Salience of Surface versus Higher-level Properties in Spatial Comparison at Different Levels of Scene Similarity

Hoang Anh Vu (hvu@fandm.edu)[1], Irina Rabkina (irina.rabkina@barnstormresearch.com)[2], Laura M. Hiatt (laura.m.hiatt.civ@us.navy.mil)[3] & Jason R. Wilson (jrw@fandm.edu)[1]

[1]Department of Computer Science, Franklin and Marshall College, USA

[2]Barnstorm Research, USA

[3]Navy Center for Applied Research in AI, US Naval Research Laboratory, USA

Abstract

Identifying similarity plays a major role in spatial reasoning. Prior work has shown that identification of similarity and difference between spatial scenes involves reasoning about their properties at two different levels–surface-level properties, such as color and size, and higher-level compositional properties. The interaction between the use of these property types during reasoning and the degree of similarity or difference between spatial scenes has, however, been largely unstudied. We presented participants with in-progress Tangram puzzles at various stages of similarity to their target image. Participants were asked to identify the puzzle being built and provide an explanation of how they came to that conclusion. Analysis of explanations found an interaction between degree of similarity and preference for reasoning over surface-level or higher-level properties.

Keywords: tangrams; spatial reasoning; similarity; difference; structural similarity.

Introduction

Spatial reasoning is the ability to understand and mentally manipulate the spatial relationships of objects in the world. It is a fundamental human ability that is prevalent in both daily life (e.g., communicating about the world, navigating, moving objects around, etc.) and in science and engineering (e.g., construction, architecture, etc.). Research has shown that spatial reasoning ability correlates with students' performance in STEM such as geometry and math (Harris, 2023; Lee et al., 2009). The ability to perform comparisons between spatial scenes, in particular, has been linked to the ability to process visual stimuli such as maps, models, and graphs (Gattis, 2002; Morita et al., 2007; Shah & Hoeffner, 2002)

There is a large body of literature that considered how people describe similarities and differences between spatial scenes (e.g., Ji et al., 2022; Sagi et al., 2012; Shepard, 1974; Tversky, 1977, etc.) In this paper, we explore how varying the similarity between two images affect these descriptions and the underlying cognitive processes that they represent.

To study this, we consider how people reason about *unfinished Tangram puzzles*. Tangrams are puzzles that are solved by arranging different baseline polygons to form a larger target shape. While assembling such puzzles, there is a large number of possible paths to completion: the assembler can place any one of 7 pieces in any orientation, and in any relation to pieces that are already on the board. Thus, it is not always trivial to predict the target shape from the unfinished state. Furthermore, incomplete puzzles may contain errors in which the shapes are assembled incorrectly, creating additional differences from the target shape.

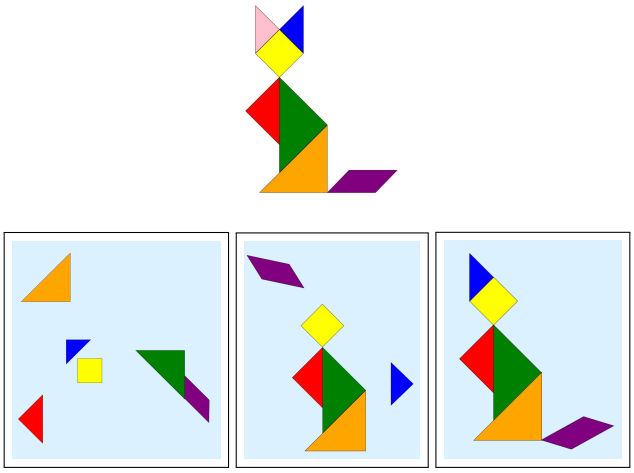


Figure 1: Illustrations of the cat puzzle at different stages. In the first image, the lack of progress leads to a puzzle highly dissimilar to the cat reference. The second puzzle contains more connections, leading to a recognizable structure resembling the cat. In the final image, most relations match that of the reference despite a clear difference with the pink piece.

