Now They See It, Now They Don't: Multimodal Reward Models Exhibit Unreliability in Physical World Constraints

Sadaf Ghaffari and Nikhil Krishnaswamy
Situated Grounding and Natural Language (SIGNAL) Lab
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
{sadafgh,nkrishna}@colostate.edu

Abstract

Generative AI systems, especially those driven by autoregressive and diffusion-based models, are known to struggle with spatial reasoning. As such, it becomes critical to understand how humans regard those failure modes. In this paper, we examine how humans judge different types of errors in images generated by a text-to-image model. We curated prompts that described common household objects with variance in number, spatial relations, and orientations, and generated a variety of images using each prompt. Humans observed pairs of images generated using the same prompt and answered a set of systematic questions about each image. Survey results showed that incorrect spatial *orientation* regularly emerges as a reason that the generated images do not accurately represent the prompt. We further investigated how RLHF-based multimodal reward models score prompt-image alignment over the same data, and whether they can reliably distinguish the better image in a pairwise setting, as humans do. We find that even though a general cross-task reward model may output alignment scores that accord with those of humans, its reasoning traces are flawed with respect to spatial orientational and relational indicators—the very factors that human annotators rated as the most consequential errors in generated images. Our results show that human annotators regard spatial reasoning errors as highly impactful on the correctness of generated images, and undermine the reliability of multimodal reward model scores as a baseline for evaluating image quality.

1 Introduction

There is increasing interest in deploying sophisticated AI in physically-grounded scenarios where spatial reasoning is a core requirement, such as collaboration with humans, graphic design or autonomous navigation. A core assumption in such collaborations is that the AI system will be able to engage in the same kind of situated, contextually-grounded spatial reasoning that comes naturally to humans. However, this remains to yet be borne out in reality.

Figure 1: DALL-E 3 generated image given the prompt "an upside-down bowl in the colander".

For instance, text-to-image generative models, including both autoregressive and diffusion-based methods, have seen rapid progress in recent years (Ramesh et al., 2021; Ding et al., 2021; Esser et al., 2021; Yu et al., 2022; Saharia et al., 2022; Ramesh et al., 2022). Given a prompt, these models can generate high-fidelity images. But despite this improvement, they still lag considerably in their ability to capture realistic physical and spatial relationships that are obvious to humans. For example, Fig. 1 shows a DALL-E 3 generated image the prompt "an upside-down bowl in the colander". In the generated image, the meaning of the configurational cue "upside down" has been ignored or misinterpreted.

In order to bridge this gap, it is critical to understand not only the kinds of errors that generative AI systems make in the domain of situated reasoning, but also how human observers regard or respond to those errors, and to enable some level of theoretically-grounded explanation as to *why* generative AI systems make these errors. In this paper, we conduct an examination of how humans rate the severity of different types of generation er-

rors made by a text-to-image generative model, and establish that errors in spatial reasoning, and particularly those pertaining to object orientation, have a strong effect on human ratings of image correctness. Further, we assess the alignment of independent multimodal reward model's assessments of image correctness with human judgments, and show that even if raw scalar rewards assigned to an image by a reward model roughly aligns with scores give by humans, the underlying reasoning generated by the reward model may contain significant errors, particularly in the cases of images that do not adequately match the prompts. Our results show that not only do humans consider spatial reasoning errors significantly impactful on image correctness, but modern multimodal reward models are sensitive to the same issues with spatial reasoning and configuration seen in image generators. Our data and code is available at: https://github.com/csu-signal/rel_multrewardmodel

2 Related Work

Spatial reasoning as demonstrated by humans is a sophisticated and multifaceted capability (Stock et al., 2022). Spatial reasoning systems emerge in infancy and then remain roughly constant over a human lifetime (Shusterman and Spelke, 2005). On the other hand, the development of spatial reasoning capabilities in young children is thought to be a crucial prerequisite for other types of reasoning such as mathematical reasoning (Davis et al., 2015), linguistic reasoning (Simms and Gentner, 2019), and cultural reasoning (Li and Gleitman, 2002; Li et al., 2011).

Computer systems, up to and including modern neural language models, do not undergo the same developmental phases as humans, and thus they encode spatial information differently. There have been a number of prior approaches to this problem in particular. Pre-neural language modeling, these ranged from broad formal and cognitive-based approaches (Bateman et al., 2010; Moratz et al., 2001), to **qualitative spatial calculi**, which render continuous point and space distributions in discrete categorical terms more amenable to computational processing while retaining human interpretability (Randell et al., 1992; Moratz et al., 2002; Scivos and Nebel, 2004; Moratz and Ragni, 2008; Albath et al., 2010).

Pustejovsky (2013) forwarded the related notion of object *habitats* or conditioning environments that enable or disable afforded object behaviors (Gibson, 1977). Pustejovsky and Krishnaswamy (2016) extended this to a theoretically-grounded encoding in which a habitat is defined as a precondition $\mathcal{C}$ wherein if a program π is enacted, the deterministic result $\mathcal{R}$ is realized ($\mathcal{C} \rightarrow [\pi]\mathcal{R}$). The habitat also encodes the notion that certain objects have intrinsic surfaces and directedness (e.g., the top of a bottle). Empirically, these preconditions are frequently encapsulated within the orientation of the object, and with *canonical* vs. *non-canonical* orientations (e.g. an upright vs. an inverted bottle or one on its side (Krishnaswamy and Pustejovsky, 2022)). Ghaffari and Krishnaswamy (2024a) and Ghaffari and Krishnaswamy (2024b) showed how a large cross section of multimodal LLMs struggled to reason correctly about habitats and orientation, even when they frequently succeeded in their reasoning about related concepts like affordances, and hypothesize that this is due in part to a bias toward canonical object orientation in training data.

