Visuospatial Perspective Taking in Multimodal Language Models

Jonathan Prunty[*,1], **Seraphina Zhang**[1,2], **Patrick Quinn**[1], **Jianxun Lian**[3], **Xing Xie**[3] **& Lucy Cheke**[1,2]
[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[2]Department of Psychology, University of Cambridge
[3]Microsoft Research Asia

Abstract

As multimodal language models (MLMs) are increasingly used in social and collaborative settings, it is crucial to evaluate their perspective-taking abilities. Existing benchmarks largely rely on text-based vignettes or static scene understanding, leaving visuospatial perspective-taking (VPT) underexplored. We procedurally generate large stimulus batteries for two evaluation tasks adapted from human studies: the Rotating Figure Task, probing perspective-taking across angular disparities, and the Director Task, assessing VPT in a referential communication paradigm. Across both tasks, MLMs show pronounced deficits in Level 2 VPT, with failure patterns indicating reliance on simple mirroring heuristics rather than genuine perspective transformations. These results expose critical limitations in current MLMs' ability to represent and reason about alternative perspectives, with implications for use in collaborative contexts.

Keywords: Artificial Intelligence; Social Cognition; Theory of Mind; Spatial Cognition; Language Understanding

Introduction

Generative models are increasingly embedded within social domains, taking on roles as teachers, colleagues, therapists, friends, and romantic partners (Collins et al., 2024; Shevlin, 2024). As deployment expands, reliable methods for understanding the strengths and limitations of their social cognition become essential – both to identify safe contexts of use and to anticipate dangerous or costly failure cases (Bengio et al., 2025; Department for Science, Innovation and Technology, 2025; Massachusetts Institute of Technology, 2025).

One core feature of human social cognition is Theory of Mind (ToM): the capacity to reason about, explain, and predict others' behaviour based on their beliefs, intentions, and emotions (Baron-Cohen et al., 1985; Frith & Frith, 2007, 2012; Premack & Woodruff, 1978; Wimmer & Perner, 1983). There is growing debate about whether and to what extent such capabilities exist in artificial systems, and how they should be measured (Hu et al., 2025). Most existing evaluations rely on vignette-style tasks adapted from cognitive science (Baron-Cohen et al., 1999; Happé, 1994), in which participants read short stories and reason about characters' mental states. These are easily formatted for language models, but can often be solved by exploiting shallow textual associations rather than understanding underlying causal factors (Hernández-Orallo, 2019; McCoy et al., 2023), and performance frequently fails to generalise to novel or atypical scenarios (Kim et al., 2023; Shapira et al., 2023; Ullman, 2023). Multimodal capabilities open new evaluation routes (Frank, 2023; Voudouris et al., 2025): paradigms grounded in visual context more closely resemble the inferential demands of real social interaction, where what others know must be inferred from gaze, occlusion, and spatial layout rather than stated explicitly in a narrative.

*Corresponding Author: jep84@cam.ac.uk

Visuospatial perspective taking

Visuospatial perspective taking (VPT) – representing what someone else sees and how they see it – is a core component of everyday social cognition. It supports shared attention, referential communication, and joint action across the lifespan (Clark & Brennan, 1991; Frith & Frith, 2012; Sacheli et al., 2022; Tomasello et al., 2005). By adopting another's viewpoint, we can infer what they know, what they want, or what they are referring to when they speak.

Both developmental and adult research distinguish two levels of VPT (Flavell, Everett, et al., 1981; Flavell, Flavell, et al., 1981; Lempers et al., 1977; Quesque et al., 2024). Level 1 concerns *whether* someone sees something: a judgement made rapidly and spontaneously through line-of-sight computation (O'Grady et al., 2020; Samson et al., 2010). Level 2 concerns *how* something appears: judgements that are effortful, scale with angular disparity, and require inhibiting one's own egocentric perspective and mentally rotating visual information into the other's viewpoint (De Lillo & Ferguson, 2023; Keysar et al., 2003; Michelon & Zacks, 2006; Surtees & Apperly, 2012; Surtees et al., 2013, 2016). Neuroimaging supports this dissociation: only Level 2 reliably recruits the right temporo-parietal junction (rTPJ), implicated in self-other distinction and egocentric inhibition (Martin et al., 2020; Perner et al., 2006; Schurz et al., 2014).

VPT also distinguishes *visual* judgements (what a scene looks like from another viewpoint) from *spatial* judgements (where things are located relative to that viewpoint) (Kessler & Rutherford, 2010; Quesque et al., 2024; Surtees et al., 2013). Both components are recruited together in referential communication (Dumontheil et al., 2010; Keysar et al., 2003). Disambiguating "Can you pass me the jar on the left?" depends on whether speaker and listener share visual access (the speaker would not refer to a jar only the listener can see) and whether they share a spatial frame (if the speaker faces the listener, *their* left is the listener's right).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Visuospatial perspective taking in multimodal language models

Recent advances have enabled language models to process multiple input modalities (Alayrac et al., 2022; Hurst et al., 2024; Liu et al., 2023), supporting their deployment in interactive multimodal settings such as instruction following, embodied reasoning, and collaborative task execution (see Zou et al. (2025) for a recent survey). Accordingly, growing attention has focused on whether multimodal language models (MLMs) can support visuospatial perspective taking.