Examples of incomplete Tangrams are shown in Figure 1. In order from left to right, the images display puzzles at different stages of assembly towards the same goal (i.e., a cat), which corresponds to different levels of similarity. While people may correctly identify all 3 puzzles as a(n incomplete) cat, the process by which they reach this conclusion may differ among these levels of completion. Previous work has suggested that the presence of salient higher-order shapes in images affects reasoning paths (Hashemi & Tubau, 2025; Li et al., 2019). In Tangram puzzles, salient higher-order shapes may be compositional *semantic* components, such as an animal's head or ears. On the other hand, *surface-level* properties such as color or shape of individual pieces would not be considered salient or higher-order. Participants' choice to use semantic or surface-level properties when describing an incomplete puzzle is likely to be reflective of the reasoning they use to identify it. Here, we explore whether the language, and thereby the reasoning process, varies as a function of similarity of the puzzle to the target shape.

In our study, we presented participants with incomplete

Tangram puzzles and asked them to identify which of three images was being assembled. We also asked them to explain how they made their decision. We then coded responses based on the languages used, including semantic and surface-level feature descriptions. We found that, when explaining their decision, people were more likely to reference surface properties where there are very few to no pieces connected or when the puzzle is nearly complete (i.e., the puzzle being assembled is very dissimilar or very similar to the correct target puzzle). In between these stages, we found that people preferred to refer to semantics. These findings suggest that the level of similarity between spatial puzzles affects the reasoning processes used to compare them.

Background

This paper focuses on the spatial reasoning used by adult observers of Tangram puzzles in various states of similarity to their target images. While, to the best of our knowledge, this paradigm has not been previously studied, there is existing literature on spatial reasoning more broadly. We discuss that work here.

Reasoning About Spatial Similarity

Tversky (1977) viewed the reasoning process about spatial similarity as the process of matching surface-level properties, known as features, between two objects. He argued that, during comparison, objects are broken down to their component properties. These properties are then matched one by one to identify similarities and differences. The perceived similarity between two objects is proportional to the linear combination of their similar and different features. Interestingly, this measure is not necessarily symmetric–a leopard may be considered to be more similar to a tiger than a tiger is to a leopard due to the difference in saliency of their properties.

Similarly, Sagi et al. (2012) found that people can identify whether two images are different much more quickly when the images are highly dissimilar than when they are highly similar, suggesting that a large number of different surface-level properties is sufficient to identify a mismatch. However, they found that identifying *what* is different between two images is slower for highly dissimilar images than for similar ones. They argued that this is because a full structural comparison is necessary to identify specific differences. Shared higher-level properties support the comparison, making it faster. Yet, the differences themselves, referred to as alignable differences (Markman & Gentner, 1993), are typically at the surface level.

Hashemi and Tubau (2025) directly studied whether and when people focus on surface-level properties (referred to as the object path) versus higher-order properties (the global path) in their reasoning. While the research found that participants overall prefer reasoning over higher-order properties, this is modulated by the properties' saliency. In other words, the extent to which the higher-order structures are obvious and recognizable affects the likelihood that participants will rely on them. This suggests that, in the present work, participants should be most likely to refer to semantic properties in their reasoning when they are most salient.

Spatial Reasoning with Tangrams

Tangram puzzles in particular have previously been used to study spatial reasoning. For example, Ji et al. (2022) asked participants to annotate the semantics of segments and overall shapes in complete Tangrams. In this case, *semantic* refers to salient compositional higher-level properties (e.g., the head of a person). They showed that participants are quite accurate in their annotations, selecting relevant areas and labeling them appropriately. However, such annotations are not evidence for how they may be used when performing spatial reasoning. Our study explores whether the proclivity for identifying higher-level properties affects identification of puzzles at different levels of similarity to their target scene. Thus, we hope to shed light on how and when higher-level versus surface-level properties are used during spatial reasoning.

Methods

ToMCAT Dataset

Our study utilized the Theory of Mind of Children Assembling Tangrams (ToMCAT) dataset (Wilson et al., 2025). The dataset originates from an experiment conducted by Langer et al. (2026) that videotaped 34 children, between the ages of 5 and 8 years old, assembling puzzles from start to finish, based on a grayscale version of a cat or a rabbit image. The ToMCAT dataset contained 436 frames of videos recorded in the experiment. Each frame was captured as an observation describing the changes in puzzle state (whether correct or incorrect) as a child assembled the puzzle. Each puzzle in the dataset has 7 pieces: 2 small triangles–pink and blue, 1 medium red triangle, 2 large triangles–green and orange, 1 purple parallelogram, and 1 yellow square. However, the assemblers did not receive the small pink triangle until the later stages of assembly, and so 75.9% of the frames in the ToMCAT dataset include only 6 pieces.