Likewise, recent approaches have showcased the struggles modern language models (including multimodal variants) continue to have with spatial reasoning and related tasks. Chen et al. (2024) and Cheng et al. (2024) highlight how a lack of 3D knowledge in training limits LLMs' and VLMs' spatial reasoning abilities, and Saad et al. (2025) show that underlying spatial reasoning limitations in SOTA LLMs limit the effectiveness of Gricean understanding and other tasks. Conwell et al. (2024) shows, among other things, that text-to-image generative models are susceptible to hallucinating extra objects when the prompt describes plurals. Conwell and Ullman (2023) also emphasizes that text-guided visual generation models such as DALL-E 2 fail at correctly representing simple relations involving objects and agents.

Building on RLHF, reward modeling has expanded from text to visual generation. ImageReward (Xu et al., 2023), HPSv2 and HPSv3 (Wu et al., 2023; Lin et al., 2024), and PickScore (Kirstain et al., 2023) learn human preferences for text-to-image generation, improving correlation with human judgments and guiding diffusion models beyond CLIP-based evaluations. Despite progress, multimodal reward models remain task-specific and lack a rigorous evaluation. In this work, we stress-test multimodal reward models' ability to evaluate text-image alignment on a spatial reasoning task.

2.1 Human Preference Alignment in Text-to-Visual Generative Models

Given this apparent difference in human vs. generative AI spatial reasoning abilities, we can examine how these differences affect humans' and AIs' assessments of spatial reasoning.

Various evaluation protocols (Madhyastha et al., 2019; Yu et al., 2022; Hessel et al., 2021) have been proposed to measure alignment between images and text. Some prior works (Radford et al., 2021; Yu et al., 2022) use the alignment score of image and text embeddings determined by pretrained multi-modal models, such as CLIP (Radford et al., 2021). However, because scores from pre-trained models tend to be misaligned with human intent, researchers have also introduced human evaluation (Saharia et al., 2022).

ImageReward (Xu et al., 2023), trained and evaluated on 137k pairs of expert comparisons in total, is one example of generative text to visual models using human feedback to align multi-modal text-to-image models with human preferences. Although such approaches seem to be more effective than previous image-text alignment evaluation methods such as CLIPScore (Hessel et al., 2021), collecting human preferences is a costly and challenging process. Like other reward models (e.g., PickScore (Kirstain et al., 2023) and HP3 (Wu et al., 2023)), ImageReward is also designed for only image generation tasks. This limitation has motivated the most recent adaptable and generalizable reward models such as UnifiedReward-Think (Wang et al., 2025) with the capability of being utilized across multiple tasks, e.g., generation and understanding.

3 Methodology

3.1 Prompt Selection and Data Generation

We follow the prompting scheme described in Ghaffari and Krishnaswamy (2024b). Each prompt describes a common object or objects in a configuration that primarily concerns spatial relation(s), cardinality (number) and orientation. The list of objects mentioned in the prompt is given in the appendix. The number of objects described may vary from one to five. Spatial relations include *in*, *inside*, *on*, *under*, or *on top of*. We then create three variants of each prompt, each of which describes a different orientation of the object: (i) no stated orientation, (ii) a canonical orientation (e.g., an *upright* cup), and (iii) a non-canonical orientation (e.g., an *inverted* or *upside down* cup). This process resulted in a total of 99 distinct prompts.

We chose DALL-E 3 API (Betker et al., 2023) as the text to visual image generator. DALL-E 3 represents one of the most recent OpenAI text-to-image models not native to the proprietary ChatGPT platform. Though sophisticated, it generates a sufficient density and diversity of errors for meaningful analysis, thus satisfying our research goals, not of improving text-to-visual generative model performance in spatial reasoning tasks (which is not possible for 3rd party researchers with closed models), but rather of analyzing errors made and human assessments of those selfsame errors. This choice also gives us a unique opportunity to investigate a range of multimodal reward models designed for text-to-visual generation and understanding tasks. For each prompt, we generated 5 images using the default settings, resulting in 495 total generated images.

3.2 Human Survey and Evaluation

Using the Prolific platform, we recruited a total of 199 participants to take a survey in which each task presented them with two images generated using the same prompt. Each sample, consisting of a prompt and an image pair, was presented to two participants. Participants were asked to answer a set of systematic questions:

1. How many objects were depicted in each image (excluding surface and background objects (Talmy, 1975; Rubin, 2001a,b))?;
2. How well did each image in the pairwise comparison align with the given prompt (1 being no alignment and 10 being perfect alignment)?;
3. Which of the two images better represented the contents of the prompt (in preference alignment terminology, a "winner", although ties were allowed if both images were equally representative of the prompt)?; and
4. What factors made the generated images unrepresentative of the prompts (a multi-answer question allowing for incorrect number of objects, incorrect spatial relation or orientation, hallucinated objects, or none)?

Collectively, these questions address the qualities of number, orientation, and spatial relation as they relate to the generated images' overall representativeness of the descriptions they purport to

represent, as well as distinct factors that detract from that representativeness. Fig. 13 shows a sample comparison presented in the survey. Specific image placements (left vs. right) were randomized across individual survey instances to control for positional bias. A research protocol was approved by the local Institutional Review Board for this study. Participants accepted and consented to the study via the Prolific interface. They were paid $10 per hour, and were taken from the UK and USA and were required to speak English. An image of the survey is given in the appendix.