Early investigations adapted Piaget's *Three Mountains Task* (Piaget & Inhelder, 1948) to assess Level 1 and Level 2 VPT (Gao et al., 2024; Linsley et al., 2024). While foundational, these benchmarks emphasise recognition of static scene layouts rather than the dynamic egocentric-to-allocentric transformations required for real-world interaction. Goral and colleagues (2024, 2025) adapted the *Dot Task* (Samson et al., 2010), distinguishing visual and spatial content but limiting their evaluation to Level 1 VPT. Leonard and colleagues (2024, 2025) extended this line of inquiry by adapting the *Rotating Figure Task* (Surtees et al., 2013), which probes both Level 1 and Level 2 VPT across viewpoint disparities ranging from 0° to 180°. They reported floor performance for non-shared perspectives, though their conclusions were limited by reliance on a single model (GPT-4o) and sparse trial diversity (eight trials per condition). Notably, chain-of-thought prompting improved performance only at 180°, suggesting reliance on a simple mirroring heuristic (e.g., "swap left and right if the figure faces me") rather than a general VPT mechanism – a pattern deserving closer examination.

The current study extends this work along two complementary paradigms. The *Rotating Figure Task* isolates the geometric components of VPT under controlled stimulus variation: by independently manipulating angular disparity, content type (visual vs. spatial), and perspective level (1 vs. 2), it allows us to characterise whether models support genuine perspective transformations or instead rely on simple heuristics. We expand prior work by evaluating a broader range of models, trials, and conditions, including reasoning-optimised models that test whether inference-time reasoning yields genuine improvements. The *Director Task* (Keysar et al., 2003) complements this by embedding VPT in a referential communication paradigm, testing whether models can deploy perspective information in service of an interactive goal – disambiguating referents based on a partner's occluded view. Related work shows MLMs also struggle with perspectival language, with prompting interventions only partially closing the gap (Dong et al., 2025). Together, the two tasks dissociate the underlying *capacity* for perspective transformation from its *deployment* in functional social interaction.

Experiments

We adopt a procedural approach to evaluating VPT in MLMs, generating large, controlled stimulus batteries through systematic and stochastic variation of key parameters (Fränken et al., 2024; Gandhi et al., 2023; Prunty et al., 2025). This enables fine-grained manipulation of perspective level, content type, and viewpoint disparity alongside matched controls, while reducing benchmark contamination through novel stimulus generation. Using the VIGNET framework (Prunty et al., 2025), we construct *Rotating Figure Tasks* (Surtees et al., 2013) – substantially expanding prior MLM evaluations (Leonard et al., 2024, 2025) – and *Director Tasks* (De Lillo & Ferguson, 2023; Keysar et al., 2003), a referential communication paradigm not previously applied to MLMs. We test frontier OpenAI models with (o3, o4-mini) and without (GPT-4o, GPT-4o-mini) inference-time reasoning. Full implementation details and supplementary analyses are provided in the appendix.[1]

Rotating Figure Task

In the *Rotating Figure Task*, participants view a top-down image of a person in a room containing a reversible symbol (e.g., 6/9) placed on the floor (Figure 1). The critical manipulation is the angular disparity between figure and viewer: at 0° perspectives are fully shared, and at 180° they are opposite. By varying orientation from 0° to 359°, we sample the full continuum of shared and non-shared perspectives. Each condition is defined by manipulating the symbol's placement and orientation relative to the figure, and probes either basic visual understanding (Controls 1–2) or perspective taking (Tests 1–3).

- **Control 1 (Symbol identification)** probes symbol form and location from the viewer's perspective. *Visual:* "What number or letter can you see in the image?" *Spatial:* "Is the number or letter on the left or right side of the image?"
- **Control 2 (Figure orientation)**[2] probes the direction the figure is facing. *Visual:* "What colour is the wall directly in front of the person?" *Spatial:* "What side of the image is directly in front of the person?"
- **Test 1 (Level 1 VPT: Visibility)** probes what the figure can see. The symbol is placed in front of or behind the figure, with the visual field represented by an (invisible) cone (Figure 1, left). *Visual:* "Can the person see the number or letter?" *Spatial:* "Is the number or letter in front of or behind the person?"
- **Test 2 (Level 2 VPT: Appearance)** probes how a symbol appears from the figure's viewpoint and its location relative to them. The symbol is placed within the figure's visual field, offset slightly to their left or right, and oriented to be legible from their perspective (Figure 1, centre). *Visual:* "What number or letter can the person see?" *Spatial:* "Is the number or letter on the person's left or right?"
- **Test 3 (Level 2 VPT: Integrated visual–spatial)** requires combining visual and spatial perspective taking. Two distinct

[1] Appendix, framework, and analysis code: https://github.com/Kinds-of-Intelligence-CFI/visuospatial-perspective-taking

[2] During piloting, models struggled to infer viewing direction from top-down images. We therefore added a line-of-sight arrow (Figure 1) that scaffolds viewing-direction parsing without revealing symbol position, identity, or appearance.