ToMCAT represents the frames in 3 forms: cropped screenshot of the video frames, abstract images of the screenshots replicated with draw.io, and logical descriptions of the spatial relationships between puzzle pieces. We used 427 of the abstract images as the visual stimuli for our experimental task. Due to a technical error, eight images were excluded from the Qualtrics survey and subsequent analysis; one image was excluded because it was not meaningfully different from its neighboring images (i.e., only one piece was moved by millimeters).

Tangram recognition task

We designed an experiment prompting participants to identify what Tangram shape was being constructed given an incomplete puzzle state.

When presented with the abstract puzzle image, the adult observer was asked to indicate which of 3 reference images the child was constructing–a cat, rabbit, or sailboat (see Figure

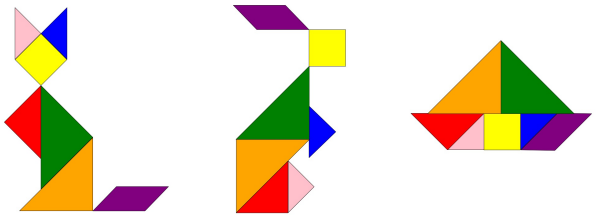


Figure 2: The reference images of a cat, rabbit, and a sailboat Tangrams, from left to right, provided to survey participants.

2). The sailboat was introduced as a false option in order to increase the complexity of the task–as described above, all children were assembling either a rabbit or a cat. Next, we requested a detailed explanation of the reasoning behind their choice: “Please explain your answer in detail. We are interested in how people reason about these puzzles. A clear description of how you came to your conclusion would be helpful. If the puzzle is complete, explain why you believe that to be true." Finally, participants were asked to rate their confidence in their answers using a percentage scale between 0 and 100.

Human data

To collect sufficient data, we conducted two rounds of human studies. The experimental study was approved by the IRB of the institute of the lead author.

704 participants were initially recruited for the study. Following data screening, 209 participants (29.7%) were excluded from the final analysis for failing attention checks, exhibiting insufficient effort (e.g., rapid completion times), or suspected use of generative AI for their responses. The researchers identified AI responses by examining their open-ended explanations. These answers often contained lengthy descriptions, mentions of puzzle pieces irrelevant to the task, and particular languages that are characteristics of AI.

The final analytical sample consisted of 495 participants, including Prolific participants (N = 415, M_{age} = 40.4) and college students pooled from introductory psychology courses (N = 80, M_{age} = 19.4). Prolific participants received an average compensation of $2.50, and college students received 0.5 hour research credit towards their psychology course. Overall, 244 participants identified as female, 245 as male, 2 as other gender identity, and 4 chose not to self-disclose. Most participants identified as White (68% Prolific, 59% college), followed by Black (19% Prolific, 6% college), Asian (5% Prolific, 15% college), Mixed (4% Prolific, 8% college), and other or undisclosed (4% Prolific, 9% college).

Prolific participants identified the puzzle in three different images, while the college students identified the puzzle in six different images. This resulted in 1718 total responses.

Coding Explanations

To analyze the types of reasoning used by participants, we coded their responses into categories that best described what they referenced. Coders labeled each explanation according to eight categories. In this paper, we focus our analysis on two categories in particular: "shapes/colors" and "semantics." Shapes and colors referred to responses that mentioned the surface-level properties of Tangram pieces. Typically, the explanations mentioned the polygonal shape and/or its color. A number of participants also combined relative sizes with polygon names when describing various triangles. The semantic category contained responses that named the higher-order properties represented by the puzzle (e.g., the head, body, sail, etc.), or the action/behavior depicted in the puzzle (e.g., sitting, standing).

Due to time constraint, we randomly selected 1105 of the 1718 responses from participants for coding and data analysis. The explanations were divided among 5 coders, including 3 authors included in the paper and 2 undergraduate research assistants. After coding, the primary researcher determined the final code for all responses.

