Testing Spatial Intuitions of Humans and Large Language and Multimodal Models in Analogies

Ivo Bueno
Technical University of Munich
ivo.bueno@tum.de

Anna Bavaresco
University of Amsterdam
a.bavaresco@uva.nl

João Miguel Cunha
University of Coimbra,
CISUC/LASI, DEI
jmacunha@dei.uc.pt

Philipp Wicke
LMU, Munich
Munich Center for Machine Learning (MCML)
philipp.wicke@lmu.de

Abstract

Language and Vision-Language Models exhibit impressive language capabilities akin to human reasoning. However, unlike humans who acquire language through embodied, interactive experiences, these models learn from static datasets without real-world interaction. This difference raises questions about how they conceptualize abstract notions and whether their reasoning aligns with human cognition. We investigate spatial conceptualizations of LLMs and VLMs by conducting analogy prompting studies with LLMs, VLMs, and human participants. We assess their ability to generate and interpret analogies for spatial concepts. We quantitatively compare the analogies produced by each group, examining the impact of multimodal inputs and reasoning mechanisms. Our findings indicate that generative models can produce and interpret analogies but differ significantly from human reasoning in their abstraction of spatial concepts – variability is influenced by input modality, model size, and prompting methods, with analogy-based prompts not consistently enhancing alignment. Contributions include a methodology for probing generative models through analogies, a comparative analysis of analogical reasoning among models, and humans, and insights into the effect of multimodal inputs on reasoning.[1]

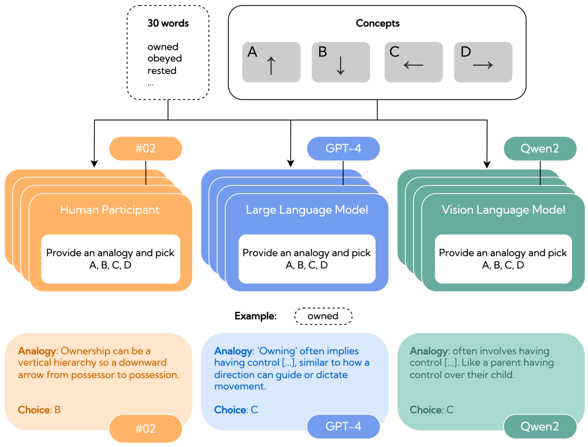

Figure 1: Human participants (e.g., participant #02), LLMs (e.g., GPT-4) and VLMs (e.g., Qwen2-VL) are prompted to provide an analogy for their choice of 1 of 4 items (↑, ↓, ←, →) that best represents 1 of 30 words.

1 Introduction

Large language models (LLMs) have revolutionized natural language processing, achieving remarkable language proficiency and emergent abilities that seem to parallel human reasoning (Brown et al., 2020; Achiam et al., 2023; Kojima et al., 2022). Trained on vast corpora of text – or paired text and images for vision-language models (VLMs) – these models' learning paradigms fundamentally differ from human language acquisition, raising questions about how they represent meaning, form abstract ideas, and structure knowledge.

LLMs learn from static, digital artifacts, processing accumulated language data without real-time interaction or sensory experience. Their training spans weeks to months using massive computational resources (Hoffmann et al., 2022; Scao et al., 2023). In contrast, human language acquisition is an embodied process: children learn through dynamic interactions with their environment – observing, testing, and experiencing the world around them (Mandler, 1992). First words emerge around 12 months, alongside nonverbal communication (Bretherton and Bates, 1979; Iverson, 2010), and foundational language abilities develop over approximately five years, with sensory experiences and social interactions playing crucial roles (Clark and Casillas, 2015).

Despite these differences, both LLMs and humans produce language artifacts and exhibit reasoning grounded in language. This raises a fundamental question: **How can LLMs exhibit rea-**

[1] Code available at: https://github.com/cisnlp/spatial_intuitions

Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 108–132

soning abilities seemingly analogous to human cognition when their training procedures are so fundamentally different? Addressing this is crucial as we integrate LLMs/VLMs into systems where reasoning and understanding are essential.

Studies have highlighted limitations in LLM reasoning capabilities – they often struggle with complex reasoning tasks (Mondorf and Plank, 2024; Stechly et al., 2023), arithmetic operations (Gambardella et al., 2024), planning (Valmeekam et al., 2022), and other challenges (Sobieszek and Price, 2022). One potential issue is how LLMs abstract from their knowledge. It is argued that human cognition largely relies on analogical reasoning, i.e., understanding abstract concepts by relating them to familiar ones (Gentner, 1983; Hofstadter, 2001). Analogies facilitate learning and are a crucial component for human cognitive development (Vosniadou and Ortony, 1989; Holyoak, 2012).

We focus on analogical reasoning to investigate whether LLMs and VLMs can generate and interpret analogies like humans to understand abstract spatial associations. Specifically, we address: (**RQ1**): How do LLMs and VLMs conceptualize semantic notions through spatial analogies compared to humans? (**RQ2**): How do multimodal inputs (e.g., text and images) affect the models' analogical reasoning? To answer these questions, we conduct analogy prompting studies (i.e., requiring to produce an analogy to answer a question) with LLMs, VLMs, and human participants. We systematically categorize and compare the analogies generated by each group. Our experiments examine the influence of different modalities, testing state-of-the-art VLMs with image inputs to assess how sensory information impacts reasoning outcomes. Our contributions are:

1. Methodology for probing conceptualization in models through analogy generation;
2. Comparative analysis of analogical reasoning abilities of LLMs, VLMs and humans, using both quantitative and qualitative approaches;
3. Insights into how multimodal inputs influence models toward human-like reasoning;
4. Evaluation of whether different types of models, e.g., those with enhanced reasoning, improve analogy and conceptual understanding.

2 Related Work

2.1 Analogical Reasoning in Cognition

Analogical reasoning is a key cognitive strategy which allows individuals to draw parallels between disparate domains by mapping relational structures. Gentner's structure-mapping theory posits that analogy involves aligning relational structures from a base domain to a target domain, emphasizing the importance of systematic correspondences over mere attribute similarities. Gust et al. (2008) argue that analogies underpin key cognitive abilities – memory adaptation, transfer learning, reasoning, and creativity – by enabling the application of prior knowledge to novel contexts; they propose that analogical reasoning is fundamental for integrating diverse cognitive processes in large-scale systems. Evidence for the connection between human reasoning and analogies comes from several psycholinguistic studies (Richardson et al., 2001; Beitel et al., 2001; Gibbs et al., 1994). They provide evidence that certain linguistic representations are grounded in spatial schemas, which operate as analogical structures for language comprehension.

2.2 Analogical Reasoning in AI Models

Analogical reasoning in AI has gained attention through various benchmarks and methodologies, revealing both the strengths and limitations of LLMs. Sultan and Shahaf (2022) detail a mechanism to extract analogies from a corpus of data describing a situation or a process. The entities of these texts are extracted and a mapping between these entities, or a cluster of entities, is build, connecting two texts in an analogy-like relation. Sourati et al. (2024) introduce the Analogical Reasoning on Narratives (ARN) benchmark, which extends traditional analogy evaluations by integrating narrative elements. This framework distinguishes near from far analogies, demonstrating LLMs' proficiency in surface mappings yet exposing their limitations with abstract, far analogies under zero-shot conditions. Another benchmark is the AnaloBench (Ye et al., 2024), which tests the capabilities of LLMs to find analogies in a large dataset of texts. Short sentence analogies and analogies contained in a larger paragraph of text are tested, and the authors demonstrate that models like GPT3.5 and GPT4 still struggle to recognize analogies, especially with an increase in text size.

In this context, Yu et al. (2023) propose Thought Propagation (TP), a method that leverages the generation and resolution of analogous problems to iteratively refine model outputs, thereby achieving significant improvements over conventional baselines. Furthermore, Yuan et al. (2024) develop a knowledge base containing analogies, and show that training language models on this database improves the model's ability to recognize and generate analogies. Complementing these approaches, Webb et al. (2023) compare LLM performance with human reasoning across varied analogy tasks, showing that while models like GPT-3 rival humans in structured analogies, they struggle with causal and cross-domain reasoning. Furthermore, Petersen and van der Plas (2023) align model evaluations with human-like paradigms, and Hu et al. (2023) show how encoding visual information into textual representations enhances LLMs' performance on visual analogical reasoning, as they demonstrate with Raven's Progressive Matrices.

Chain-of-thought (CoT) prompting encourages step-by-step reasoning in zero-shot settings (Kojima et al., 2022). In few-shot settings, when examples contain analogies, the model is explicitly guided to apply analogical reasoning (Wei et al., 2022b,a). Moreover, the term "analogy prompting" has already been used by Yasunaga et al. (2024), albeit in a different context. The authors further the idea of chain-of-thought by prompting the model to find similar math or coding problems in its knowledge base before trying to solve a given problem. They show that this methodology improves the ability of the model to solve math and coding problems in comparison to zero-shot and few-shot CoT. In the context of this paper, however, "analogy prompting" refers to prompting the model to generate analogies.

2.3 Spatial Schemas

Understanding how LLMs and VLMs conceptualize foundational spatial schemas is crucial for robust, intelligent systems. These schemas are the basic building blocks that infants use for spatial integration – a process described by Mandler (1992) as synthesizing perceptual experiences into conceptual representations via analogical reasoning.

Zhang et al. (2025) test the spacial reasoning of VLMs by asking spacial-related questions about a given image, i.e., "Is the blue ball in front of the red ball?", and "From the blue ball's point of view, is the red ball to the right of the blue ball?". They find that VLMs' answers tend to not be robust and consistent, especially when they are asked to adopt a different frame of reference.

Richardson et al. (2001) study spatial schemas in adults and finds that commonly used verbs are consistently associated with a specific spatial direction (horizontal vs. vertical), which highlights the importance of spatial schemas in semantic representations even after the developmental stage.

Wicke and Wachowiak (2024) and Wicke et al. (2024) focus on the same stimuli used in Richardson et al. (2001) and assess whether a suite of LLMs and VLMs exhibits word-direction associations similar to humans'. Our work substantially extends their effort by using analogy-based prompting to gain deeper insights into model reasoning, incorporating state-of-the-art VLMs, and conducting a human subject study that not only validates previous results but also provides human analogies for direct comparison with those of models.