Figure 1: Example stimuli from the *Rotating Figure Task* (visual field cone shown in yellow for illustration). Left: the figure *can see* the symbol *in front* of them (Test 1). Centre: the symbol `p` appears as a `d` from the figure's perspective, and is located on their *right* (Test 2). Right: If asked about how the symbol on the figure's right appears to them, the correct response would be `6`, requiring both visual and spatial perspective taking (Test 3).

symbols are placed within the figure's visual field, one to their left and one to their right; the model must locate the symbol on the queried side (spatial PT) and report how it appears (visual PT) from the figure's viewpoint (Figure 1, right). *Visuospatial:* "What number or letter can the person see on their {`side`} side?"

Adult humans perform this task with high accuracy, with Level 2 trials showing response-time costs that scale with angular disparity (Surtees et al., 2013). We extend this paradigm in two ways: by adding Test 3 to assess joint visual-spatial integration under Level 2 VPT, and by introducing controls that explicitly assess precursor perceptual abilities in MLMs that are typically assumed in humans (see Millière & Rathkopf, 2024). We additionally collected human baseline data on our extended paradigm (see Appendix).[3]

Results Control conditions verified MLMs' baseline perceptual competencies. Models achieved high accuracy on symbol identification (Control 1), though o3 showed elevated invalid responses. Models struggled with viewing direction extraction (Control 2), particularly when figures faced corners. Excluding ambiguous corner trials substantially improved performance for most models except GPT-4o-mini, suggesting fundamental limitations in extracting viewing direction from static images.

For test conditions, Angular Disparity – the absolute angular difference between viewer and figure perspectives – was binned into four 45° intervals (0–45°, 45–90°, 90–135°, 135–180°). We analysed performance using mixed-effects logistic regression with Type III Wald tests and post-hoc comparisons (full models and analysis by question type (visual/spatial) in Appendix). All main effects and interactions were significant ($p < .001$) except where noted.

[3]The final battery comprised 15k images (3k per stimulus set), yielding 27k trials due to dual visual and spatial questions in Controls 1–2 and Tests 1–2. All four models (o3, o4-mini, GPT-4o, GPT-4o-mini) completed the full battery. The human study ($N = 30$ adult participants) sampled 48 trials per condition (240 trials total) from this battery.

Table 1: Angular Disparity across test conditions

Test	Subject	0–45	45–90	90–135	135–180	χ^2	p
1-V	Human	0.97	0.98	0.97	0.98	1.4	0.715
	GPT-4o-mini	0.39	0.62	0.71	0.77	203.7	<.001
	GPT-4o	0.72	0.70	0.77	0.81	23.0	<.001
	o4-mini	0.99	0.84	0.86	0.98	145.5	<.001
	o3	0.80	0.82	0.80	0.78	4.2	0.242
1-S	Human	0.98	0.97	0.99	0.98	5.1	0.166
	GPT-4o-mini	0.58	0.56	0.52	0.53	5.3	0.151
	GPT-4o	0.82	0.75	0.72	0.78	17.0	<.001
	o4-mini	0.99	0.85	0.87	0.98	128.9	<.001
	o3	0.76	0.79	0.78	0.77	1.2	0.756
2-V	Human	0.95	0.90	0.85	0.77	38.3	<.001
	GPT-4o-mini	0.80	0.62	0.31	0.10	705.9	<.001
	GPT-4o	0.33	0.41	0.54	0.71	228.5	<.001
	o4-mini	0.46	0.41	0.65	0.88	368.3	<.001
	o3	0.65	0.55	0.50	0.66	55.7	<.001
2-S	Human	0.99	0.94	0.93	0.90	33.5	<.001
	GPT-4o-mini	0.72	0.53	0.43	0.25	321.4	<.001
	GPT-4o	0.78	0.59	0.44	0.19	494.2	<.001
	o4-mini	0.87	0.67	0.63	0.80	135.0	<.001
	o3	0.61	0.56	0.50	0.66	42.4	<.001
3-VS	Human	0.94	0.89	0.81	0.73	31.9	<.001
	GPT-4o-mini	0.79	0.33	0.09	0.01	689.4	<.001
	GPT-4o	0.79	0.40	0.13	0.01	647.3	<.001
	o4-mini	0.84	0.47	0.21	0.23	489.7	<.001
	o3	0.81	0.52	0.23	0.27	415.6	<.001

Values are estimated marginal probabilities of correct responses.
$\chi^2(3)$ tests the simple main effect of angular disparity within each subject and task.
Test labels use N-T (N = 1–3; V = visual, S = spatial, VS = visuospatial).