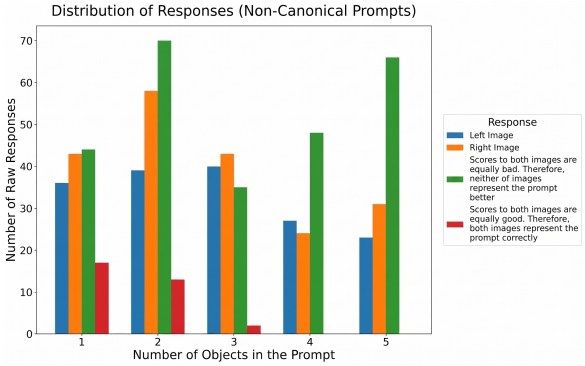

Figure 2: Participants response distribution reporting tie and preferred images depending on the pair prompt-images.

Participants were allowed to report a tie when assessing paired images, as shown in Fig. 2. This facilitated a focus on the objective and logical correctness properties of the images with respect to factors like number, orientation, and spatial relations, rather than aesthetic factors (illumination, patterns, etc.). Furthermore, given the 5 images generated with each prompt, and the full Cartesian pairings thereof that were presented to participants, ties allowed them to indicate where both images represented the prompt equally well or badly. The remaining subset of participants responses follows the strict Bradley-Terry pairing scheme. Pairing the images as we do here replicates the Bradley-Terry (BT) modelization assumption of paired preferences that is common in modern language model alignment (Bradley and Terry, 1952; Ji et al., 2023), and thus facilitates BT modeling of rewards. Additionally, it represents a well-known compromise between the noise inherent in gathering scores of single samples in isolation (Sun et al., 2024) and the complexity of collecting ranked orderings of samples a la the Plackett-Luce model (Plackett, 1975; Luce et al., 1959). We analyze the participants observations and responses in Sec. 4.1.

3.3 Multimodal Reward Model Evaluation

Reward models play a key role in developing and improving LLMs (Xu et al., 2025; Lambert et al., 2025). They can support scalable model performance evaluation and identifying systematic weaknesses (Zheng et al., 2023). They can also serve as a measure of data quality, which is essential for building synthetic-data pipelines (Wang et al., 2023). Motivated by the success of RLHF in language modeling (Stiennon et al., 2020; Nakano et al., 2021; Ouyang et al., 2022), there is growing interest in utilizing reward models to align text-guided vision-generation models with human preferences. Current reward modeling approaches in text to visual generation models can be divided into three types:

1. Fixed, discriminative scorers (e.g., CLIP (Radford et al., 2021), BLIP (Li et al., 2022), CLIP-Score Hessel et al., 2021) that assign a single global score to each prompt–image pair, with limited sensitivity to prompt-specific evaluation criteria;

2. Models based on BT pairwise preference modeling that learn rewards from relative comparisons and induce preferences via score difference (e.g. HPSv3 (Ma et al., 2025));

3. Approaches such as UnifiedReward-Think (Wang et al., 2025) that leverage generative VLMs to produce richer textual judgment to incorporate Chain-of-Thought (CoT) reasoning into reinforcement fine-tuning.

From the wide spectrum of available multimodal reward models in the text-to-visuals generation, we choose two models that are based on RLHF—UnifiedReward-Think and ImageReward (Xu et al., 2023)—to assess their judgment performance on our task. ImageReward is based on point scoring approach, while UnifiedReward-Think is based on both point scoring and pair ranking approach and provides a CoT reasoning justification to the score. The aim of long CoT incorporation in the UnifiedReward-Think reward model was to enhance its robustness and reliability (Wang et al., 2025). We analyze the outputs of both reward models in Sec. 4.2.

4 Results

4.1 Participants' Responses

Fig. 3 shows the average reported count of objects in generated images plotted vs. the number of objects described in the prompt (results of survey question 1). There is a clear monotonic increase in the number of reported objects in the generated images as the number of objects in the prompt increases, and the model is more prone to generate extraneous images as more objects are described in the prompt, which reinforces an observation by Conwell et al. (2024).

However, if we break the responses down by images where the object(s) in the prompt were described using no explicit orientation, an explicitly canonical orientation, or an explicitly non-canonical orientation, a further interesting observation emerges. As shown in Fig. 4, the prompts containing non-canonical orientations are less aligned with generated images than the remaining prompts (results of survey question 2). As the number of objects in the prompt increases, the human-assessed alignment of the image with the prompt decreases, which reflects the effect of extraneous objects appearing in the image (Fig. 3). But when the prompts are disaggregated by orientation type in the prompt, we see that canonical orientations' alignment scores decline rapidly as the number of objects in the prompt increases, which is inversely proportional to the average number of extraneous objects appearing in the generated image, but non-canonical orientations' scores have a much lower starting point, even when the number of objects in the prompt is only one.

Indeed, Fig. 5 shows that incorrect object orientation and incorrect number of objects are the top two frequent errors present in all prompts (results of survey question 4), with incorrect object orientation reported as a factor about 17 percentage points more than number of objects. Fig. 6 shows that for *all* prompts, the frequency of object count errors dramatically increases as the number of objects in the prompt increases beyond 3 (cf. Conwell et al. (2024)), while the frequency of orientation errors is much more consistent across the number of objects in the prompt.

4.2 Multimodal Reward Models

Similar to the human survey to assess alignment between prompt and paired images, we assessed image-to-prompt alignment with two RLHF based multimodal reward models. We chose the CLIP-based evaluator version of ImageReward and the Qwen3-VL-7B-based evaluator version of UnifiedReward-Think. We selected the UnifiedReward-Think reward model to leverage

Figure 3: Participants reported object counts in generated images versus number of objects in prompt.