3 Methods

3.1 Experimental Setup

Our aim is to explore spatial intuitions in both humans and multimodal models by bridging a psycholinguistic study with computational modeling. We build upon the original study by Richardson et al. (2001), which provides the experimental stimuli of words and schematic directions (up, down, left, right) but has not been reproduced in over 20 years and did not explore the use of analogies. We conduct a human subject experiment where participants associate words with schematic directions and, additionally, provide the analogies they use for these associations (see Fig. 1). We query a variety of LLMs and VLMs – including GPT-4o (OpenAI, 2024a), Llama3 (AI@Meta, 2024), Molmo (Deitke et al., 2024), Qwen2-VL (Wang et al., 2024b), and others – with *regular* and *analogy* (i.e., explicitly asking the model to provide an analogy and to use it to provide its answer) prompts. We quantify the correlation between model and human schema selection in both prompting conditions and systematically compare the analogies generated by humans and models. These comparisons provide insights into how prompting strategies, modalities, and model architectures affect spatial associations.

Stimuli and Modalities In order to keep our results comparable to those by Richardson et al. (2001), we use the same stimuli as the original study (depicted in Fig. 2). The original stimuli include 30 verbs and pictures showing arrows. In Richardson et al.'s study, participants were asked

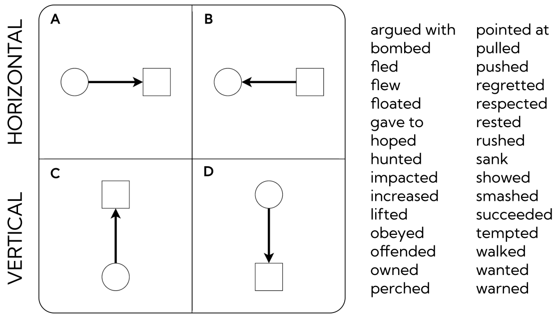

Figure 2: Left: Schematic directions used in all experiments. Right: Action words as experimental items. Both sets are adapted from Richardson et al. (2001).

to choose a preferred arrow (spatial schema) to represent each verb. In case of our studies, we present these spatial schemas in three different renderings: i) a reproduction of the original images (*visual condition*), ii) an equivalent Unicode version (↑, ↓, ←, →) of the arrows (*pseudo-visual condition*), and iii) a textual description (up, down, left, right) of the spatial schemas (*textual condition*).

3.2 Human Subject Study

We replicate the experiment by Richardson et al. (2001) with two key modifications designed to enhance both the task setup and subsequent analysis.

First, we introduce a one-shot example that diverges from the original relational schema (up, down, left, right) but retains a similar structure, designed to familiarize participants with the task without revealing the target relations (see App. Fig. 6). Second, we ask participants to provide an analogy explaining their choice before selecting one of the four options (see App. Fig. 7). Participants are asked to provide informed consent and demographic information (reported in App. A.2). We recruit 24 native English speakers, resulting in a total of 240 responses (30 items with 8 responses per item).

Schema Choice Evaluation To compare the results of our human study with those of Richardson et al. (2001), we calculated *item-level agreement* using a normalized concentration metric. This metric is based on the squared proportions of values within each distribution, ensuring it ranges from 0 (complete disagreement) to 1 (complete agreement). To account for sample size differences, scores were weighted by the number of observations (N) in both datasets. Overall agreement was computed as the weighted average across all items, with variability assessed via standard deviation, offering insights into the consistency of item-level distributions.

Labeling Analogies To facilitate comparisons between human and model-generated analogies, we design a classification schema that categorizes them into four types (more details in App. A.4):

- Physical Action Representation
- Interaction or Relationship Between Entities
- Cultural or Conventional Associations
- No Analogy or Direct Explanation Provided

The creation of these labels was guided by prior NLP work in analogy classification (Mikolov et al., 2013; Gladkova et al., 2016; Drozd et al., 2016), as well as recent advancements in analogy evaluation (e.g., Wijesiriwardene et al., 2025). With guidance from these sources and insights from their analysis, our labels account for semantic and pragmatic influences on the structure of the analogy.

To label our dataset of +7,000 analogies, we employ LLMs as judges while acknowledging their limitations in reliability (Zheng et al., 2023; Bavaresco et al., 2024). On samples of 3x30 analogies from both human and LLM data, two annotators achieve an agreement of Cohen's $\kappa = 0.6277$ after three annotation schema revisions, indicating their substantial agreement (Cohen, 1960).

When prompted according to this revised schema, GPT-4o achieves an agreement with two human annotators of Fleiss' $\kappa = 0.6024$ (Fleiss and Cohen, 1973) (see details in App. A.4).

3.3 Generative Model Study

Large Language Models We select a diverse set of state-of-the-art LLMs, including both open-source and proprietary architectures. As open-source models, we include two variants of Llama 3.1 – Llama-70B and Llama-70B-Instruct (AI@Meta, 2024) – and DeepSeek's R1-Distill-Llama (DeepSeek-AI et al., 2025), based on Llama-3.3-70B-Instruct. As proprietary models, we evaluate GPT-3.5-Turbo (OpenAI, 2023), GPT-4o, GPT-4o-Mini (OpenAI, 2024a), and GPT-o1-Preview (OpenAI, 2024b), accessed via the OpenAI API. LLMs were prompted by passing schemas as *textual* and *pseudo-visual* renderings.

Vision-language Models Given the documented limitations of vision-language models in spatial reasoning (Kamath et al., 2023; Wang et al., 2024a), we conduct a preliminary analysis to verify their ability to correctly process the input images used in the main experiment (see App. B.2 for more details). VLMs from the LLaVA family (Liu et al., 2024c,a,b) were found to be incapable of reliably identifying our stimuli, and therefore excluded from our main experiment. Our selection of VLMs includes Molmo-7B, Molmo-72B (Deitke et al., 2024), Qwen2-VL-7B, and Qwen2-VL-72B (Wang et al., 2024b). These models were prompted with schemas in their *visual* rendering (as images).

Prompts We test both LLMs and VLMs in two prompting conditions (all with temperature 0, except for GPT-o1). In *regular* prompting, models are simply asked to provide their chosen schema for each verb; in the *analogy* prompting condition, they are asked to rely on an analogy to choose a schema, and to include both analogy and chosen schema in their response. Both kinds of prompts are one-shot, i.e., they include an example question, in-line with the human subject study. The complete list of prompts used for all models is provided in App. B.3. As suggested by Aher et al. (2023), we employ prompt validation to enhance the validity of model responses (see App. B.1 for more details). Despite these mitigation efforts, some invalid responses persisted (see App. B.4 for details).

3.4 Evaluation Metrics

We evaluate our models along two main dimensions: schema selection (textual, pseudo-visual, and visual) and labeled analogies.

For both dimensions, we compare model outputs and human responses with Spearman correlations and F1 scores (see App. A.6 and B.5 for more details). While the schema selection evaluation was performed against both human datasets, the one regarding analogy labels is only applicable to our dataset, because Richardson et al.'s data does not include human-generated analogies.

In addition to these task-level comparisons, we perform item-level analyses. For the human data, we assess the agreement between our human samples and the original data using item-level agreement measures. Moreover, we examine the item-level correlations of analogy types between selected models by comparing their outputs to our human-sampled analogies.

4 Results and Discussion

4.1 Human Subject Study

Our human study partially aimed to replicate Richardson et al. (2001), albeit with significant procedural differences. The item-level agreement analysis that we performed to compare Richardson et al.'s results to ours yields an overall weighted agreement of 0.49 (±0.15) for Richardson et al.'s schema choices and 0.62 (±0.26) for ours. Notably, items such as *pointed at* (0.80), *pushed* (0.78), and *bombed* (0.76) obtain the highest agreement in the Richardson dataset, whereas our dataset shows perfect agreement for items like *fled*, *pulled*, *sank*, and *increased*, albeit with a smaller sample size.

Altogether, our results indicate that the overall item-level agreement for our data is higher than that reported by Richardson et al. (2001). For further details, please refer to App. Tab. 2. We interpret the higher agreement in our dataset as suggesting that analogy prompting induces participants to deeply engage their knowledge about spatial schemas, as opposed to relying on simpler associations.

4.2 Generative Model Study

Our study with generative models focuses on comparing model outputs with human responses on two levels. First, we investigate how strongly the spatial schemas chosen by models align with those chosen by human participants from both our experiment and Richardson's. Second, we explore the similarity between analogies generated by models and those provided by participants in our study.

4.2.1 Alignment of Spatial Schema Selection

We quantify alignment between models' and humans' schema choices by computing Spearman correlations and F1 scores. The former are shown in Fig. 3 and consider answer distributions aggregated per main direction ('horizontal' vs. 'vertical'); this choice was favored over considering all four spatial schemas as it yielded more statistically significant correlations. F1 scores are reported in Tab. 1 and were calculated considering all four spatial schemas (up, down, left, right). Both Spearman correlations and F1 scores were computed per prompting condition (regular and analogy) and input type (textual, pseudo-visual, and visual).

Regular vs. analogy prompting Since we explicitly instructed our participants to employ analogical reasoning while Richardson et al. did not, we expected analogy-prompting model responses

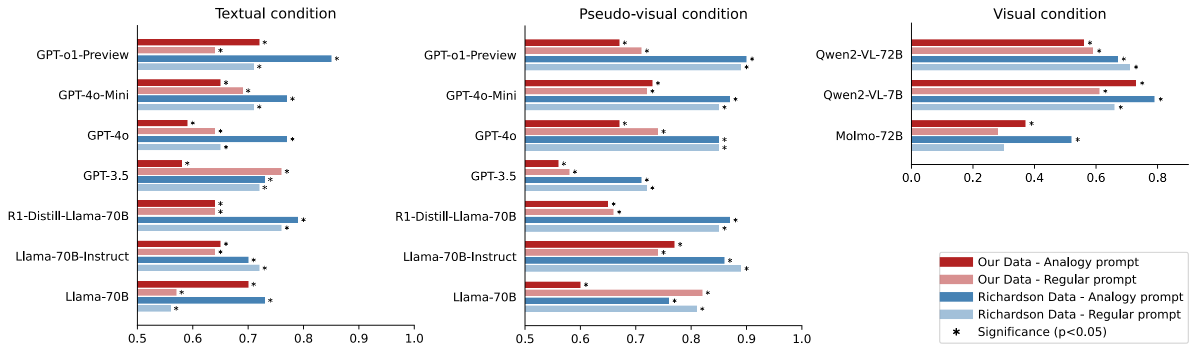

Figure 3: Spearman correlations between model and human chosen-concept distributions in the textual, pseudo-visual, and visual condition for our and Richardson et al.'s data. Values were computed per direction ('horizontal': up/down and 'vertical': left/right). Note that the x-axis range in the visual condition is different from the other two.

to align more closely with our dataset, and regular-prompting ones to be more aligned with Richardson et al.'s dataset. However, the Spearman correlations visualized in Fig. 3 indicate that none of the prompting conditions results in systematically stronger correlations with human responses. Moreover, the effect of the prompting condition is inconsistent even when the same model outputs are compared with different human datasets. As an example of this, in the textual condition, analogy prompting results in GPT-4o correlating more strongly with Richardson's data than ours ($\rho_{Rich.} = 0.45 > \rho_{Ours} = 0.29$). A similar effect can also happen for the same model in different experimental conditions – e.g., for Llama-70B analogy prompting yields higher correlations with our dataset than regular prompting in the textual condition ($\rho_{Analog.} = 0.70 > \rho_{Reg.} = 0.57$), but the reversed trend is observed in the pseudo-visual condition ($\rho_{Analog.} = 0.60 < \rho_{Reg.} = 0.82$). Regarding the schema-wise F1 scores reported in Tab. 1, they do not indicate a systematic advantage of analogy prompting for our human data. However, an interesting trend is that, albeit with a few exceptions, analogy prompting tends to result in higher F1 scores for Richardson et al.'s data. Taken together, these findings suggest that models may process analogical relationships differently from humans, potentially relying more on learned associative patterns than true analogical reasoning.