Test 1 (Level 1 VPT) revealed disparity-dependent patterns across models, in contrast to humans, who remained at ceiling regardless of figure rotation (> 97% for both visual and spatial questions). GPT-4o-mini showed monotonically increasing visual accuracy with disparity, performing below chance for aligned perspectives but above chance when inverted, while its spatial performance remained near chance. GPT-4o, o4-mini, and o3 maintained relatively stable performance across most rotations but showed a pronounced dip specifically at left-facing orientations (−90°; Figure 2) – a pattern not predicted by symmetric rotation costs. Although o4-mini approached human ceiling at most rotations, all models fell well short of human stability across the full range of viewpoints.

Test 2 (Level 2 VPT) exposed broader and qualitatively

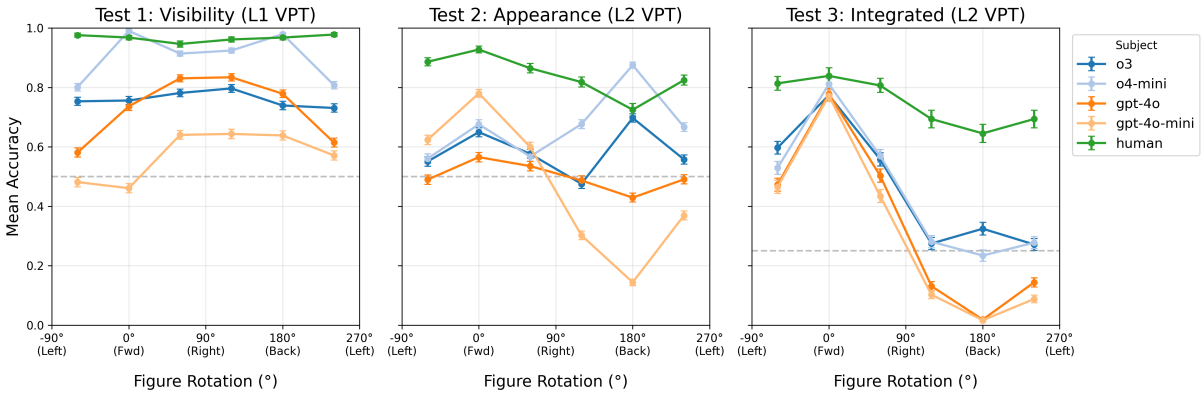

Figure 2: Mean MLM accuracy in Test conditions of the *Figure Rotation Task*, binned by figure rotation angle (12 bins) and aggregated across visual and spatial questions in Tests 1–2. 0° indicates a shared perspective and 180° an opposite perspective; 90° and −90° correspond to right- and left-facing figures (270° is the left wrap-around point). The dashed line indicates chance.

distinct limitations. Humans showed a gradual accuracy decline with increasing disparity (visual: 0.95 to 0.77; spatial: 0.99 to 0.90), consistent with the response-time costs reported in the human VPT literature (Surtees et al., 2013). Models, in contrast, displayed sharply non-human profiles. GPT-4o-mini's accuracy declined steeply from aligned to inverted viewpoints, indicating systematic failure to adopt perspective. GPT-4o exhibited dissociated patterns for visual versus spatial questions: spatial accuracy decreased monotonically while visual accuracy *improved* with inversion, suggesting inconsistent transformation strategies. Reasoning models o3 and o4-mini showed M-shaped performance – elevated accuracy at aligned and inverted viewpoints but reduced accuracy at intermediate disparities, with o4-mini's visual accuracy at 135°–180° (0.88) even exceeding the human baseline (0.77). This profile is inconsistent with the graded disparity cost expected of mental rotation, instead suggesting reliance on simple mirroring strategies that succeed only at fully shared (0°) or fully opposite (180°) viewpoints.

Test 3 (Integrated L2 VPT) combined Level 2 visual and spatial perspective-taking. Humans showed a graded decline across angular disparity (0.94 to 0.73), with overall accuracy depressed relative to Test 2, confirming that integrating visual and spatial perspective-taking imposes a measurable cost beyond either component alone. Models, in contrast, fell from above-chance at aligned perspectives to near floor for GPT-4o variants and approximately chance for reasoning models at maximum disparity – a magnitude of decline far exceeding the human baseline. Notably, the reasoning models (o3, o4-mini) were not spared despite their strong Test 2 performance on 180° figure rotations: integration exposed limitations that single-component tasks did not. None of the tested models reliably supports flexible, integrated visuospatial perspective-taking.