Results

Accuracy

Participants’ accuracy was calculated with 1718 responses. Overall, participants correctly identified the puzzle’s target 77.84% of the time. The participants performed slightly better on the Rabbit puzzle compared to the Cat, with accuracies of 80.86% and 75%, respectively.

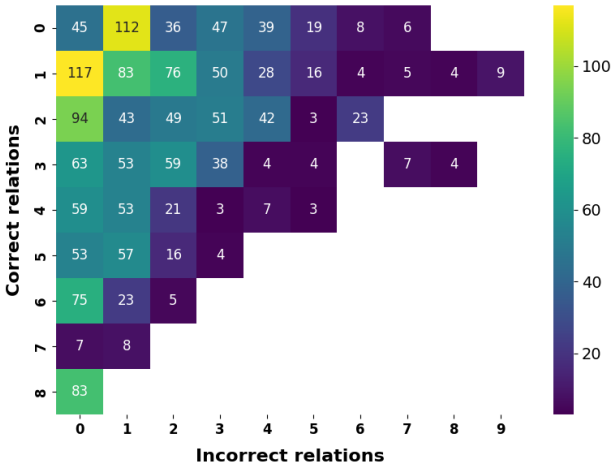


Figure 3: Heatmap showing distribution of observations across numbers of correct and incorrect relations (n = 1718).

To see how the reasoning and explanations changed with various states of the puzzles, we described each puzzle state using the number of correct and incorrect relations presented in the image. A correct relation occurs when two pieces are adjacent to each other in the right manner. An incorrect relation includes pairs of pieces that should not be adjacent to each other or were connected on the wrong edge or vertex (Wilson et al., 2025). Figure 3 shows the distribution of responses along these two dimensions. The disproportionate

Table 1: Logistic Regression Predicting Correct Inference

Predictor	*B*	*SE*	*z*	OR	95% CI for OR	*p*
Correct relations	0.63	0.06	10.85	1.87	[1.67, 2.10]	< .001
Incorrect relations	0.04	0.04	1.05	1.04	[0.97, 1.12]	.294
Confidence	0.02	0.003	8.99	1.02	[1.02, 1.03]	< .001
Constant	-1.52	0.19	-8.16	0.22	[0.15, -0.31]	< .001

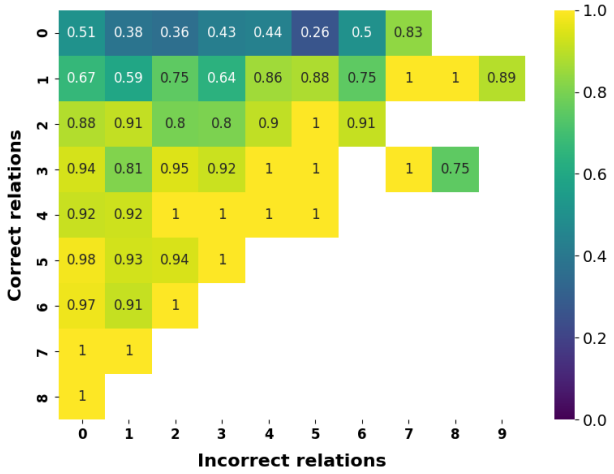


Figure 4: Heatmap showing average accuracy of participants across numbers of correct and incorrect relations (n = 1718).

number of responses with 0 or 1 correct relationships was due to the higher density of puzzles in the ToMCAT dataset that had 0 or 1 correct relationships (i.e., initial states of nearly all puzzles included at most 1 correct or incorrect relation; in addition, many children also decided to restart the puzzle during assembly, leading to more puzzles with few correct relationships).

We fitted a logistic regression model to predict whether participants correctly identified the puzzle (see Table 1). The number of incorrect relations was not significantly associated with correctly identifying the puzzle. Certain mistakes are closer to being correct than others (e.g., the right pieces were used but placed together incorrectly). Given this, it is unlikely that the *number* of incorrect relations would affect whether the participants chose the right image. On the other hand, the number of correct relations had a strong association with correctly identifying the puzzle. It is expected that people would correctly identify the puzzle when there were more correct relations because the puzzle was closer to completion and more closely resembled the reference image. The participant's confidence was also significantly related to correctly identifying the puzzle. Participants were more likely to have answered correctly when they felt confident, suggesting that they were able to assess how challenging the problem was.