Effect of input type Spearman correlations visualized in Fig. 3 allow a comparison among between input types (textual, pseudo-visual, visual). Overall, we observe stronger correlations in the pseudo-visual condition ($\rho = 0.56$–0.90) than in the textual condition ($\rho = 0.58$–0.85), but the trend is not systematic. A similar trend can be detected in the F1 scores (Tab. 1), whose range is 0.29–0.63 in

Textual condition				
Model	**Our**		**Richardson**	
	R	A	R	A
GPT-3.5	0.46	0.49	0.60	**0.63**
GPT-4o	0.33	0.29	0.40	**0.45**
GPT-4o-Mini	**0.46**	0.35	0.45	0.40
GPT-o1-Preview	0.35	0.44	0.35	**0.49**
Llama-70B	0.50	0.38	**0.51**	0.40
Llama-70B-Inst	0.33	0.37	0.41	**0.48**
R1-Distill-Llama-70B	0.45	0.41	0.53	**0.58**
Pseudo-visual condition				
Model	**Our**		**Richardson**	
	R	A	R	A
GPT-3.5	0.35	0.50	0.53	**0.61**
GPT-4o	0.41	0.42	0.58	**0.63**
GPT-4o-Mini	0.48	0.45	**0.64**	0.63
GPT-o1-Preview	0.50	0.46	0.64	**0.67**
Llama-70B	0.34	0.47	0.44	**0.51**
Llama-70B-Inst	0.46	0.49	0.6	**0.63**
R1-Distill-Llama-70B	0.49	0.45	**0.69**	0.63
Visual condition				
Model	**Our**		**Richardson**	
	R	A	R	A
Molmo-72B	0.05	**0.16**	0.05	0.15
Qwen2-VL-7B	0.23	0.22	0.18	**0.34**
Qwen2-VL-72B	0.35	0.38	0.41	**0.51**

Table 1: Weighted F1 scores between human and models' concept preferences in the textual, pseudo-visual and visual conditions. Scores are reported for both our collected dataset and Richardson's, and for the two different prompting conditions (**R** indicates regular prompting and **A** analogy prompting). Figures were computed concept-wise, i.e., considering all four spatial schemas.

the textual condition and 0.34–0.69 in the pseudo-visual condition. One plausible explanation for this is that Unicode symbols reduce semantic ambiguities – particularly for words like "right" – which, in textual contexts, could be conflated with its "correctness" meaning. By providing a less ambiguous representation, pseudo-visual prompts may thus facilitate more accurate analogical mappings. Finally, correlations achieved by VLMs in the visual condition are, in general, lower than those achieved by LLMs in the other conditions ($\rho = 0.28$–0.79). This may be due to the visual condition posing the extra challenge of decoding the content of the visual stimuli. In other words, while LLMs receive abstract textual or pseudo-visual stimuli – which they can directly combine with their pretraining knowledge – VLMs are first tasked with mapping the different image(s) to abstract spatial notions and, only after completing this initial step, can they engage with their pretraining knowledge.

F1 scores and unbalanced concept productions For some models, we observe systematic concept over- and underproductions, which affect the weighted F1 scores provided in Tab. 1. For example, Molmo-72B never produces 'down' and 'right' in the regular-prompt setup, while overproducing the answer 'up' (in 97% of its outputs); this results in an extremely low F1 score (0.05) for both our human responses and Richardson et al.'s. Similarly, Qwen2-VL-7B generates 'up' in 73% of the cases in the regular-prompting setup. Across all LLMs, there is a systematic trend to underproduce the concept 'left', and in some cases 'down'. This tendency is especially extreme, e.g., for GPT-3.5 regular-prompted in the pseudo-visual condition (5% of 'left' responses), GPT-4o analogy-prompted in the textual condition (9% of 'left' responses), and Llama-70B regular-prompted in the pseudo-visual condition (8% of 'down' responses); in these cases, unbalanced model responses are again reflected in comparatively low F1 scores. Notably, while human participants also underproduce 'left' (19% in both datasets), this imbalance is not substantial enough to suggest a bias in the stimuli themselves. Instead, the models' consistent underrepresentation of 'left' is more likely an artifact of biases in training data.

4.2.2 Human- vs. Model-generated Analogies

The Spearman correlations quantifying the similarity between analogies provided by human participants and models are visualized in Fig. 4. Although correlations are non-significant, some interesting trends emerge. First, the types of analogies generated by VLMs are the most aligned with those provided by humans ($\rho = 0.23$–0.55). Second, LLMs do not systematically generate more human-like analogies in the textual vs. pseudo-visual condition ($\rho_{Text.} = 0.00$–0.17, $\rho_{Pseudo-vis} = 0.00$–0.20). Finally, it is interesting that the types of analogies produced by GPT-o1-Preview – the only reasoning model – are the least similar to the human-provided ones, with a Spearman correlation of 0 in the pseudo-visual condition. These findings suggest that multimodal pretraining, while not resulting in models closely mirroring human schema choices, may help VLMs generate analogy types that are more similar to human ones than LLMs' (examples of generated analogies in App. Tab. 4).

In a more focused analysis, we pick one LLM (GPT-4o) and check whether the items where its schema preferences align with the human ones are also those for which it generates more human-like analogy types. The results of this analysis are displayed in Fig. 5, which shows item-wise Spearman correlations with spatial schemas and analogy labels for the pseudo-visual condition. The correlations reveal a marked divergence between the models' analogical mappings and schema selections for several verbs (e.g., *gave to*, *impacted*, *obeyed*).

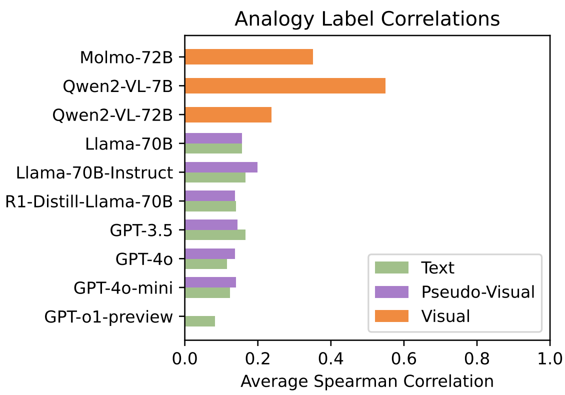

Figure 4: Correlations of the model's chosen analogy types with those analogy types chosen by humans.

These differences may be due to two possible scenarios. First, a model might produce analogies similar to human analogical associations while choosing different spatial schemas; this would suggest a decoupling between analogical similarity and spatial mapping within the model's reasoning process. Alternatively, a model might arrive at a

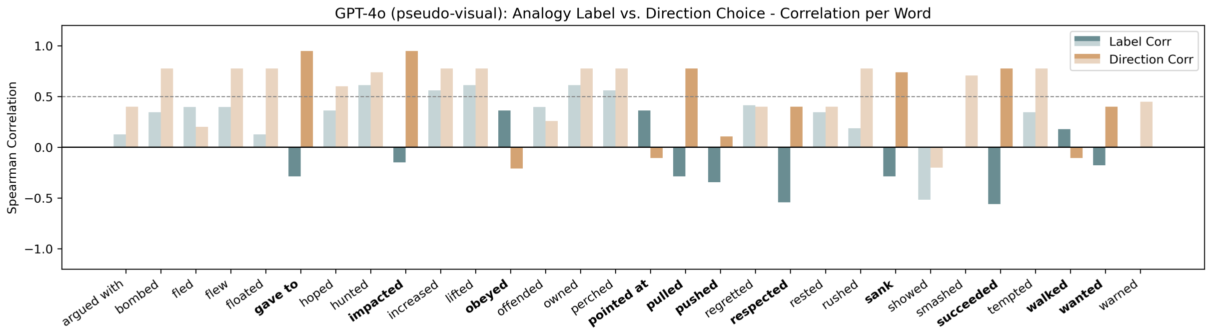

Figure 5: Spearman correlations for GPT-4o in the pseudo-visual condition, comparing human-model alignment on analogy labels (teal) and schema selection (ochre) responses for 30 words. Highlighted bars and labels denote words where analogy and direction correlations are opposed, showing cases of potential decoupling of the two.

similar directional assignment as humans, yet the underlying analogical reasoning, as reflected in the label correlation, diverges markedly from human responses. Both of these scenarios occur 8 times in our example (highlighted bars and words in Fig. 5).

Overall, both model-wise correlations (Fig. 4) and the item-level analysis (Fig. 5) seem to point towards a similar conclusion, i.e., that models' ability to produce analogies that resemble human ones does not necessarily result in human-aligned spatial-schema choices, and vice versa. This divergence is especially critical given that the words span abstract to concrete concepts, suggesting that the integration of analogical and spatial reasoning may be more fragile in contexts where multiple interpretative routes coexist.

4.3 Summary of Findings

Our analyses compare humans' and generative models' spatial intuitions on multiple levels (schema selection & analogy types) and consider two main experimental factors (prompting & modality). We now turn to our research questions.

RQ1 – Conceptualization of Abstract Notions through Analogies Our experiments reveal substantial discrepancies between models' and humans' spatial conceptualizations. At the level of alignment between spatial choices, we do not observe a systematic improvement associated with analogy vs. regular prompting. These findings, together with a comparison between analogy types generated by humans and models, show that, even when models generate analogies similar to the human ones, these do not result in more human-aligned spatial schema choices. More importantly, this is true even when considering our human dataset, which was collected by explicitly asking participants to rely on analogical reasoning. The discrepancies we document suggest that the profound differences between humans' and models' concept-learning processes are indeed reflected in spatial schemas, which appear to be supported by analogical reasoning in humans and simpler associations in models.

