some mismatches may stem from ambiguity or subjectivity in the human annotation itself rather than errors by the LLM.

These results suggest that LLMs can generate plausible literal equivalents of metaphorical language while preserving image schematic structure. Moreover, cases of disagreement between LLM and human annotations offer valuable insights for refining schema classification frameworks. Future work will focus on expanding expert evaluation, investigating inter-annotator agreement, and exploring how LLM-generated alternatives can contribute to iterative improvement of IS resources.

Declaration on Generative AI

This work made use of generative AI to conduct analysis as detailed in the paper, and for general language refining.

References

[1] J. M. Mandler, C. Pagán Cánovas, On defining image schemas, Language and Cognition (2014) 1–23. doi:10.1017/langcog.2014.14.

[2] L. Talmy, The fundamental system of spatial schemas in language, in: B. Hampe, J. E. Grady (Eds.), From perception to meaning: Image schemas in cognitive linguistics, volume 29 of *Cognitive Linguistics Research*, Walter de Gruyter, 2005, pp. 199–234.

[3] M. M. Hedblom, O. Kutz, F. Neuhaus, Image schemas in computational conceptual blending, Cognitive Systems Research 39 (2016) 42–57.

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

[5] G. Righetti, O. Kutz, The moving apple: An image-schematic investigation into the leuven concept database, in: Proceedings of The Seventh Image Schema Day co-located with The 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023), Rhodes, Greece, September 2nd, 2023, CEUR-WS, 2023.

[6] G. Righetti, D. Porello, N. Troquard, O. Kutz, M. M. Hedblom, P. Galliani, Asymmetric hybrids: Dialogues for computational concept combination, in: Formal Ontology in Information Systems, IOS Press, 2021, pp. 81–96.

[7] M. M. Hedblom, M. Pomarlan, R. Porzel, R. Malaka, M. Beetz, Dynamic action selection using image schema-based reasoning for robots (2021).

[8] M. Pomarlan, G. Righetti, J. A. Bateman, It is what it tends to do: Defining qualitative parameter regions by their effects on physical behavior, in: M. M. Hedblom, O. Kutz (Eds.), Proceedings of

the Sixth Image Schema Day, Jönköping, Sweden, March 24-25th, 2022, volume 3140 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3140/paper7.pdf.

[9] M. Pomarlan, S. De Giorgis, R. Ringe, M. M. Hedblom, N. Tsiogkas, Hanging around: Cognitive inspired reasoning for reactive robotics, in: Formal Ontologies for Information Systems (FOIS) 2024, 2024.

[10] F. Olivier, Z. Bouraoui, Grounding Agent Reasoning in Image Schemas: A Neurosymbolic Approach to Embodied Cognition, arXiv preprint arXiv:2503.24110 (2025).

[11] M. M. Hedblom, O. Kutz, T. Mossakowski, F. Neuhaus, Between contact and support: Introducing a logic for image schemas and directed movement, in: Conference of the Italian Association for Artificial Intelligence, Springer, 2017, pp. 256–268.

[12] S. De Giorgis, A. Gangemi, D. Gromann, ImageSchemaNet: A framester graph for embodied commonsense knowledge, Semantic Web 15 (2024) 1417–1441.

[13] S. De Giorgis, A. Gangemi, D. Gromann, Introducing ISAAC: The Image Schema Abstraction And Cognition Modular Ontology, in: Proceedings of the 8th Joint Ontology Workshops, 2022.

[14] J. Hurtienne, J. H. Israel, Image schemas and their metaphorical extensions: intuitive patterns for tangible interaction, in: Proceedings of the 1st international conference on Tangible and embedded interaction, 2007, pp. 127–134.

[15] J. Hurtienne, S. Huber, C. Baur, Supporting user interface design with image schemas: The iscat database as a research tool., in: ISD, 2022.

[16] E. K. Dodge, J. Hong, E. Stickles, Metanet: Deep semantic automatic metaphor analysis, in: Proceedings of the Third Workshop on Metaphor in NLP, 2015, pp. 40–49.

[17] M. Johnson, The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason, The University of Chicago Press, Chicago and London, 1987.

[18] G. Lakoff, M. Johnson, et al., Philosophy in the flesh: The embodied mind and its challenge to western thought, volume 640, Basic books New York, 1999.

[19] E. Dodge, G. Lakoff, Image schemas: From linguistic analysis to neural grounding, From perception to meaning: Image schemas in cognitive linguistics (2005) 57–91.

[20] S. De Giorgis, G. Righetti, "The Time for Action has Arrived": Extending the IS Catalogue Leveraging Large Language Models, in: M. M. Hedblom, O. Kutz (Eds.), Proceedings of The Eighth Image Schema Day co-located with The 23rd International Conference of the Italian Association for Artificial Intelligence(AI*IA 2024), Bozen-Bolzano, Italy, November 27-28th, 2024, volume 3888 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3888/Paper_7.pdf.