Mirroring heuristic The M-shaped profiles in Test 2 – peak accuracy at 0° and 180°, with failures at intermediate rotations – are consistent with a mirroring strategy that succeeds only at canonical (shared or fully opposed) viewpoints. However, the two-way reversible symbols used in Test 2 (e.g., 6/9, p/d) cannot fully diagnose this account, as there is no valid response option for 90° rotations. To probe model strategies directly, we ran a follow-up using the `m` symbol, which affords four distinct rotational interpretations (`m`, `3`, `w`, `E`). We replicated Test 2 visual using this symbol, with all four interpretations as response options. If models rely on a mirroring strategy, accuracy should peak at 0° and 180° and drop sharply at ±90°, rather than showing a graded decline across angular disparity.

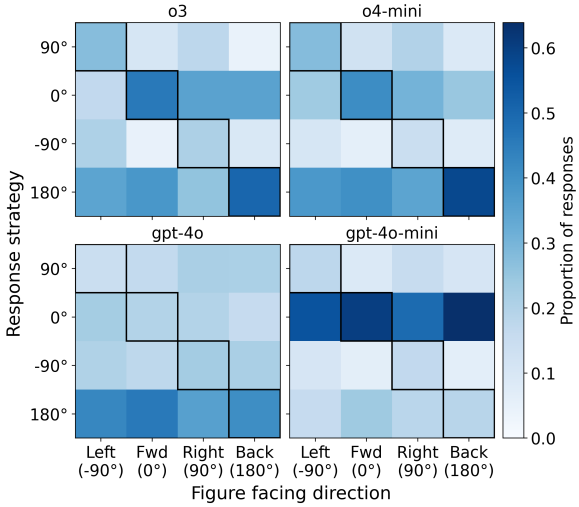

Figure 3: Response strategies in the diagnostic-symbol follow-up. Each cell shows the proportion of responses corresponding to a given interpretation of the symbol (y-axis: 0° = as displayed to viewer; 180° = flipped; 90°/−90° = rotated clockwise/anti-clockwise) at each figure rotation bin (x-axis). Bold borders mark the correct response strategy for each figure orientation.

The accuracy data dissociated reasoning from non-reasoning models. Reasoning models (o3, o4-mini) showed the predicted

M-shape, with peaks at 0° and 180° and troughs at ±90°. GPT-4o and GPT-4o-mini instead showed single peaks – at 180° and 0° respectively – suggesting commitment to a single fixed strategy rather than switching between heuristics (see Appendix). Figure 3 confirms these strategies directly: GPT-4o-mini responded egocentrically across rotations; GPT-4o applied a consistent 180° flip; while o3 and o4-mini concentrated correct responses at 0° (egocentric) and 180° (flipped), with responses at ±90° dispersed across categories. 90° transformations were rare across all models, consistent with the heuristic account of Test 2.

Director Task

To assess MLM VPT in an applied communicative setting, we adapted the *Director Task* (Keysar et al., 2003). Participants follow instructions from a director positioned opposite a 4×4 grid, selecting target objects by cell location. Some cells have opaque backs blocking the director's view, creating visual-shared (mutually visible) and visual-different (participant-only) perspectives. On visual-different trials, instructions like "select the smallest star" require responding from the director's limited perspective rather than the viewer's privileged one (Figure 4).

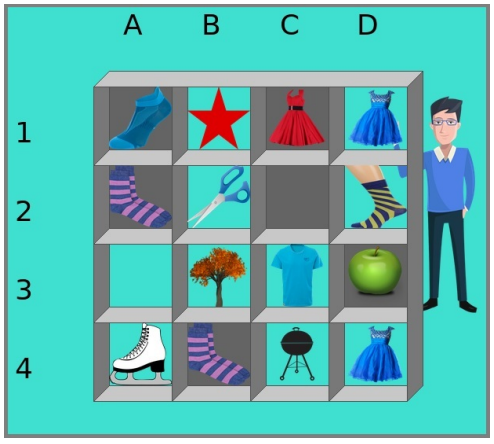

Figure 4: Example stimulus from the *Director Task*. If the director says "Please select the rightmost blue, non-striped item of clothing from my point of view", he would be referring to the item in `C3`. The answer is not `D1` or `D4`, as the director specified the rightmost item from *his* perspective. The answer is also not `A1` as the item is occluded from the director's view. All other items do not fit the object specification.

Grid configurations are procedurally generated, systematically varying *visual perspective* – whether target items have occluded alternatives – and *relative adjective* type:

- **None.** Only one object matches the director's description (e.g., "red book").
- **Size.** Multiple matches, target specified by size (e.g., "the largest red book").
- **Spatial.** Multiple matches, target specified by a spatial term:
  - **Shared**. Vertical terms ("topmost"/"bottommost"), invariant across viewpoints.
  - **Different**. Horizontal terms ("leftmost"/"rightmost"), reversed across viewpoints.