While the overall accuracies are relatively high, there were situations in which many participants struggled to identify the puzzle. In particular, early stages of the puzzle, where few

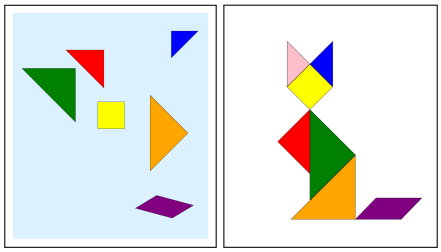


Figure 5: An example where no connections were present. Participants correctly inferred the puzzle by reasoning about the purple parallelogram placed at the bottom right area, similar to the cat image.

pieces have been correctly placed, resulted in a high number of errors. This is seen in the top two rows of heatmap in Figure 4, which shows the accuracy for each correct/incorrect combination. Once four correct relations were present, participants were at least 92% accurate. With only two correct relations, participants were still at least 80% accurate.

Participants were surprisingly accurate (51%) when there were no correct and incorrect relations, meaning no two pieces were connected. In these cases, we would expect participants to be at chance (33%). One reason participants were well above chance is because of cases where pieces were close to their relative position in the reference image (see Figure 5).

Explanations

Of the 1105 participant explanations, 399 used semantics, 463 referenced the shape or color of the pieces, and 181 used both. The remaining 62 responses were cases where participants did not use either descriptions (e.g., *The puzzle is the same as the cat*). Responses which exclusively used semantics inferred correctly 89.45% of puzzles. Those that only mentioned shapes and colors were 73.05% accurate. Lastly, when shapes and colors were used together with semantics, participants inferred correctly 88.95% of puzzles.

Table 2 displays the most common semantics used by participants. The mentions are not mutually exclusive, meaning participants can reference more than one feature in a response. Head, ear were overall the most frequently mentioned names (33.77% in cat puzzle and 45.29% in rabbit puzzle). When participants selected the cat as their inference, they most frequently mentioned its tail (43.86%). Only the cat image contains a tail, making it a clear indicator of the puzzle. Yellow and purple were among the most common surface-level properties mentioned by participants. These pieces are associated

with the head and tail in the cat puzzle and the head and ear of the rabbit puzzle. These distributions further suggest that people found the ear, head, and tail to be the highly salient features of the images.

Table 2: Distribution of Visual Feature Mentions by inference

	Cat		Rabbit	
Feature	Count	%	Count	%
Body Parts				
Head	77	33.77	77	45.29
Body	62	27.19	65	38.24
Tail	100	43.86	3	1.76
Ear	92	40.35	98	57.65
Leg	3	1.32	9	5.29
Face	14	6.14	9	5.29
Shapes and Colors				
Triangle	68	31.19	56	24.56
Square	25	11.47	38	16.67
Yellow	57	26.15	78	34.21
Purple	60	27.52	75	32.89
Red	31	14.22	28	12.28
Orange	32	14.68	42	18.42
Green	49	22.48	52	22.81
Blue	41	18.81	24	10.53
Pink	15	6.88	21	9.21

Note. Percentages indicate the proportion of mentions within each category.

The following analysis focuses on correct relations. Since this number progresses from zero to eight as the child assembles the puzzle, examining the content of explanations in relation to the number of correct relations allows us to see how observers use different types of description at different stages of the task. To examine the relative frequency of semantics or shapes and colors, we calculated the percentage of explanations that used each of these categories. For explanations that used both categories, the explanation was included in both percentages.

To examine the differing functional relationships between correct relations and percentage across categories, we conducted a polynomial regression. The analysis revealed a significant interaction between the quadratic term (correct relations^2) and condition (semantics vs shapes and colors), $B = 2.18$, $SE = 0.60$, $t(13) = 3.66$, $p = 0.003$. This interaction confirmed that the curvature of the relationship differed significantly by category. Specifically, the shapes and colors category exhibited a U-shaped curve, while the semantics category followed an inverted U-shaped pattern (see Figure 6). This suggests that shapes and colors are the focus when the images are highly dissimilar (below 3 correct relations) or when they are highly similar (above 5 correct relations), and semantics are the focus the rest of the time.