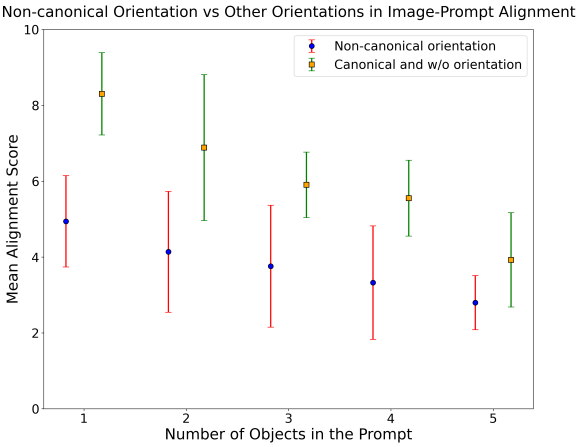

Figure 4: Participant-reported image-prompt alignment scores for prompts describing non-canonical orientations vs. prompts that have no-orientation and stated canonical orientation.

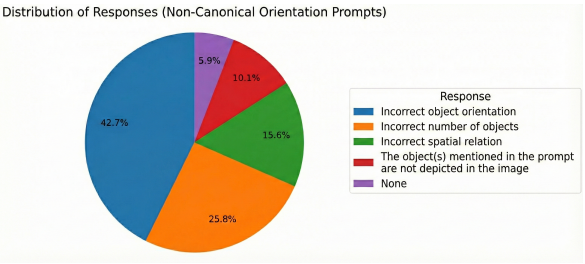

Figure 5: Distribution of different types of errors in generated images from prompts with non-canonical orientation.

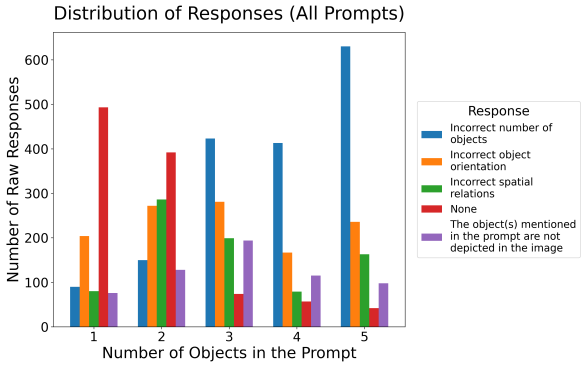

Figure 6: Distribution of different types of errors present in generated images from all prompts.

its generalizable capability for both understanding and generation tasks.

While ImageReward returns a simple scalar score, UnifiedReward-Think model provides pointwise and pairwise ranking (scores from 1 to 10), along with chain-of-thought generated rationales across multiple dimensions including but not limited to the following items: *Semantic consistency*, *Aesthetics*, *Authenticity*, *Object coherence*, and *Composition*. For all multimodal reward models, we use default sampling parameters from the official implementations; in most cases, the sampling temperature is 1.0.

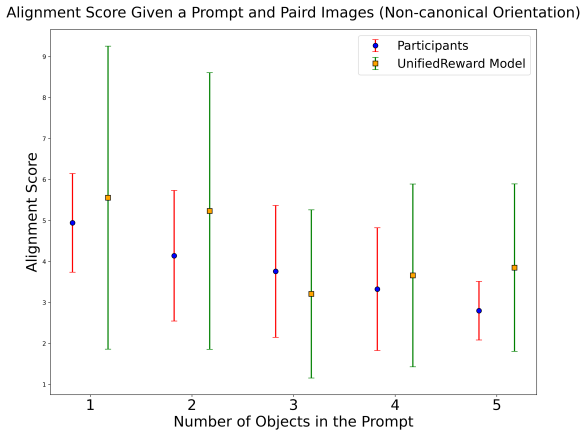

Figure 7: Alignment scores comparison between UnifiedReward-Think reward model and survey participants.

Fig. 7 compares the image-prompt alignment scores returned by UnifiedReward-Think to those of the human annotators for objects in non-canonical orientations, plotted against the number of objects described in the prompt. We considered the *semantic consistency* dimension, which is consistently present across all the prompts in the model output, unlike other dimensions. The distribution of scores for UnifiedReward-Think model is roughly aligned with the scores reported by human evaluators.

Then, for the set of prompts that described objects in non-canonical orientation, we took the image pairs where the two participant who took the DALL-E 3 survey agreed on which specific image was more representative of the prompt. Only samples where a specific images was agreed to be the winner was considered ("left image" or "right image" in Fig. 2, resolved to the specific image name, as image placement was randomized across individual surveys). Five prompts were not included in this experiment, as the images generated from them were scored as ties instead of indicating a distinct winner. For this subset of pairs, we compare the multimodal reward models' assessments of which image is better fit for the prompt, and compare them to the human responses. The winner of the image pair according to UnifiedReward-Think was chosen based on the sum of scores across all dimensions given by the model. In the case that the ImageReward model did not explicitly state which image is better, we infer the preferences from the scores assigned by ImageReward to each of the paired images.

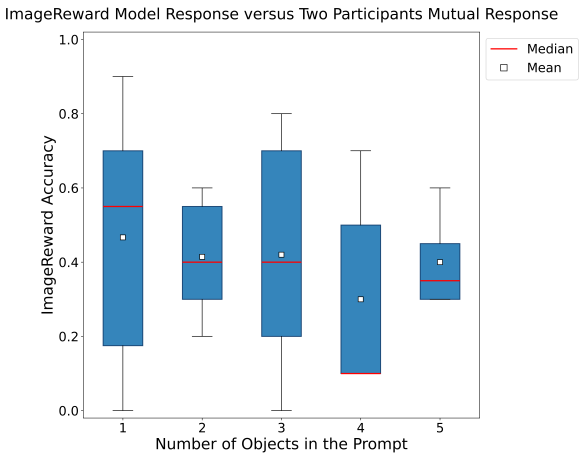

Figure 8: ImageReward accuracy in matching responses from two survey participants.