Instructions are framed as "from your/my point of view", defining *perspective reversal* conditions that prevent spatial-different trials from being solved by automatic left–right swapping. Importantly, this framing affects only horizontal adjectives, not visual perspective: the director cannot see occluded objects regardless of the framing.

To isolate VPT from basic perceptual demands (e.g., object identification), we developed a text-based (ASCII) version structurally identical to the image-based task, providing convergent evidence across modalities.[4]

Results We analysed MLM performance using mixed-effects logistic regression. Full analyses (see appendix) established that: 1) Image-based trials showed lower accuracy than ASCII, but VPT deficits were robust across both formats and cannot be attributed to image processing alone. 2) Relative adjective type (none, size, spatial-shared) did not affect baseline performance. 3) Spatial PT impairments emerged exclusively in the critical cell predicted by the design – horizontal-adjective trials from the director's point of view – with robust performance elsewhere. We therefore focus on the visual × spatial PT interaction, collapsed across task format.

We modelled this interaction as AI Model × Visual Perspective × Spatial Perspective with Task Format as a covariate, restricted to spatial relative adjective trials from the director's point of view. All main effects and interactions, including the three-way interaction, were significant ($p < .001$); MLM performance is summarised in Table 2 and Figure 5.

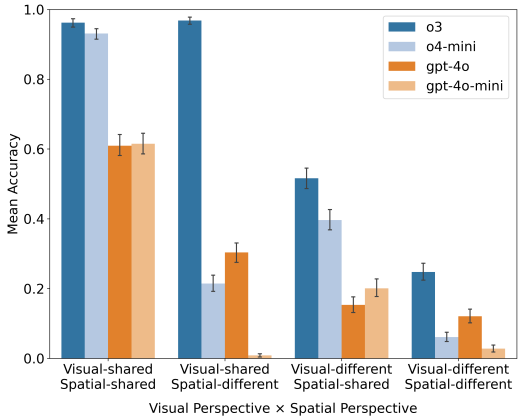

Figure 5: MLM mean accuracy in the *Director Task* for visual (occluded vs. no occluded alternatives) and spatial (horizontal vs. vertical adjectives) perspective taking trials, aggregated across task formats and restricted to spatial relative adjective trials from the director's perspective. Error bars are 95% confidence intervals.

[4]The final battery comprised 8k images (2k per dataset across four datasets with related item proportions: 0.3, 0.5, 0.7 and 0.9), yielding 16k samples across image and ASCII versions.

Table 2: Visual and spatial perspective contrasts (shared vs. different) by model, on director-POV spatial-adjective trials.

Perspective	Model	Shared	Different	OR	SE	p
Visual	GPT-4o-mini	0.10	0.08	1.37	0.36	0.242
	GPT-4o	0.45	0.13	5.30	1.03	<.001
	o4-mini	0.66	0.17	9.46	1.91	<.001
	o3	0.97	0.37	52.87	11.39	<.001
Spatial	GPT-4o-mini	0.39	0.01	43.62	11.60	<.001
	GPT-4o	0.35	0.19	2.27	0.44	<.001
	o4-mini	0.76	0.11	25.04	5.05	<.001
	o3	0.85	0.77	1.69	0.36	0.015

Shared/Different are estimated marginal probabilities. OR = odds ratio, SE = standard error and p from Shared/Different pairwise contrasts.

A strong main effect of AI model, $\chi^2(3) = 529.91$, revealed substantial performance differences, with reasoning models (o3, o4-mini) near ceiling when VPT was not required. Both Visual PT, $\chi^2(1) = 50.91$, and Spatial PT, $\chi^2(1) = 161.97$, showed significant main effects: performance declined when perspectives differed, with larger costs for spatial PT.

Crucially, the magnitude of these effects varied across models (AI Model × Visual PT, $\chi^2(3) = 85.18$; AI Model × Spatial PT, $\chi^2(3) = 449.09$; three-way, $\chi^2(3) = 134.52$), though much of this variation reflects ceiling and floor effects. GPT-4o-mini was at floor on visual trials regardless of perspective, and GPT-4o approached floor under either VPT demand, leaving little room to dissociate the two costs. Only o3 supported a clean dissociation: near-ceiling performance on shared trials gave way to a pronounced visual PT cost (0.97 to 0.37, OR = 52.87) but only a small spatial PT cost (0.85 to 0.77, OR = 1.69), with combined demands producing the largest decrement. o4-mini showed an intermediate profile, impaired under both forms of VPT but more so under spatial PT. The cleanest case (o3) thus dissociates the two demands: visual PT (occlusion tracking) is severely impaired while spatial PT (left/right reversal) is largely preserved.