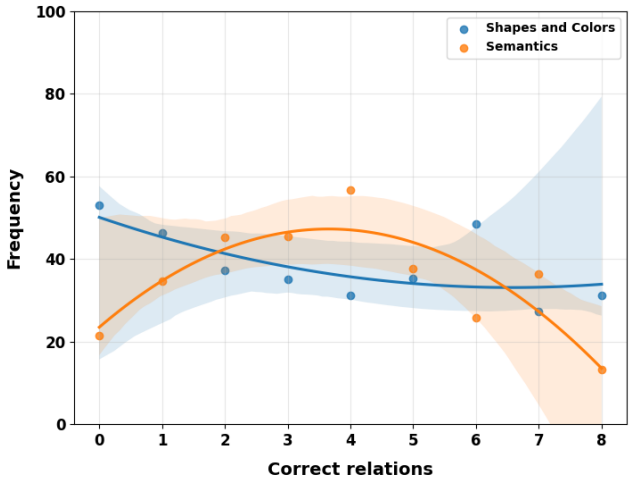


Figure 6: Quadratic model fitting Shapes & Colors, Semantics references in explanations (n = 1105).

Post-hoc

Given that people are able to identify the semantic parts of a tangram puzzle, particularly complete ones (Ji et al., 2022), we investigated why participants frequently used shapes and colors when the puzzles are highly similar or highly dissimilar to the reference image. We further coded each of the explanations that fell into "shapes and colors" category to capture whether participants were *explicitly* referring to similarities or differences between the child's puzzle and the reference images. It is worth noting that various responses implicitly made without clarifying how they related to the target puzzle (e.g., *the triangles make up the body*). To maintain consistency, responses making implicit similarity comparisons did not fall into the category.

Figure 7 shows the percentage of all explanations that used shapes and colors and referred to either similarities or differences. We focus our analysis on when shapes and colors were used more frequently than semantics. When puzzles were highly similar (6 or more correct relations), participants referred to differences more than similarities. When the puzzles were highly dissimilar (2 or fewer correct relations), participants referred to similarities more than differences. This suggests that references to shapes and colors were used to focus on the few similarities (in the dissimilar images) or few dissimilarities (in the similar images), thus revealing the scene attributes that are most salient to participants at different puzzle states.

Discussion

Insights from examining Tangram puzzle assembly can provide evidence for the types of spatial representations, abstractions, and reasoning processes that people perform.

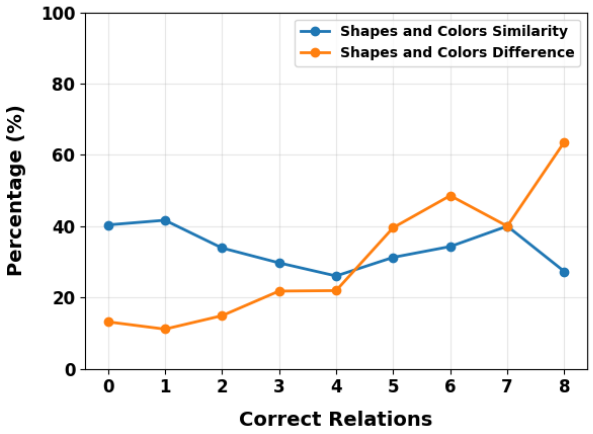


Figure 7: Percentage of explanations referring to shapes and colors similarities and differences (n = 1105).

Finding similarity in Unfinished spatial scenes

Incomplete spatial scenes affect their similarity to the target object. However, our findings showed that people can still accurately identify Tangram puzzles despite as few as 2 correct connections being shown in the observation. There are a few reasons to explain this result. In some cases, while the placement of pieces was not correct, they were approximately close to their correct position; thus, the mistakes might not have been detractors, but actually indicators of which puzzle the child was assembling. When there are relatively few similarities, people often focus on features that distinguished each scene. We found that particular structures are often identified by participants, which indicates their saliency in people's mapping process. These structures are familiar (e.g., head, ear, body) or distinct (e.g., the cat is the only puzzle with a tail) components of the scene. Our findings suggest that low-similarity in spatial scenes would be less likely to interfere with people's ability to make inferences when salient higher-order properties are present.