Using the human responses as ground truth, Figs. 8 and 9 show the respective accuracy of ImageReward and UnifiedReward-Think responses, respectively, plotted against the number of objects described in the prompt. The average accuracy across all prompts for ImageReward is consistently lower than the average accuracy of UnifiedReward-Think model across the prompts with varied numbers of objects.

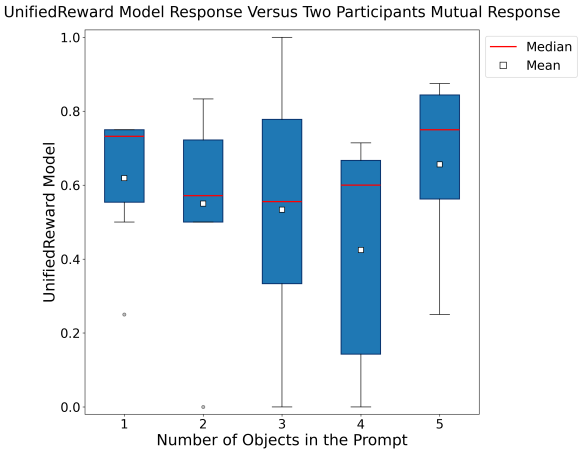

Figure 9: UnifiedReward-Think accuracy in matching responses from two survey participants.

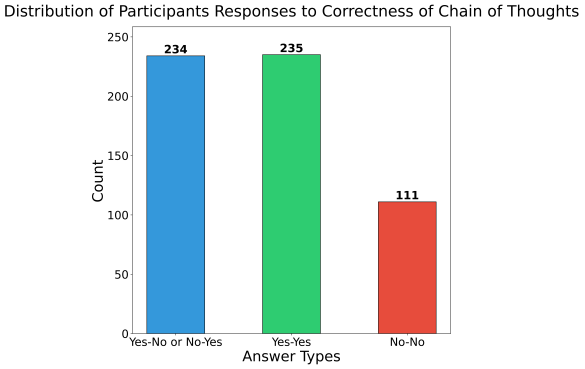

Figure 10: Participant judgment of UnifiedReward-Think reasoning traces in secondary survey.

4.3 Divergences Between Humans and Reward Models

Even if the high-level scalar assessments of a reward model broadly align with those of humans, the underlying reasoning for the assessment may diverge, thus calling into question the validity of the model assessments. This is a well-known problem in domains like scientific review and education (Alper et al., 2024; Joachim et al., 2025; Wetzler et al., 2025). Here we perform a similar examination in the fundamental capacity of spatial reasoning.

We therefore launched a secondary survey among 99 Prolific participants to validate the correctness of reasoning traces provided by the UnifiedReward-Think model. Similar to the initial main survey described in Sec. 3, participants from UK and USA were recruited under the same human subjects research protocol and consent structure. Participants in this survey were compensated $12 per hour. In the study, participants were provided with the prompt-image pairs along with UnifiedReward-Think's reasoning about the image's representativeness of the prompt for each image. They were asked to answer "Does the reward model reason correctly overall?" for each image and its associated CoT rationale. For each prompt and pair of images, three participant responses were collected, each of which may constitute a pair *yes-yes*, *no-no* or *yes-no/no-yes*. An image of this survey is given in the appendix.

Fig. 10 presents statistics over the responses of the two participants who most consistently agreed that *both* reasoning traces in the pair were correct, across all responses for all prompts that described objects in non-canonical orientation. As shown, 345 responses are either *no-no* or one *yes* and one *no* (in other words, at least two participants did not agree that both reasoning traces were correct), while there are only 235 *yes-yes* responses that confirm the correctness of the reasoning traces. Overall, this result indicates that in more than 50% of reasoning traces by UnifiedReward-Think, at least one of the reasoning traces for the paired images is considered incorrect, and close to 20% are consistently incorrect.

We additionally investigated the types of mistakes in the reasoning traces of the UnifiedReward-Think model for prompts involving objects in non-canonical configurations. Given the DALL-E3 generated image, its associated prompt and UnifiedReward-Think model's CoT rationale, we utilized the vision-language model, *Gemini-3-flash-preview*, as a judge to identify which categories were incorrectly stated in the reasoning traces and to quantify their frequency of occurrence. An example of the question we prompted for object orientation in *Gemini-3-flash-preview* was *"Look at this image. Does it show the upside-down bowl? Please answer with 'Yes' or 'No' and provide a short, specific explanation."*. The similar approach was taken for the case of spatial relation and number of objects. In other words, object configuration, spatial relation and number of objects stated in the reasoning traces were extracted to check their correctness against the prompt and generated image.

We observed that similar to DALL-E3 generated images, the CoT rationale generated by the UnifiedReward-Think model contain numerous errors.[1] Examples of such errors include objects

[1]The outputs generated by the *Gemini-3-flash-preview* model were subsequently reviewed by an annotator to remove potential model-induced errors, ensuring the correctness of

lying on their sides being incorrectly interpreted as upside down, and object quantities exceeding three being miscounted. Fig. 11 illustrates the accumulation of errors and their frequencies across all reasoning traces, highlighting the two most frequent mistake types—spatial orientation and object count—in the UnifiedReward-Think model's reasoning traces.

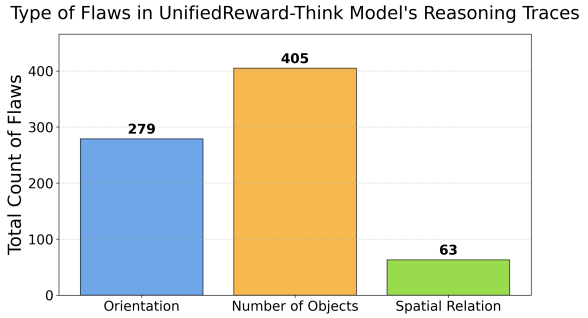

Figure 11: Frequency and type of mistakes present in UnifiedReward-Think model's reasoning traces.