RQ2 – Effect of Multimodal Inputs on Analogical Reasoning Our comparisons between experimental conditions employing different input types (textual, pseudo-visual, and visual) reveal three interesting trends. First, LLMs tend to produce more human-aligned schema choices in the pseudo-visual condition, which is likely due to reduced semantic ambiguity. Second, VLMs' schema choices are, in general, less human-aligned than LLMs' ones. Indeed, while images should be, in principle, the least semantically ambiguous input type, they still posit the extra challenge of extracting abstract meaning from the input stimuli. Finally, we observe that VLMs tend to generate types of analogies that are more similar to the human ones than LLMs. Taken together, these findings suggest that VLMs' ability to process visual inputs proves advantageous in terms of producing human-like analogical reasoning. However, when focusing solely on associations between words and spatial schemas, Unicode arrows are the stimulus type associated with the most human-like choices; this may be due to them being abstract enough to not require perceptual processing and, at the same time, being less semantically ambiguous tokens than words.

5 Conclusions

Our study evaluates a suite of LLMs and VLMs concerning their ability to use analogical reasoning to support associations between verbs and spatial schemas, a core component of human concept learning processing. We employ regular and analogy prompts to elicit these associations and compare them with human data from Richardson et al. (2001) and a set of newly collected human responses which, in contrast to Richardson et al., include human-written analogies. In addition, we explore how stimulus types varying in their degree of abstractness (textual, pseudo-visual, visual) influence model responses. Our experiments reveal substantial discrepancies between models' ability to generate analogies similar to the human ones and their ability to associate verbs to spatial schemas in a human-like way. LLMs and VLMs are increasingly applied in domains beyond language, including robotics, navigation, medicine, scientific discovery, and autonomous systems. However, their limitations in complex tasks suggest that performance gaps cannot be solely attributed to model size. While scaling improves alignment with human responses, our findings indicate that underlying analogical structures and spatial intuitions may diverge from human reasoning. This study highlights the need to examine fundamental conceptualization mechanisms to better understand these discrepancies and refine future models accordingly.

Limitations

A key limitation of our study is the potential for data contamination in Richardson et al.'s dataset. While it is unlikely that proprietary LLMs were explicitly fine-tuned on this dataset, it is possible that Richardson et al.'s paper was included in the pretraining data of certain models. This raises concerns that some observed correlations may not reflect genuine analogical reasoning, but rather memorized associations from training corpora. At present, a key mitigation effort is the dataset collected in our study, which was not publicly available during our evaluation phase and thus was not included in the training data of any model.

Additionally, differences in experimental design between our dataset and Richardson et al.'s may introduce confounds. Our explicit analogy-based prompting method engages different cognitive strategies than the spontaneous associations likely employed in Richardson et al.'s experiment. While we anticipated that this methodological distinction would result in stronger correlations for analogy-prompted responses in our dataset, our findings did not consistently support this hypothesis. This discrepancy highlights the need for further research into how different prompting strategies interact with model architectures and training data to shape analogical reasoning performance.

We employed LLMs as annotation judges to assist in labeling our analogy dataset. This process followed an iterative refinement of the label classification schema, involving two human annotators, three rounds of revision, and the development of a carefully engineered prompt to ensure substantial agreement (Cohen, 1960). While we acknowledge the reliability limitations of LLM-based annotation (Zheng et al., 2023; Bavaresco et al., 2024), this approach offered certain advantages over human annotators, particularly in mitigating inconsistencies that arose even within the same annotator.

While our study examines the reasoning capabilities of models, we include only a single designated "reasoning model" (o1-Preview). We acknowledge that such models may provide additional insights into underlying reasoning processes. However, as of now, they rely on advanced, predefined reasoning templates that are non-deterministic and not openly accessible. Furthermore, our focus is on capturing the models' intuitions after a single analogical reasoning step, rather than tracing multiple, potentially opaque reasoning iterations.

Responsible Research

Use of Artifacts We use both open and proprietary language models in our work. For all models, we include model cards or references to their respective providers, which specify their licenses and intended usage. Additionally, we use GitHub Copilot, powered by OpenAI Codex, and ChatGPT to generate code snippets. These tools provide outputs that are licensed for free use, ensuring compliance with their intended access conditions.

We also utilize research data from Richardson et al. (2001) and Wicke and Wachowiak (2024), which are publicly available research papers. The data derived from these sources is used strictly within research contexts, in accordance with their original access conditions. To the best of our knowledge, the use of all artifacts aligns with their specified terms, ensuring compliance with licensing and intended use policies.

Use of AI Assistance We used AI assistance tools (ChatGPT, OpenAI Playground, and GitHub Copilot) to aid in rewriting code, filter large datasets to identify additional trends, and refining our labeling schema. All AI-generated content was thoroughly reviewed and verified by the authors. AI was not used to generate new research ideas or original findings; rather, it served as a support tool to improve clarity, efficiency, and organization. In accordance with ACL guidelines, our use of AI aligns with permitted assistance categories, and we have transparently reported all relevant usage in this paper. While AI contributed to enhancing the quality of the work, no direct research outputs are the result of AI assistance.

Acknowledgments

J.M.C's work is partially financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025; and supported by the Portuguese PRR through project C645008882-00000055, Center for Responsible AI. A.B. was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

AI@Meta. 2024. Introducing llama 3.1: Our most capable models to date. Blog post.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, ..., and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Dinara A Beitel, Raymond W Gibbs Jr, and Paul Sanders. 2001. The embodied approach to the polysemy of the spatial preposition on. In *Polysemy in cognitive linguistics*, pages 241–260. John Benjamins.

Inge Bretherton and Elizabeth Bates. 1979. The emergence of intentional communication. *New Directions for Child and Adolescent Development*, 1979(4):81–100.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, ..., and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Bangkok, Thailand. Association for Computational Linguistics.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Raymond W Gibbs, Dinara A Beitel, Michael Harrington, and Paul E Sanders. 1994. Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of semantics*, 11(4):231–251.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Helmar Gust, Ulf Krumnack, Kai-Uwe Kühnberger, and Angela Schwering. 2008. Analogical reasoning: a core of cognition. *Künstliche Intell.*, 22(1):8–12.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Douglas R. Hofstadter. 2001. Epilogue: Analogy as the core of cognition. In *The Analogical Mind: Perspectives from Cognitive Science*. The MIT Press.

Keith J Holyoak. 2012. Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, pages 234–259.

Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. In-context analogical reasoning with pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.

Jana M Iverson. 2010. Developing language in a developing body: The relationship between motor development and language development. *Journal of child language*, 37(2):229–261.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jean M Mandler. 1992. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. In *First Conference on Language Modeling*.

OpenAI. 2023. Gpt-3.5 turbo fine-tuning and api updates.

OpenAI. 2024a. Gpt-4o model card.

OpenAI. 2024b. Introducing openai o1-preview.

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16414–16425, Singapore. Association for Computational Linguistics.

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*, volume 23.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, ..., and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Working paper or preprint.

Adam Sobieszek and Tadeusz Price. 2022. Playing games with ais: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Stella Vosniadou and Andrew Ortony. 1989. *Similarity and analogical reasoning*. Cambridge University Press.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2024a. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Philipp Wicke, Lea Hirlimann, and João Miguel Cunha. 2024. Using analogical reasoning to prompt llms for their intuitions of abstract spatial schemas. In *First Workshop on Analogical Abstraction in Cognition, Perception, and Language at IJCAI 2024*.

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*, pages 6102–6117, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Sreeram Reddy Vennam, Vinija Jain, Aman Chadha, Amitava Das, Ponnurangam Kumaraguru, and Amit Sheth. 2025. Knowledgeprompts: Exploring the abilities of large language models to solve proportional analogies via knowledge-enhanced prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3979–3996.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. (arXiv:2310.01714). ArXiv:2310.01714 [cs].

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. Analobench: Benchmarking the identification of abstract and long-context analogies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.

Junchi Yu, Ran He, and Zhitao Ying. 2023. Thought propagation: An analogical approach to complex reasoning with large language models. In *The Twelfth International Conference on Learning Representations*.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. (arXiv:2410.17385). ArXiv:2410.17385 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Human Study

A.1 Survey Design

The survey was conducted using *Google Forms*. All participants provided their informed consent to participate in our study. No names, addresses, IPs or traceable information was collected, and the participants could decide to end the study at any point. In order to familiarize the participants with the task, an example task was provided (Fig. 6). The example task used the same format as the real task, but the symbols and the direction (diagonal as opposed to vertical/horizontal) were different. We tested the survey design with peers before collecting responses from non-peers. The test responses have not been included in the final data collection.

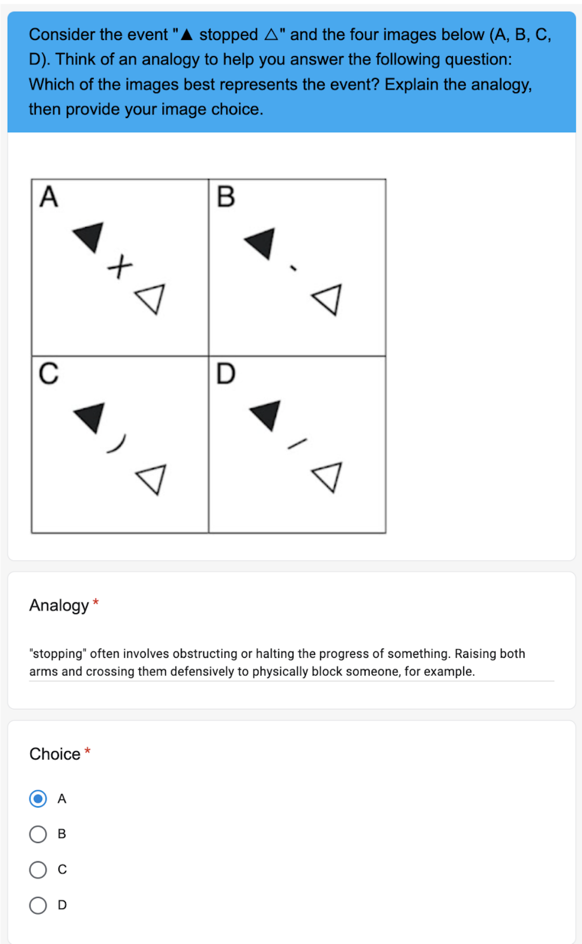

Figure 6: All participants in the study are presented with an example item (one-shot) at the start of the questionnaire. This allows the participants to familiarize themselves with the task, while not providing a priming effect due to the use of a different directionality (diagonal as opposed to vertical/horizontal) and different symbols (triangles as opposed to circle/square).