Discussion

We evaluated visuospatial perspective-taking (VPT) in frontier multimodal language models using two complementary paradigms. Across both, models showed persistent VPT deficits despite strong perceptual baselines, with failure patterns pointing to reliance on shallow heuristics rather than genuine perspective transformations.

The clearest evidence comes from the M-shaped angular profiles in the *Rotating Figure Task* – accurate at 0° and 180° but failing at intermediate rotations – a signature inconsistent with the graded angular cost expected of mental rotation. The diagnostic-symbol follow-up confirmed this directly: even at 90° rotations, models defaulted to egocentric or 180°-flipped responses rather than producing the correct rotated interpretation. Such heuristics can succeed at canonical viewpoints but lack the representational machinery for arbitrary perspective transformations. Extended inference-time reasoning, chain-of-thought, or in-context learning interventions may improve performance through heuristic application, but it remains an open empirical question whether these methods will lead to the more flexible VPT that humans demonstrate.

These limitations matter most in applied settings. The *Director Task* showed all models approaching floor when VPT had to be deployed in service of referent identification, with combined visual and spatial demands proving especially costly. Comparable patterns across image-based and ASCII versions indicate that these failures are not artefacts of image processing: the same VPT bottleneck appears regardless of input format. Notably, only o3 showed robust spatial VPT in the *Director Task*, suggesting that even similar Rotating Figure profiles can yield different generalisation to applied contexts.

Related work

Our findings substantially expand prior work investigating VPT in MLMs. Leonard et al. (2024, 2025) found that GPT-4o showed improved performance at 180° disparity with chain-of-thought prompting, indirectly suggesting a mirroring strategy. Our work demonstrates this directly and across reasoning and non-reasoning models. Góral et al. (2024, 2025) showed impaired Level 1 VPT in GPT-4o and older models, with particularly poor spatial VPT. Our GPT-4o findings are consistent but reveal underlying mechanisms, including orientation-specific blind spots and divergent trajectories for visual versus spatial content. Critically, we demonstrate that these VPT deficits translate to failures in referential communication.

Comparisons with human VPT further highlight these differences. In humans, Level 2 judgements are effortful, scale with angular disparity, and recruit distinct neural mechanisms including rTPJ to inhibit egocentric representations (Martin et al., 2020; Surtees et al., 2013), whereas Level 1 visibility judgements are comparatively automatic (Samson et al., 2010). Adults in our paradigm replicated this graded angular signature, while models showed non-human-like profiles – most diagnostically, failures at *intermediate* rather than maximal disparities. These tasks offer a basis for future comparative work across human and artificial systems, including mechanistic analyses on open-source models.

Conclusion

Perspective-taking underpins referential communication, spatial coordination, and joint action (Clark & Brennan, 1991; Frith & Frith, 2012). Models that cannot flexibly represent what others can see will struggle with collaborative task execution, instruction-following in shared environments, and embodied reasoning (Zou et al., 2025). The failure modes identified here are easily hidden in casual interaction but emerge sharply under specific geometric configurations or combined VPT demands. Procedurally generated paradigms, which support novel item generation at scale, offer a precise diagnostic for distinguishing genuine perspective-taking capacity from heuristic shortcuts – a distinction that increasingly matters as MLMs are deployed in human social and physical environments.

Acknowledgements

The authors acknowledge the support of Accenture (https://www.accenture.com/us-en/services/data-ai) and Microsoft Research Asia (https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/) to this research. This research project has also benefitted from the Microsoft Accelerate Foundation Models Research (AFMR) grant program.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716–23736.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of autism and developmental disorders*, *29*, 407–418.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., et al. (2025). International ai safety report. *arXiv preprint arXiv:2501.17805*.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, L. B., M. John, S. Teasley, & D. (Eds.), *Perspectives on socially shared cognition* (pp. 13–1991). American Psychological Association.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

De Lillo, M., & Ferguson, H. J. (2023). Perspective-taking and social inferences in adolescents, young adults, and older adults. *Journal of Experimental Psychology: General*, *152*(5), 1420.

Department for Science, Innovation and Technology. (2025). Safety and security risks of generative artificial intelligence to 2025.

Dong, D. T., Luo, Y., Wang, P.-Y. A., Ozyurek, A., & Rubio-Fernández, P. (2025). You prefer this one, i prefer yours: Using reference words is harder than vocabulary words for humans and multimodal language models. *arXiv preprint arXiv:2506.00065*.

Dumontheil, I., Küster, O., Apperly, I. A., & Blakemore, S.-J. (2010). Taking perspective into account in a communicative task. *Neuroimage*, *52*(4), 1574–1583.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the level 1–level 2 distinction. *Developmental psychology*, *17*(1), 99.