5 Discussion

In the first phase of our study, incorrect object orientation emerged as a critical factor affecting human participants' assessment of image alignment with prompt. Incorrect number of objects was a secondary factor, aligning with findings by Conwell et al. (2024). Ostensibly, human evaluations may consider images generated from prompt describing n objects that contain $> n$ objects as *technically* containing *at least* n objects, and thus satisfying the n objects constraint. In model-theoretic terms, a scene containing 3 objects is a semantic consequence of ($\models$) the scene containing 5 objects, even though this is pragmatically misleading. However, orientational cues are mutually exclusive; for instance, an object cannot be both upright or on its side *and* inverted at the same time. Human evaluators strongly rate such discrepancies as a factor in making the generated image unrepresentative of the prompt. This is particularly true of images generated from a prompt containing an object described in a non-canonical orientation, and such spatial reasoning errors correlate with lower overall alignment score of the image to the prompt.

When we consider the performance of multimodal reward models in assessing the text-to-image outputs, we observe that although the image-prompt alignment scores returned by a modern model such as UnifiedReward-Think may roughly

(a) *The image perfectly depicts an inverted strainer basket on a table. The object is clearly a colander, it is placed upside down, and it is on a wooden table, matching all elements of the caption.*

(b) *This image also perfectly matches the caption. It shows a strainer basket, it is inverted, and it is resting on a wooden table surface.*

Figure 12: UnifiedReward-Think reasoning traces provided for semantic consistency for generated paired images and prompt "An inverted strainer basket on a table."

align with those assigned by human evaluators, the reasoning given for the score is significantly flawed, especially when the objects in the prompt are described in non-canonical orientations. Fig. 12 shows one such indicative example of images generated from the prompt "an inverted strainer basket on the table". In both cases, the reward model claims that the image aligns very well with the prompt and proceeds to conflate the orientation of the objects (on their sides) with being inverted or upside down. Thus our qualitative and quantitative results show that even if the scalar reward aligns with that assigned by humans, the generated reasoning traces might be erroneous in the presence of images that are misaligned with the prompt.

6 Conclusion

In this paper, we have examined both humans' and multimodal reward models' assessments of rea-

Fig. 11.

soning errors in generative text-to-image models. Our results show that spatial orientation errors are particularly impactful on human ratings. We also show that multimodal reward models are sensitive to the same types of errors, and that even though UnifiedReward-Think more frequently agreed with human assessments of image-prompt alignment than ImageReward, the underlying reasoning that led to the assessment frequently contains significant flaws, especially when reasoning about images in which the image generator include spatial reasoning errors about orientation.

The reliability and robustness of multimodal reward models for assessing text-image alignments, like that of generative text-to-image models themselves, appears to be strongly challenged by spatial reasoning tasks, especially those related to reasoning about objects in non-canonical configurations.

Limitations

This study did not assess the performance of newer text-to-image models such as the GPT-Image or Nano Banana families due to access constraints imposed by the Gemini and OpenAI APIs and web interfaces. While newer closed models *possibly* output fewer errors for prompts that are structurally more simple, they still are not error-free. Our experiment using Gemini-3-flash-preview, a newer closed vision-language model, also demonstrates that same spatial reasoning errors are still present despite of recent advances. At the end, we must reiterate that our research question in this paper was directed not at how to improve text-to-image model performance but rather toward how humans perceive different kinds of generation, especially spatial reasoning errors, and how human and AI assessments of text-to-image outputs align or differ across a variety of examples, including instances of both text-image alignment and misalignment.

Acknowledgments

This material is based in part upon work supported by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program and by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO) Knowledge Systems program. Views expressed herein do not reflect the policy or position of the Department of Health and Human Services, the Department of Defense, or the U.S. Government. We would also like to thank the anonymous reviewers whose feedback helped improve the final copy of this manuscript. Any remaining errors are the responsibility of the authors.

References

Julia Albath, Jennifer L Leopold, Chaman L Sabharwal, and Anne M Maglia. 2010. Rcc-3d: Qualitative spatial reasoning in 3d. In *CAINE*, pages 74–79.

Ayfer Alper and 1 others. 2024. Evaluating the evaluators: A comparative study of ai and teacher assessments in higher education. *Digital Education Review*, (45):124–140.

John A Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Colin Conwell, Rupert Tawiah-Quashie, and Tomer Ullman. 2024. Relations, negations, and numbers: Looking for logic in generative text-to-image models. *arXiv preprint arXiv:2411.17066*.

Colin Conwell and Tomer Ullman. 2023. A comprehensive benchmark of human-like relational reasoning for text-to-image foundation models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Brent Davis and 1 others. 2015. *Spatial reasoning in the early years*. Taylor & Francis.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and 1 others. 2021. Cogview:

Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2024a. Exploring failure cases in multimodal reasoning about physical dynamics. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 105–114.

Sadaf Ghaffari and Nikhil Krishnaswamy. 2024b. Large language models are challenged by habitat-centered reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13047–13059, Miami, Florida, USA. Association for Computational Linguistics.

James J Gibson. 1977. The theory of affordances. *Perceiving, acting, and knowing: toward an ecological psychology*, pages pp–67.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Michael V Joachim, Thomas B Dodson, and Amir Laviv. 2025. How artificial intelligence differs from humans in peer review. *Journal of Oral and Maxillofacial Surgery*, 83(8):1040–1050.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nikhil Krishnaswamy and James Pustejovsky. 2022. Affordance embeddings for situated language understanding. *Frontiers in artificial intelligence*, 5:774752.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Peggy Li, Linda Abarbanell, Lila Gleitman, and Anna Papafragou. 2011. Spatial reasoning in tenejapan mayans. *cognition*, 120(1):33–53.