For each of the 30 items, we generated a question shown in Fig. 7. We use the same visual stimuli as Richardson et al. (2001) for our human subject study. We note that in the original study, the participants were presented with the entire list of 30 items at once (next to the same picture, which we repeat for each item).

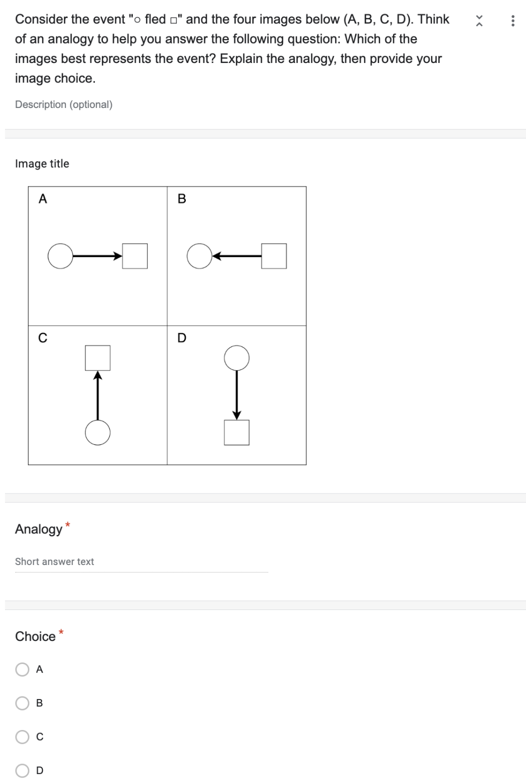

Figure 7: Example item presented to the participants. First, they are asked to provide an analogy, then they are asked to choose one of four images that best relates to the options (A, B, C, D).

A.2 Demographics

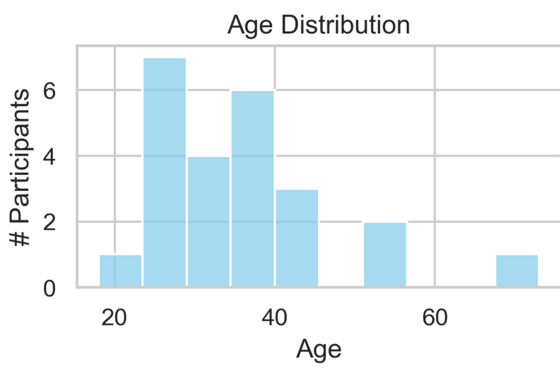

Figure 8: Distribution of age for $N = 24$ participants. Average age is 35.54.

We sampled $N = 24$ participants with two restrictions: (i) Native English speakers, (ii) no prior knowledge about this research. To the best of our knowledge, no participant self-reported significant or severe visual or cognitive impairments.

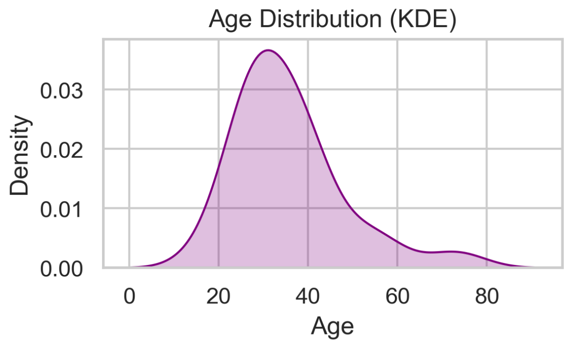

Figure 9: Kernel density estimate (KDE) to represent participants' ($N = 24$) age as spectrum, with an average around 35 years.

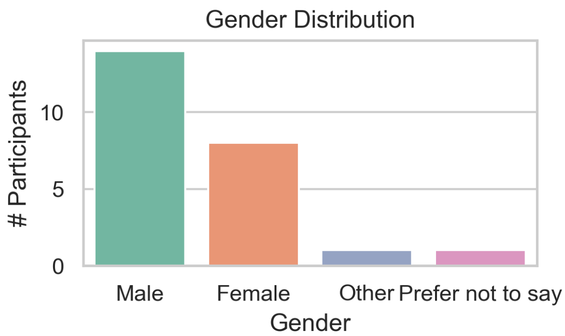

Figure 10: Gender distribution of all $N = 24$ participants: Male: 14 participant(s), female: 8 participant(s), other: 1 participant, prefer not to say: 1 participant.

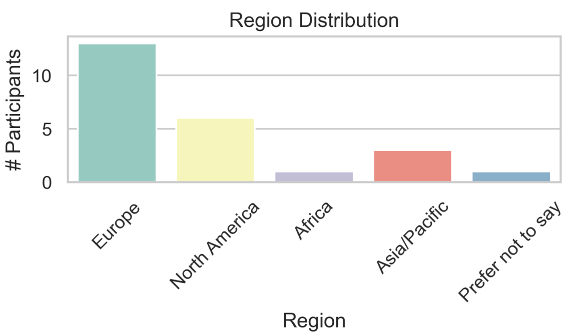

Figure 11: All participants declared that they are native English speakers. The regional distribution is as follows: Europe: 13 participant(s), North America: 6 participant(s), Africa: 1 participant, Asia/Pacific: 3 participant(s), Prefer not to say: 1 participant.

A.3 Human Study Results

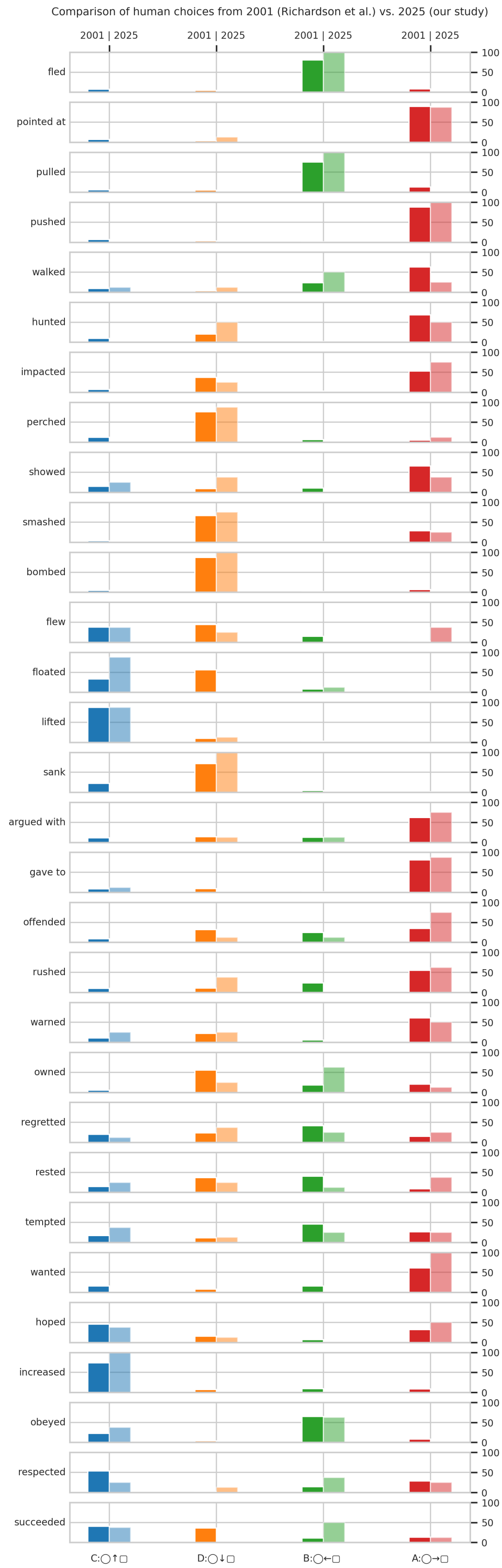

Figure 12: Comparison of the data by Richardson et al. (2001) with the human choices gathered in our study.

A.4 Analogy Annotation Methodology

We sampled 30 analogies (15 human-created, 15 GPT-4o-generated) and classified them into four categories: “Physical Action,” “Cultural/Convention,” “Interactive Entities,” or “No Analogy/Explanation.” In a second round, two authors annotated a different set of 30 analogies using this scheme. Annotator agreement was measured using Cohen’s κ (Cohen, 1960). After three revisions of the annotation scheme, we achieved $\kappa = 0.6277$, indicating substantial agreement. All annotation schema versions are available in the code repository. The final schema, incorporating these revisions as additional rules, was then formalized into a prompt:

> **Task:** You will be provided with an explanation that uses a directional or movement analogy to describe an event, action, or reaction. Your job is to carefully read the explanation, assess the type of analogy it employs, and select one of the following labels that best corresponds to it:
>
> - **Physical Action** – This label applies if the explanation relies on tangible movements, forces, or physical processes.
> - **Cultural/Convention** – This label applies if the explanation relies on societal norms, symbolic interpretations, or culturally shared meanings related to direction or movement.
> - **Interactive Entities** – This label applies if the explanation emphasizes the interaction or relationship between distinct entities (e.g., square and circle).
> - **No Analogy/Explanation** – This label applies if the explanation is purely descriptive, with no directional, movement-based, or analogical content.
>
> **Additional rules:**
>
> - If the explanation mentions “square” or “circle,” it is always labeled **Interactive Entities**.
> - If the explanation does not mention these shapes implicitly or explicitly, and no entities are present, then it is not **Interactive Entities**.
> - If the explanation mentions “culture,” it is always **Cultural/Convention**.
> - If the explanation includes technical or scientific analogies (e.g., diagrams or systems), it is always **Cultural/Convention**.
> - If the explanation references gravity, understand gravity as a physical action and assign **Physical Action**.
>
> Here is the explanation: `Explanation`
>
> Based solely on your analysis of the explanation above, provide only one label from the following: **Physical Action, Cultural/Convention, Interactive Entities, or No Analogy/Explanation**.