Flavell, J. H., Flavell, E. R., Green, F. L., & Wilcox, S. A. (1981). The development of three spatial perspective-taking rules. *Child Development*, 356–358.

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*(8), 451–452.

Fränken, J.-P., Gandhi, K., Qiu, T., Khawaja, A., Goodman, N. D., & Gerstenberg, T. (2024). Procedural dilemma generation for evaluating moral reasoning in humans and language models. *arXiv preprint arXiv:2404.10975*.

Frith, C. D., & Frith, U. (2007). Social cognition in humans. *Current biology*, *17*(16), R724–R732.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, *63*(1), 287–313.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518–13529.

Gao, Q., Li, Y., Lyu, H., Sun, H., Luo, D., & Deng, H. (2024). Vision language models see what you want but not what you see. *arXiv preprint arXiv:2410.00324*.

Góral, G., Ziarko, A., Miłoś, P., Nauman, M., Wołczyk, M., & Kosiński, M. (2025). Beyond recognition: Evaluating visual perspective taking in vision language models. *arXiv preprint arXiv:2505.03821*.

Góral, G., Ziarko, A., Nauman, M., & Wołczyk, M. (2024). Seeing through their eyes: Evaluating visual perspective taking in vision language models. *arXiv preprint arXiv:2409.12969*.

Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*(2), 129–154.

Hernández-Orallo, J. (2019). Gazing into clever hans machines. *Nature Machine Intelligence*, *1*(4), 172–173.

Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, *380*(1932), 20230499.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kessler, K., & Rutherford, H. (2010). The two forms of visuospatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in psychology*, *1*, 213.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Kim, H., Sclar, M., Zhou, X., Bras, R., Kim, G., Choi, Y., & Sap, M. (2023). Fantom: A benchmark for stress-testing machine theory of mind in interactions. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14397–14413.

Lempers, J. D., Flavell, E. R., & Flavell, J. H. (1977). The development in very young children of tacit knowledge con-

cerning visual perception. *Genetic Psychology Monographs*, *95*(1), 3–53.

Leonard, B., Woodard, K., & Murray, S. (2025). Why multimodal models struggle with spatial reasoning: Insights from human cognition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Leonard, B., Woodard, K., & Murray, S. O. (2024). Failures in perspective-taking of multimodal ai systems. *arXiv preprint arXiv:2409.13929*.

Linsley, D., Zhou, P., Ashok, A. K., Nagaraj, A., Gaonkar, G., Lewis, F. E., Pizlo, Z., & Serre, T. (2024). The 3d-pc: A benchmark for visual perspective taking in humans and machines. *arXiv preprint arXiv:2406.04138*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.

Martin, A. K., Kessler, K., Cooke, S., Huang, J., & Meinzer, M. (2020). The right temporoparietal junction is causally associated with embodied perspective-taking. *Journal of Neuroscience*, *40*(15), 3089–3095.

Massachusetts Institute of Technology. (2025). The GenAI Divide. State of AI in Business 2025.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & psychophysics*, *68*(2), 327–337.

Millière, R., & Rathkopf, C. (2024). Anthropocentric bias and the possibility of artificial cognition. *arXiv preprint arXiv:2407.03859*.

O'Grady, C., Scott-Phillips, T., Lavelle, S., & Smith, K. (2020). Perspective-taking is spontaneous but not automatic. *Quarterly Journal of Experimental Psychology*, *73*(10), 1605–1628.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social neuroscience*, *1*(3-4), 245–258.

Piaget, J., & Inhelder, B. (1948). *La représentation de l'espace chez l'enfant*. Presses universitaires de France.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Prunty, J., O'Flynn, A., Quinn, P., & Cheke, L. G. (2025). Intuit: Investigating intuitive reasoning in humans and language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H., Bernstein, D., Bertoux, M., Bird, G., Bukowski, H., et al. (2024). Defining key concepts for mental state attribution. *Communications Psychology*, *2*(1), 29.

Sacheli, L. M., Arcangeli, E., Carioti, D., Butterfill, S., & Berlingeri, M. (2022). Taking apart what brings us together: The role of action prediction, perspective-taking, and theory of mind in joint action. *Quarterly journal of experimental psychology*, *75*(7), 1228–1243.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of experimental psychology: human perception and performance*, *36*(5), 1255.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Shevlin, H. (2024). All too human? identifying and mitigating ethical risks of social ai.

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426–438.

Surtees, A., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child development*, *83*(2), 452–460.

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–691.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Voudouris, K., Cheke, L., & Schulz, E. (2025). Bringing comparative cognition approaches to ai systems. *Nature Reviews Psychology*, 1–2.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Zou, H. P., Huang, W.-C., Wu, Y., Chen, Y., Miao, C., Nguyen, H., Zhou, Y., Zhang, W., Fang, L., He, L., et al. (2025). A survey on large language model based human-agent systems. *Authorea Preprints*.