Peggy Li and Lila Gleitman. 2002. Turning the tables: Language and spatial reasoning. *Cognition*, 83(3):265–294.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

R Duncan Luce and 1 others. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. 2025. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2019. Vifidel: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550.

Reinhard Moratz, Kerstin Fischer, and Thora Tenbrink. 2001. Cognitive modeling of spatial reference for human-robot interaction. *International Journal on Artificial Intelligence Tools*, 10(04):589–611.

Reinhard Moratz, Bernhard Nebel, and Christian Freksa. 2002. Qualitative spatial reasoning about relative position: The tradeoff between strong formal properties and successful reasoning about route graphs. In *International Conference on Spatial Cognition*, pages 385–400. Springer.

Reinhard Moratz and Marco Ragni. 2008. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1):75–98.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10.

James Pustejovsky and Nikhil Krishnaswamy. 2016. Voxml: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

David A Randell, Zhan Cui, Anthony G Cohn, and 1 others. 1992. A spatial logic based on regions and connection. *KR*, 92(165-176):40–40.

Edgar Rubin. 2001a. Figure and ground. *Visual perception*, pages 225–229.

Nava Rubin. 2001b. Figure and ground in the brain. *Nature neuroscience*, 4(9):857–858.

Fardin Saad, Pradeep K Murukannaiah, and Munindar P Singh. 2025. Gricean norms as a basis for effective collaboration. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1812–1820.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Alexander Scivos and Bernhard Nebel. 2004. The finest of its class: The natural point-based ternary calculus for qualitative spatial reasoning. In *International Conference on Spatial Cognition*, pages 283–303. Springer.

Anna Shusterman and Elizabeth Spelke. 2005. Language and the development of spatial reasoning. *The innate mind: Structure and contents*, pages 89–106.

Nina K Simms and Dedre Gentner. 2019. Finding the middle: Spatial language and spatial reasoning. *Cognitive Development*, 50:177–194.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Kristin Stock, Christopher B Jones, and Thora Tenbrink. 2022. Speaking of location: a review of spatial language research. *Spatial Cognition & Computation*, 22(3-4):185–224.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. 2024. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*.

Leonard Talmy. 1975. Figure and ground in complex sentences. In *Annual meeting of the Berkeley linguistics society*, pages 419–430.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Elizabeth L Wetzler, Kenneth S Cassidy, Margaret J Jones, Chelsea R Frazier, Nickalous A Korbut, Chelsea M Sims, Shari S Bowen, and Michael Wood. 2025. Grading the graders: Comparing generative ai and human assessment in essay evaluation. *Teaching of Psychology*, 52(3):298–304.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

The list of objects mentioned in our prompts used in DALL-E3 model is: *bench, bowl, bucket, can, pitcher, chair, colander, cup, desk, glass, gravy boat, jug, laundry basket, measuring cup, mug, pan, plate, pot, saucepan, strainer basket, luggage, vase, cardboard box, sofa.*

Fig. 13 shows the initial survey given to Prolific participants to assess the correctness of generated images with respect to the prompt, and the categorization of error types, as discussed in Sec. 3..

Fig. 14 shows the survey given to Prolific participants to assess the quality of UnifiedReward-Think's reasoning about image alignment with prompt, discussed in Sec. 4.3.

A sample of chain of thought traces generated by the Unified Reward-Think model is presented in Fig. 15. The reasoning associated with Image 1 in Fig. 15 is illustrative of how the model wrongly reasons about semantic consistency while generating a point-wise score of 9 out of 10 at the same time.