A.5 Choice Coherence

Item	Richardson	Our (w/ analogy)
pointed at	0.80	0.78
pushed	0.78	**1.00**
lifted	0.77	0.78
bombed	0.76	**1.00**
fled	0.67	**1.00**
gave to	0.67	0.78
perched	0.60	0.78
pulled	0.59	**1.00**
sank	0.57	**1.00**
increased	0.57	**1.00**
smashed	0.53	0.62
hunted	0.52	0.50
obeyed	0.48	0.53
walked	0.47	0.34
showed	0.47	0.34
argued with	0.44	0.59
warned	0.44	0.38
floated	0.43	0.78
wanted	0.43	**1.00**
impacted	0.42	0.62
owned	0.39	0.47
respected	0.39	0.28
rushed	0.38	0.53
flew	0.36	0.34
hoped	0.34	0.41
rested	0.32	0.28
tempted	0.32	0.28
succeeded	0.32	0.41
regretted	0.29	0.28
offended	0.29	0.59
Overall	**0.49 (±0.15)**	**0.62 (±0.26)**

Table 2: Item-wise agreement scores for the choice (of direction) measure computed using a normalized concentration metric (i.e., squared proportions weighted by the number of observations, yielding values from 0 to 1). This metric quantifies how concentrated the responses are for each item -— scores near 1 signify that nearly all raters converge on the same label (indicating high consensus), whereas lower values reflect greater variability in judgments. “Richardson” refers to the human data reported by Richardson et al. (2001) and “Our” refers to the data collected in the present study. The final row gives the overall weighted agreement and its standard deviation.

A.6 Label Evaluation

For each word, we first compute frequency distributions over the four label categories from human responses (8 responses) and model responses (24 responses). These distributions are then converted into ranked vectors by ordering categories according to their frequencies. Spearman correlation is computed between the human and model ranked frequency vectors, quantifying the monotonic agreement in label usage. In parallel, for each category, the F1 score is calculated via

$$F_1 = \frac{2 \times \min(\text{count}_{\text{human}}, \text{count}_{\text{model}})}{\text{count}_{\text{human}} + \text{count}_{\text{model}}} \quad (1)$$

(with a default score of 1 when both counts are zero).

Model	Condition	Int. Coh. ↑	JS Div. ↓	Entr. ↓
Human Reference	Ref.	0.550	—	1.760
gpt-3.5	Pseudo	**0.933**	0.436	0.920
gpt-4	Pseudo	0.876	0.443	0.883
gpt-4-mini	Pseudo	0.839	0.443	0.904
llama-70b	Pseudo	0.929	**0.399**	0.830
llama-70b-inst	Pseudo	0.922	0.436	**0.642**
gpt-3.5	Text	0.907	0.417	0.813
gpt-4	Text	0.861	0.449	0.981
gpt-4-mini	Text	0.874	0.450	0.678
llama-70b	Text	0.929	0.409	0.885
llama-70b-inst	Text	0.856	0.443	0.910

Table 3: Evaluation metrics for five LLM configurations under Pseudo and Text conditions compared to a human reference. "Int. Coh." (**Internal Coherence**) is the average fraction of label agreement per item, reflecting labeling consistency. "Entr." (**Entropy**) quantifies the diversity of the label distribution, and "JS Div." (**Jensen–Shannon Divergence**) measures the similarity of the model's distribution to that of humans.

A.7 Label Examples

B Model Studies

B.1 Validation Scores

In order to improve model responses, we tested different prompt endings and calculated a validation score that measured how often the model, when given a regular prompt, produced a valid response. To achieve this, we generated a model response for each of the 30 action words using the following prompt:

```
Given the concepts: 'X', '-',
')', '/'. For the concept
that best represents the event
'stopped', what concept would
you choose?
[ending]: 'X'

Given the concepts: '[concept]'.
For the concept that best
represents the event '[event]',
what concept would you choose?
[ending]:
```

where [ending] is one possible prompt ending (e.g., "CONCEPT", "Choice", and "selection"), [concept] refers to the four spatial concepts, and [event] is an action word. We employed Llama3.1-8B (AI@Meta, 2024) as the LLM for this experiment, based on the rationale that if a smaller model can produce a valid answer with a specific ending, then larger models are likely to do so as well. As described in Wicke and Wachowiak (2024), the validity score is calculated as follows:

$$\sum_{a \in V} p(a|k) \quad (2)$$

for a specific prompt k and a set of valid answers V.

This evaluation was conducted under both textual and pseudo-visual conditions, and an average validation score was computed for each ending. As a result, the ending "concept" was selected, achieving a 74% validation score. The full list of prompt endings, along with their corresponding validation scores, is presented in Table 5.

B.2 Vision-language Models Validation

The task considered in our main experiment is challenging even for current vision-language models for two main reasons. First, it requires correctly distinguishing the arrow depictions mainly based on their spatial orientation. Second, it requires the ability to process multiple images simultaneously. Since most state-of-the-art VLMs still struggle with spatial reasoning (Kamath et al., 2023; Wang et al., 2024a) or are simply not trained for multi-image inference, we designed a simple validation task to help us select VLMs that could be suitably included in our main experiment. The task consisted in responding correctly to the four possible variations of the prompt:

```
Which of the following is the [up,
down, left, right]ward-pointing
arrow?
Image SHTL: <image up>
Image XHWK: <image down>
Image AKRC: <image left>
Image ZHRN: <image right>
Please respond with the
four-letter image id.
```

The following similar prompt was used for a single-image scenario, where all the concepts were provided within the same image:

```
Which is the [up, down, left,
right]ward-pointing arrow? SHTL,
XHWK, AKRC or ZHRN?
```

Action	Contributor	Example Analogy	Assigned Label
wanted	Human	circle wants square. as circle is first its on the left, furthermore the arrow should point towards the square as it shows circle acting towards square	Interactive Entities
fled	GPT-3.5 (pseudo-visual)	'Fled' often involves moving quickly away from a dangerous or threatening situation. An analogy could be a person running away from a predator in a straight line to escape.	Physical Action
pointed at	GPT-4o-mini (pseudo-visual)	'Pointing at' something typically involves directing attention towards a specific object or direction, often using a finger or hand to indicate where to look. The action of pointing is about indicating a specific location or item in space.	Physical Action
pulled	GPT-4o (pseudo-visual)	'Pulling' often involves drawing something towards oneself or in a specific direction. When you pull an object, you typically exert force to bring it closer to you or to move it in a desired direction. In a directional sense, pulling can be associated with moving something from a lower position to a higher one, as if lifting or drawing it upwards.	Physical Action
showed	GPT-3.5 (textual)	'showing' often involves presenting or revealing something in a particular direction. For example, pointing towards a specific direction to indicate where something is located.	Physical Action
obeyed	GPT-4o-mini (textual)	'obeying' often involves following directions or commands, which can be likened to moving in a specific direction as instructed. When someone is told to go 'up', they are complying with a directive, just as one would follow orders or rules in a broader sense.	Cultural/Convention
rushed	GPT-4o (textual)	'Rushing' often involves moving quickly and with urgency towards a destination or goal. It implies a sense of forward momentum and progress, similar to how one might move in a straight line without hesitation. In many contexts, moving 'up' can symbolize advancement, progress, or moving towards a goal, as it is often associated with positive movement or elevation.	Cultural/Convention
argued with	Qwen-VL-72b	'argued with' often involves opposing or challenging someone's views. A debate between two people, for example, is a common representation of this event.	Interactive Entities
hoped	Qwen-VL-7b	'hoping' involves having a desire or wish for something to happen. It's like having a goal or aspiration.	No Analogy / Explanation

Table 4: Examples of different collected analogies from different contributors. Selection was focused on representing different assigned labels. Full collection of analogies is available at https://github.com/anonymousACL/analogy_prompting.

Ending	Textual	Pseudo	Avg.
CHOICE	0.53	0.65	0.59
Choice	0.65	0.72	0.68
choice	0.70	**0.77**	0.73
SELECTION	0.66	0.73	0.69
Selection	0.69	0.75	0.72
selection	0.68	0.75	0.71
CONCEPT	0.68	0.75	0.71
Concept	0.69	0.73	0.71
concept	**0.73**	0.76	**0.74**

Table 5: Overview of the validation scores for each possible prompt-ending, for textual and pseudo-visual prompts, along with their average.

The models tested in the multi-image scenario were `Qwen2-VL-7B-Instruct`[2] and `llava-onevision-qwen2-7b-ov-hf`[3]. The models tested in the single-image scenario were: `Molmo-7B-D-0924`[4], `llama3-llava-next-8b-hf`[5], `llava-v1.6-mistral-7b-hf`[6],

[2] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[3] https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf

[4] https://huggingface.co/allenai/Molmo-7B-D-0924

[5] https://huggingface.co/llava-hf/llama3-llava-next-8b-hf

[6] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

`llava-onevision-qwen2-7b-si-hf`[7], `llava-interleave-qwen-7b-hf`[8], and `Qwen2-VL-7B-Instruct`[9].

The only models which were able to respond correctly to all variants of the prompts were `Molmo-7B-D-0924` in the single-image scenario and `Qwen2-VL-7B-Instruct` in the multi-image scenario. Given the satisfactory performance of these 7B-parameter models, we decided to include their largest versions (`Molmo-72B-0924`[10] and `Qwen2-VL-72B-Instruct-AWQ`[11]) as well in the main experiment.

B.3 Prompts

The prompts used for the LLMs and vision-language models are reported, respectively, in Tables 6, 7, and 8. To avoid selection bias (e.g., the model always choosing the option appearing as first), for each prompt we constructed variations corresponding to all the possible label permutations ($4! = 24$).

Note that, since the preview Molmo version available when experiments were conducted (Fall 2024) did not support multi-image inference, this model was prompted with a single image including all four spatial schemas. As for the Qwen2-VL models, they were found incapable of discriminating between schemas when they were provided within the same image; therefore, each schema was provided within a separate image.

B.4 Parsing of Model Outputs

Despite our efforts to validate the prompts, there were still cases where model-generated responses did not exactly match the expected structure. When this occurred, we first tried to exploit other regularities (e.g., the model outputting `choice:` instead of `concept:`) to isolate the relevant part of the output. When no such regularity was present, we adopted a simpler single-matching approach: if a single concept could be identified in the output, we considered that as a valid answer; if not, or in the case where multiple concepts were present, we considered the output invalid.

To obtain comparable label distributions, we replaced the invalid answers with the prevalent valid answer for the action word. If no valid answer was returned for a specific action word, we excluded the action word from further comparisons with human preferences. The percentage of invalid answers never exceeded 5%. We report the percentage of invalid responses yielded by each model in Table 15.

B.5 Evaluation Metrics

For each model, we obtained 24 outputs for each verb-stimulus (corresponding to all possible permutations). This allowed us to obtain a percentage of ‘up’, ‘down’, ‘left’ and ‘right’ responses for each verb. Similarly, response distributions could be obtained for the human datasets based on multiple participants’ answers. Spearman correlations quantifying the alignment between human and model responses were computed between the human and model-generated answer distributions per each spatial schema. Since most correlations were non-significant when considering all four spatial schemas, in the main paper we show correlations per direction (‘vertical’ vs. ‘horizontal’). In the following sections, we report complete results, including both per-schema and per-direction correlation.