Referring to Puzzle parts

Previous studies have highlighted the role of high-order structures in reasoning about similarity. Hashemi and Tubau (2025) suggested that while people generally prefer global relations over object relations, the saliency of such global shapes affects whether people will utilize them to reason about visual analogies. Contributing to this research, our study showed distinct trends related to people's usage of surface-level attributes and high-order structures. When explaining how they identified a puzzle, our study found a shift in the way participants refer to parts. When similarities are rare, participants are more likely to refer to the pieces by their surface-level properties. When the scene is more similar to its target, the individual pieces form identifiable parts of the puzzle. At this stage, we found that people are more frequently referring to the puzzle more abstractly–using the semantics of the object represented in the final image. Finally, at the stage where the visual stimuli are close to complete and nearly match the reference image, people return to referring to the surface-level attributes of the individual puzzle pieces. We showed that when scenes are highly dissimilar, the lack of clear structure leads to surface features being dominant. However, when structure is present, people orient towards structural alignment as a mechanism of comparison. These findings are consistent with the structural mapping theory (Gentner & Kurtz, 2006).

Referring to Similarity and Differences

Sagi et al. (2012) suggested that within high-similarity and low-similarity spatial scenes, people are equally capable of recognizing differences between the stimuli. However, when probed to describe particular differences, they more easily identify differences in high-similarity in comparison to the counterpart. The result of our study is consistent with this research. We discovered that participants identify differences more frequently in puzzles that are closer to completion. Additionally, the differences are primarily referenced through surface-level features rather than using semantic descriptors (e.g., ear, body). Although participants can describe differences using these semantic properties, we found that they more often choose to mention their basic characteristics (i.e., shape, color, size). Our results provide further evidence for the line of research that suggested people may weigh surface-level features more heavily than structure when identifying differences (Medin et al., 1990).

Conclusion

In this paper, we investigate how people describe spatial scenes at different levels of similarity. Presented with an incomplete puzzle, research participants were asked to identify the target scene, explain their reasoning process, and give a rating of their confidence in the inference. By analyzing their explanations, we found that higher-level properties are used when the scene pairs contained some similarity. However, highly similar or dissimilar scenes promote the use of surface features. Finally, when scenes are highly similar, surface-level differences become more salient in spatial comparison. These findings suggest that in assessing spatial scenes, the degree of similarity affects how people perform spatial reasoning.

Acknowledgment

Our study used generative AI for programming assistance, including data cleaning, analysis, and visualization. We greatly appreciate Anara Adylbekova and HiuChun Wong for their assistance with the coding of open-ended responses. JRW and HAV were supported by the National Science Foundation under Grant No.2338148. This research was supported by the U.S. Naval Research Laboratory (LH). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy.

References

Gattis, M. (2002). Structure mapping in spatial reasoning. *Cognitive Development*, *17*(2), 1157–1183.

Gentner, D., & Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. *Cognitive science*, *30*(4), 609–642.

Harris, D. (2023). Spatial reasoning in context: Bridging cognitive and educational perspectives of spatial-mathematics relations. *Frontiers in Education*, *8*, 1302099.

Hashemi, A., & Tubau, E. (2025). Global relations versus object relations in visual analogies. *Thinking & Reasoning*, *31*(1), 109–135.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *Proceedings of the 2022 conference on empirical methods in natural language processing*, 582–601.

Langer, A., Wilson, J. R., Howard, L., & Marshall, P. J. (2026). The influence of individual characteristics on children's learning with a social robot versus a human. *Scientific Reports*, *16*.

Lee, J., Lee, J. O., & Collins, D. (2009). Enhancing children's spatial sense using tangrams. *Childhood Education*, *86*(2), 92–94.

Li, J., Zhang, X., Zheng, H., Lu, Q., & Fan, G. (2019). Global processing styles facilitate the discovery of structural similarity. *Psychological Reports*, *122*(5), 1755–1765.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431–467.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.

Morita, J., Nagai, Y., Taura, T., et al. (2007). Learning support for composition ability: Surface and structural similarity. *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007*, 271–272.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive science*, *36*(6), 1019–1050.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational psychology review*, *14*(1), 47–69.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*(4), 373–421.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Wilson, J. R., Rabkina, I., Roberts, M., & Hiatt, L. M. (2025). ToMCAT: Benchmark for socially assistive robots with theory of mind of children assembling tangram puzzles. *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2381–2388.