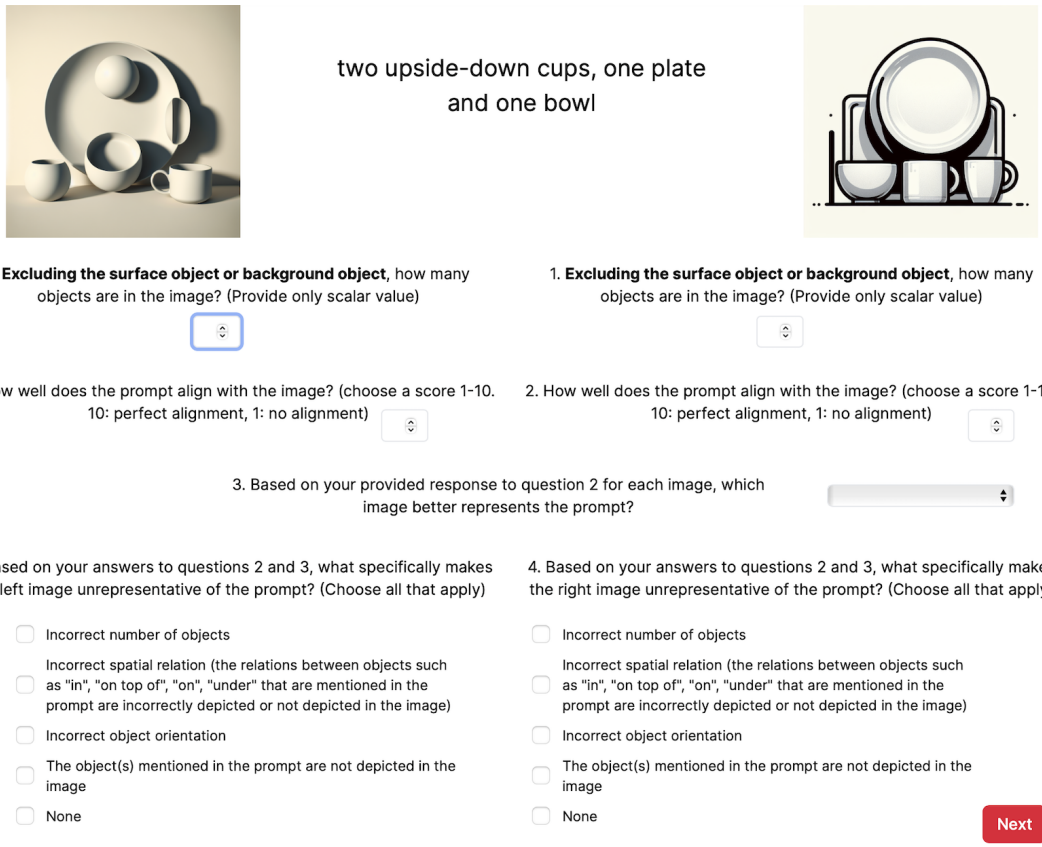

Figure 13: Sample image pair with prompt and questions in Prolific participant survey.

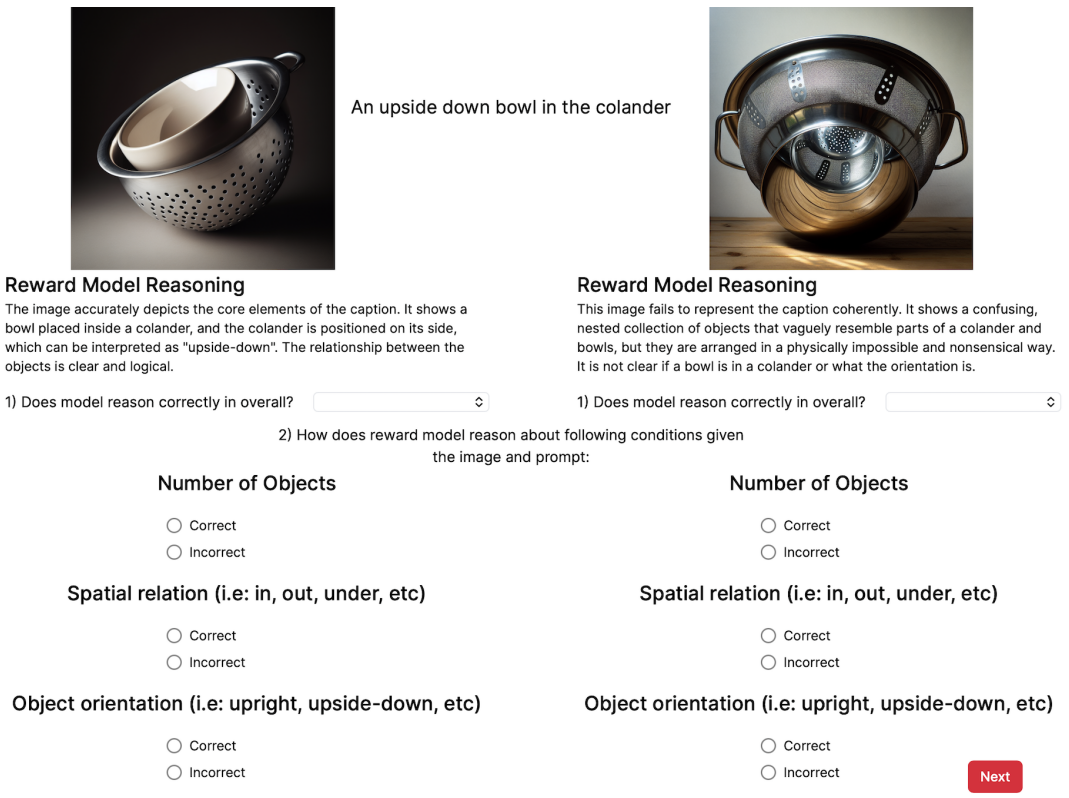

Figure 14: Survey given to Prolific participants to assess quality of UnifiedReward-Think reasoning.

Image 1	Image 2
1. **Semantic consistency (9/10)** - The image accurately depicts the core elements of the caption. It shows a bowl placed inside a colander, and the colander is positioned on its side, which can be interpreted as "upside-down". The relationship between the objects is clear and logical.	1. **Semantic consistency (2/10)** - This image fails to represent the caption coherently. It shows a confusing, nested collection of objects that vaguely resemble parts of a colander and bowls, but they are arranged in a physically impossible and nonsensical way. It is not clear if a bowl is in a colander or what the orientation is.
2. **Aesthetics (9/10)** - The image has excellent aesthetic qualities. The composition is balanced and focused, the lighting is dramatic and creates a professional, studio-like feel, and the contrast between the smooth ceramic and the brushed metal is visually appealing.	2. **Aesthetics (4/10)** - The aesthetics are poor due to the chaotic and confusing composition. While the lighting is dramatic, the jumbled nature of the objects makes the image visually jarring and unpleasant to look at. The mix of textures is interesting but poorly intograted.
3. **Authenticity (9/10)** - The image achieves a high degree of realism. The textures of the materials, the reflections on the metal, and the soft shadows all contribute to a photorealistic quality. It looks like a real product photograph.	3. **Authenticity (3/10)** - The image lacks authenticity. The objects are surreal and do not resemble real-world items. The way the mesh, solid metal, and wood-like textures are combined is physically implausible and clearly artificial.
4. **Object Coherence (10/10)** - The objects are perfectly coherent and well-formed. The colander is arl, believable forms.	4. **Object Coherence (1/10)** - The image completely lacks object coherence. Iljects are fragmented and do not exist as distinct, understandable entities.

Figure 15: Sample of complete chain of thoughts generated by UnifiedReward-Think model.