Regarding F1 scores, they were computed between human and model-generated schema preferences. For each verb, the schema preference was defined as the schema appearing in most permutations or indicated by the majority of participants.

B.6 Extended Results

In Tables 9, 10, and 11 the Spearman correlations for the LLMs and VLMs are represented. F1 scores for these models can be seen in Tables 12, 13, and 14.

B.7 Compute Resources

Running our experiments required a total of approximately 37 hours on an A100 NVIDIA GPU and 10 hours on an H100 NVIDIA GPU. Labeling the analogies with GPT-4o using OpenAI’s API access required approximately 3 hours.

[7] https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-si-hf

[8] http://llava-hf/llava-interleave-qwen-7b-hf

[9] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[10] https://huggingface.co/allenai/Molmo-72B-0924

[11] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct-AWQ

Model	Prompt
Llama (R)	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a participant in a research experiment<|eot_id|><|start_header_id|>user<|end_header_id|> Given the concepts: ’X’, ’-’, ’)’, ’/’. For the concept that best represents the event ’stopped’, what concept would you choose?<|eot_id|><|start_header_id|>assistant<|end_header_id|> concept: ’X’<|eot_id|><|start_header_id|>user<|end_header_id|> Given the concepts: ’[concept]’. For the concept that best represents the event ’[event]’, what concept would you choose? <|eot_id|><|start_header_id|>assistant<|end_header_id|> concept:
Llama (A)	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a participant in a research experiment<|eot_id|><|start_header_id|>user<|end_header_id|> Given the concepts: ’X’, ’-’, ’)’, ’/’. For the concept that best represents the event ’stopped’, what concept would you choose? Explain the analogy, then provide one choice.<|eot_id|><|start_header_id|>assistant<|end_header_id|> analogy: ’stopping’ often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example. concept: ’X’<|eot_id|><|start_header_id|>user<|end_header_id|> Given the concepts: ’[concept]’. For the concept that best represents the event ’[event]’, what concept would you choose? Explain the analogy, then provide one choice.<|eot_id|><|start_header_id|>assistant<|end_header_id|> analogy:
Llama (Z)	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a participant in a research experiment<|eot_id|><|start_header_id|>user<|end_header_id|> Given the concepts: ’[concept]’. For the concept that best represents the event ’[event]’, what concept would you choose? Give the chosen concept by surrounding it with ’##’.<|eot_id|><|start_header_id|>assistant<|end_header_id|> Let’s think step by step.

Table 6: Prompts used for the Llama 3.1 models. The R1-Distill-Llama model used the same prompt as the regular Llama models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical vs. **Z**ero-shot.)

Model	Prompt
GPT (R)	SYSTEM_PROMPT: You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure. USER_PROMPT: EXAMPLE TASK: Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose? concept: 'X' TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? concept:
GPT (A)	SYSTEM_PROMPT: You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure. USER_PROMPT: EXAMPLE TASK: Given the concepts: 'X', '-', ')', '/'. For the concept that best represents the event 'stopped', what concept would you choose? Explain the analogy, then provide one choice. analogy: 'stopping' often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example. concept: 'X' TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Explain the analogy, then provide one choice. analogy:
GPT (Z)	You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective. Follow the given structure. TASK: Given the concepts: '[concept]'. For the concept that best represents the event '[event]', what concept would you choose? Give the chosen concept by surrounding it with '##'. Let's think step by step.

Table 7: Prompts used for the GPT models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical. vs. **Z**ero-shot.)

Model	Prompt
Molmo (R)	Example task: Consider the event ‘threw’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Which of the images best represents the event? Image: XHWK Task: Consider the event ‘[event]’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Which of the images best represents the event? Image:
Qwen2-VL (R)	Example task: Consider the event ’stopped’ and these four images: SHTL [image], XHWK [image], AKRC [image], ZHRN [image]. Which of the images best represents the event? Image: SHTL Task: Consider the event ’stopped’ and these four images: [image label][image], [image label]<image>, [image label]<image>, [image label]<image>. Which of the images best represents the event? Image:
Molmo (A)	Example task: Consider the event ‘threw’ and the four images below (SHTL, XHWK, AKRC, ZHRN). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy: ‘throwing’ often involves launching an object in a horizontal direction. The trajectory followed by the object could be represented as a rightward-pointing arrow. Image: XHWK Task: Consider the event [event] and the four images below (SHTL, XHWK, AKRC, ZHRN). Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy:
Qwen2-VL (A)	Example task: Consider the event ‘stopped’ and these four images: SHTL <image>, XHWK <image>, AKRC <image>, ZHRN <image>. Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy: ‘stopping’ often involves obstructing or halting the progress of something. Raising both arms and crossing them defensively to physically block someone for example. Image: SHTL Task: Consider the event ‘[event]’ and these four images: [image label]<image>, [image label]<image>, [image label]<image>, [image label]<image>. Think of an analogy to help you answer the following question: Which of the images best represents the event? Explain the analogy, then provide your image choice. Analogy:

Table 8: Prompts used for the vision-language models. The letters in brackets after the model names refer to the experimental condition (**R**egular vs. **A**nalogical.)

		Llama-70B		Llama-70B-Inst			R1-Distill-Llama-70B	
		R	A	R	A	Z	R	A
Richardson	Up	0.45*	0.53* (+)	**0.67***	0.57* (-)	0.61* (-)	0.48*	0.63* (+)
	Down	**0.47***	0.31 (-)	0.31	0.27 (-)	0.33 (+)	0.37*	0.44* (+)
	Left	0.34	0.44* (+)	0.36	0.46* (+)	0.07 (-)	0.25	**0.47*** (+)
	Right	0.58*	0.56* (-)	0.58*	0.57* (-)	**0.62*** (+)	**0.62***	0.61* (-)
	$\uparrow$	0.67*	0.58* (-)	**0.72***	0.66* (-)	0.68* (-)	0.68*	0.57* (-)
	$\downarrow$	**0.66***	0.38* (-)	0.48*	0.49* (+)	0.48* (=)	0.58*	0.62* (+)
	$\leftarrow$	0.12	0.61* (+)	0.42*	0.44* (+)	0.33 (-)	0.43*	**0.62*** (+)
	$\rightarrow$	0.47*	0.61* (+)	0.67*	0.72* (+)	**0.77*** (+)	0.69*	0.68* (-)
	Hor./Vert.^{T}	0.56*	0.73* (+)	0.72*	0.70* (-)	0.72* (=)	0.76*	**0.79*** (+)
	Hor./Vert.^{P}	0.81*	0.76* (-)	**0.89***	0.86* (-)	0.88* (-)	0.85*	0.87* (+)
Ours	Up	0.57*	**0.58*** (+)	0.56*	0.51* (-)	0.48* (-)	0.47*	**0.58*** (+)
	Down	0.47*	0.45* (-)	0.43*	0.40* (-)	0.40* (-)	0.53*	**0.57*** (+)
	Left	0.38*	0.42* (+)	0.39*	0.36 (-)	0.17 (-)	0.36*	**0.49*** (+)
	Right	**0.47***	0.41* (-)	0.37*	0.35 (-)	0.37* (=)	0.36*	0.33 (-)
	$\uparrow$	**0.70***	0.52* (-)	0.64*	0.60* (-)	0.66* (+)	0.64*	0.50* (-)
	$\downarrow$	**0.60***	0.51* (-)	0.52*	0.53* (+)	0.52* (=)	0.50*	0.50* (=)
	$\leftarrow$	0.12	**0.59*** (+)	0.38*	0.53* (+)	0.44* (+)	0.44*	0.55* (+)
	$\rightarrow$	0.37*	0.45* (+)	0.50*	0.52* (+)	**0.56*** (+)	0.53*	0.41* (-)
	Hor./Vert.^{T}	0.57*	**0.70*** (+)	0.64*	0.65* (+)	0.62* (-)	0.64*	0.64* (=)
	Hor./Vert.^{P}	**0.82***	0.60* (-)	0.74*	0.77* (+)	0.70* (-)	0.66*	0.65* (-)

Table 9: Spearman correlations between concept distributions by humans and the open-source models (Llama3.1 and DeepSeek R1 Distill Llama). The last four rows report results aggregated into two main directions ('up' and 'down' into 'vertical' and 'left' and 'right' as 'horizontal'), for textual (T) and pseudo-visual (P) concepts. Values in the 'R' column refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting, and 'Z' indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results in an improved correlation with respect to regular prompting (+), remained the same (=), or didn't improve (-). Asterisks mark statistical significance ($p < 0.05$).

		GPT-3.5		GPT-4o		GPT-4o-Mini		GPT-o1-Preview		
		R	A	R	A	R	A	R	A	Z
Richardson	Up	**0.63***	0.48* (-)	0.59*	0.61* (+)	0.61*	**0.63*** (+)	0.60*	0.58* (-)	0.57* (-)
	Down	**0.51***	0.35 (-)	0.41*	0.45* (+)	0.26	0.22 (-)	0.41*	0.34 (-)	0.35 (-)
	Left	0.43*	**0.52*** (+)	0.32	0.45* (+)	0.36	0.47* (+)	0.35	0.45* (+)	0.26 (-)
	Right	**0.69***	0.68* (-)	0.52*	0.65* (+)	0.59*	0.60* (+)	0.59*	**0.69* (+)**	0.55* (-)
	↑	0.58*	0.47* (-)	**0.73***	0.68* (-)	0.69*	0.63* (-)	0.64*	0.69* (+)	0.66* (+)
	↓	0.55*	0.32 (-)	**0.59***	0.52* (-)	0.56*	0.36 (-)	**0.59***	0.52* (-)	0.47* (-)
	←	0.23	0.29 (+)	0.36	0.49* (+)	0.52*	0.43* (-)	0.46*	**0.53*** (+)	0.21 (-)
	→	0.69*	0.63* (-)	0.68*	0.64* (-)	0.74*	**0.76*** (+)	0.70*	0.68* (-)	0.67* (-)
	Hor./Vert.^{T}	0.72*	0.73* (+)	0.65*	0.77* (+)	0.71*	0.77* (+)	0.71*	**0.85*** (+)	0.74* (+)
	Hor./Vert.^{P}	0.72*	0.71* (-)	0.85*	0.85* (=)	0.85*	0.87* (+)	0.89*	**0.90*** (+)	0.86* (-)
Ours	Up	0.60*	0.44* (-)	**0.63***	0.58* (-)	0.61*	0.56* (-)	0.55*	0.49* (-)	0.49* (-)
	Down	**0.62***	0.44* (-)	0.49*	0.41* (-)	0.33	0.37* (+)	0.54*	0.48* (-)	0.45* (-)
	Left	0.36*	**0.56*** (+)	0.38*	0.38* (=)	0.38*	0.50* (+)	0.24	0.36 (+)	0.10 (-)
	Right	0.47*	0.50* (+)	0.37*	0.36 (-)	0.40*	0.40* (=)	0.43*	**0.57*** (+)	0.44* (+)
	↑	0.54*	0.46* (-)	0.63*	**0.67*** (+)	0.59*	0.64* (+)	0.55*	0.56* (+)	0.58* (+)
	↓	0.54*	0.36 (-)	0.55*	0.54* (-)	0.58*	0.45* (-)	**0.59***	0.51* (-)	0.39* (-)
	←	0.25	0.28 (+)	0.34	**0.54*** (+)	0.50*	0.42* (-)	0.51*	**0.54*** (+)	0.35 (-)
	→	0.52*	0.42* (-)	0.44*	0.47* (+)	0.50*	**0.54*** (+)	0.48*	0.48* (=)	0.50* (+)
	Hor./Vert.^{T}	**0.76***	0.58* (-)	0.64*	0.59* (-)	0.69*	0.65* (-)	0.64*	0.72* (+)	0.59* (-)
	Hor./Vert.^{P}	0.58*	0.56* (-)	**0.74***	0.67* (-)	0.72*	0.73* (+)	0.71*	0.67* (-)	0.65* (-)

Table 10: Spearman correlations between concept distributions by humans and the GPT models. The last four rows report results aggregated into two main directions ('up' and 'down' into 'vertical' and 'left' and 'right' as 'horizontal'), for textual (T) and pseudo-visual (P) concepts. Values in the 'R' column refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting, and 'Z' indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results in an improved correlation with respect to regular prompting (+), remained the same (=), or didn't improve (-). Asterisks mark statistical significance ($p < 0.05$).

		Molmo-7B		Molmo-72B		Qwen2-VL-7B		Qwen2-VL-72B	
		R	A	R	A	R	A	R	A
Richardson	Up	0.11	0.29 (+)	0.19	0.32 (+)	0.22	**0.56*** (+)	0.53*	0.37* (-)
	Down	0.36	-0.17 (-)	–	-0.04	0.45*	**0.52*** (+)	0.50*	0.42* (-)
	Left	–	–	-0.27	-0.07 (+)	0.05	0.11 (+)	0.31	**0.36** (+)
	Right	0.34	-0.26 (-)	–	0.15	0.19	0.22 (+)	0.44*	**0.52** (+)
	Hor./Vert.	0.33	-0.25 (-)	0.30	0.52*(+)	0.66*	**0.79***(+)	0.71*	0.67* (-)
Ours	Up	-0.05	0.03 (+)	0.17	0.30 (+)	0.30	**0.46*** (+)	0.44*	0.28 (-)
	Down	0.11	-0.08 (-)	–	-0.11	0.26	**0.44*** (+)	0.31	0.37* (+)
	Left	–	–	-0.15	-0.10 (-)	0.25	0.13 (-)	**0.41***	0.37* (-)
	Right	0.30	-0.24 (-)	–	0.05	0.06	0.06	0.30	**0.33** (+)
	Hor./Vert.	0.23	-0.22(-)	0.28	0.37* (+)	0.61*	**0.73*** (+)	0.59*	0.56* (-)

Table 11: Spearman correlations between concept distributions by humans and vision-and-language models. Results are reported both per-concept and per-direction, i.e., aggregating 'up' and 'down' into 'vertical' and 'left' and 'right' into 'horizontal'. Values in the 'R' columns refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting. The signs in brackets signal whether analogy prompting results in an improved correlation with respect to regular prompting (+) or not (-). Asterisks mark statistical significance ($p < 0.05$).

		Llama-70B		Llama-70B-Inst			R1-Distill-Llama-70B	
		R	A	R	A	Z	R	A
Richardson	Concept^T	0.51	0.40 (-)	0.41	0.48 (+)	0.36 (-)	0.53	**0.58** (+)
	Concept^P	0.44	0.51 (+)	0.60	0.63 (+)	0.60 (=)	**0.69**	0.63 (-)
	Direction^T	0.73	0.64 (-)	0.65	0.72 (+)	0.53 (-)	0.83	**0.87** (+)
	Direction^P	0.60	0.70 (+)	0.83	0.90 (+)	0.80 (-)	**0.93**	0.90 (-)
Ours	Concept^T	**0.50**	0.38 (-)	0.33	0.37 (+)	0.33 (=)	0.45	0.41 (-)
	Concept^P	0.34	0.47 (+)	0.46	**0.49** (+)	0.42 (-)	**0.49**	0.45 (-)
	Direction^T	0.67	0.71 (+)	0.58	0.72 (+)	0.67 (+)	**0.77**	0.73 (-)
	Direction^P	0.52	0.70 (+)	0.70	**0.77** (+)	0.67 (-)	0.73	0.70 (-)

Table 12: Weighted F1 scores between human and the open-source models' concept preferences. The first two rows report results considering all four concepts (up, down, left, right) for textual (T), and ($\uparrow$, $\downarrow$, $\leftarrow$, $\rightarrow$) for pseudo-visual (P), while the last two rows aggregating them into two main directions (horizontal and vertical). Values in the 'R' column refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting, and 'Z' indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results improved F1 score with respect to regular prompting (+), remained the same (=), or didn't improve (-).

		GPT-3.5		GPT-4o		GPT-4o-Mini		GPT-o1-Preview		
		R	A	R	A	R	A	R	A	Z
Richardson	Concept^T	0.60	**0.63** (+)	0.40	0.45 (+)	0.45	0.40 (-)	0.35	0.49 (+)	0.40 (+)
	Concept^P	0.53	0.61 (+)	0.58	0.63 (+)	0.64	0.63 (-)	0.64	**0.67** (+)	**0.67** (+)
	Direction^T	0.87	**0.90** (+)	0.76	0.76 (=)	0.55	0.68 (+)	0.55	0.64 (+)	0.60 (+)
	Direction^P	0.80	**0.90** (+)	**0.90**	0.87 (-)	**0.90**	0.76 (-)	0.80	**0.90** (+)	0.83 (+)
Ours	Concept^T	0.46	**0.49** (+)	0.33	0.29 (-)	0.46	0.35 (-)	0.35	0.44 (+)	0.35 (=)
	Concept^P	0.35	**0.50** (+)	0.41	0.42 (+)	0.48	0.45 (-)	**0.50**	0.46 (-)	0.46 (-)
	Direction^T	**0.80**	0.63 (-)	0.62	0.55 (-)	0.62	0.61 (-)	0.62	0.71 (+)	0.67 (+)
	Direction^P	0.67	0.76 (+)	**0.77**	0.67 (-)	0.76	0.69 (-)	0.73	0.70 (-)	0.70 (-)

Table 13: Weighted F1 scores between human and GPT's concept preferences. The first two rows report results considering all four concepts (up, down, left, right) for textual (T), and ($\uparrow$, $\downarrow$, $\leftarrow$, $\rightarrow$) for pseudo-visual (P), while the last two rows aggregating them into two main directions (horizontal and vertical). Values in the 'R' column refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting, and 'Z' indicates *zero-shot* prompting. The signs in brackets indicate whether analogy prompting results improved F1 score with respect to regular prompting (+), remained the same (=), or didn't improve (-).

		Molmo-7B		Molmo-72B		Qwen2-VL-7B		Qwen2-VL-72B	
		R	A	R	A	R	A	R	A
Rich.	Concept	0.30	0.15 (-)	0.05	0.15 (+)	0.18	0.34 (+)	0.41	**0.51** (+)
	Direction	0.39	0.25 (+)	0.33	0.68 (+)	0.60	0.55 (-)	0.60	**0.90** (+)
Ours	Concept	0.20	0.12 (-)	0.05	0.16 (+)	0.23	0.22 (-)	0.35	**0.38** (+)
	Direction	0.44	0.32 (-)	0.40	0.61 (+)	0.60	0.62 (+)	0.52	**0.69** (+)

Table 14: Weighted F1 scores between VLM and human concept preferences from both Richardson's and our dataset. Results are reported for both concept preferences and direction preferences. Values in the 'R' columns refer to the *regular* prompting condition, while 'A' indicates *analogy* prompting. The signs in brackets signal whether analogy prompting results in an improved F1 score with respect to regular prompting (+) or not (-).

Model	% Inv. Resp. ↓			# AWs w/ Inv. Resp. ↓			# Removed AWs ↓		
	R	A	Z	R	A	Z	R	A	Z
Llama-70B^T	9.44	13.89	–	14	18	–	**0**	**0**	–
Llama-70B^P	2.50	9.72	–	10	14	–	**0**	**0**	–
Llama-70B-Inst^T	**0**	0.69	1.94	**0**	2	9	**0**	**0**	**0**
Llama-70B-Inst^P	**0**	0.28	6.94	**0**	2	16	**0**	**0**	**0**
R1-Distill-Llama-70B^T	**0**	0.28	–	**0**	1	–	**0**	**0**	–
R1-Distill-Llama-70B^P	0.14	0.69	–	1	2	–	**0**	**0**	–
GPT-3.5^T	0.14	1.53	–	1	3	–	**0**	**0**	–
GPT-3.5^P	**0**	0.42	–	**0**	1	–	**0**	**0**	–
GPT-4o^T	2.22	**0**	–	1	**0**	–	**0**	**0**	–
GPT-4o^P	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–
GPT-4o-Mini^T	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–
GPT-4o-Mini^P	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–
GPT-o1-Preview^T	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**
GPT-o1-Preview^P	**0**	**0**	**0**	1	**0**	**0**	**0**	**0**	**0**
Molmo-7B^V	17	**0**	–	5	**0**	–	5	**0**	–
Molmo-72B^V	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–
Qwen2-VL-7B^V	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–
Qwen2-VL-72B^V	**0**	**0**	–	**0**	**0**	–	**0**	**0**	–

Table 15: Overview of invalid responses in the **R**egular, **A**nalogy, and **Z**ero-shot prompting conditions, for the textual (T), pseudo-visual (P), and visual (V) conditions. The first column contains the overall percentage of invalid responses, the second the number of action words for which at least one invalid response was generated, and the last the number of action words that were removed because none of the generated answers was valid. A "–" indicates that the model was not evaluated under the corresponding prompting condition.