

# Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models

Sherzod Hakimov<sup>1</sup>, Yerkezhan Abdullayeva<sup>1</sup>, Kushal Koshti<sup>1</sup>, Antonia Schmidt<sup>1</sup>, Yan Weiser<sup>1</sup>, Anne Beyer<sup>1</sup>, David Schlangen<sup>1,2</sup>

<sup>1</sup>Computational Linguistics, Department of Linguistics  
University of Potsdam, Germany

<sup>2</sup>German Research Center for Artificial Intelligence (DFKI), Berlin, Germany  
firstname.lastname@uni-potsdam.de

## Abstract

While the situation has improved for text-only models, it again seems to be the case currently that multimodal (text and image) models develop faster than ways to evaluate them. In this paper, we bring a recently developed evaluation paradigm from text models to multimodal models, namely evaluation through the goal-oriented game (self) play, complementing reference-based and preference-based evaluation. Specifically, we define games that challenge a model’s capability to represent a situation from visual information and align such representations through dialogue. We find that the largest closed models perform rather well on the games that we define, while even the best open-weight models struggle with them. On further analysis, we find that the exceptional deep captioning capabilities of the largest models drive some of the performance. There is still room to grow for both kinds of models, ensuring the continued relevance of the benchmark.

## 1 Introduction

Large *multimodal* models (LMMs; such as GPT4o,<sup>1</sup> InternVL (Chen et al., 2023)) that can handle images as input together with text seem poised to play a significant role in constructing a new, more capable kind of situated interactive agent. What is particularly exciting about them is that they, in contrast to earlier attempts at building such systems, as surveyed by Suglia et al. (2024), promise to be *generalist* models that can be adapted to tasks at hand through methods that require few or even no data and training cost. Current methods for evaluating them, however, largely do not address this potential, following mostly the *reference-based evaluation* paradigm and probing for reasoning and recognition capabilities in static contexts.

In this paper, we investigate whether a recent new evaluation paradigm for text-based

<sup>1</sup><https://openai.com/index/hello-gpt-4o/>

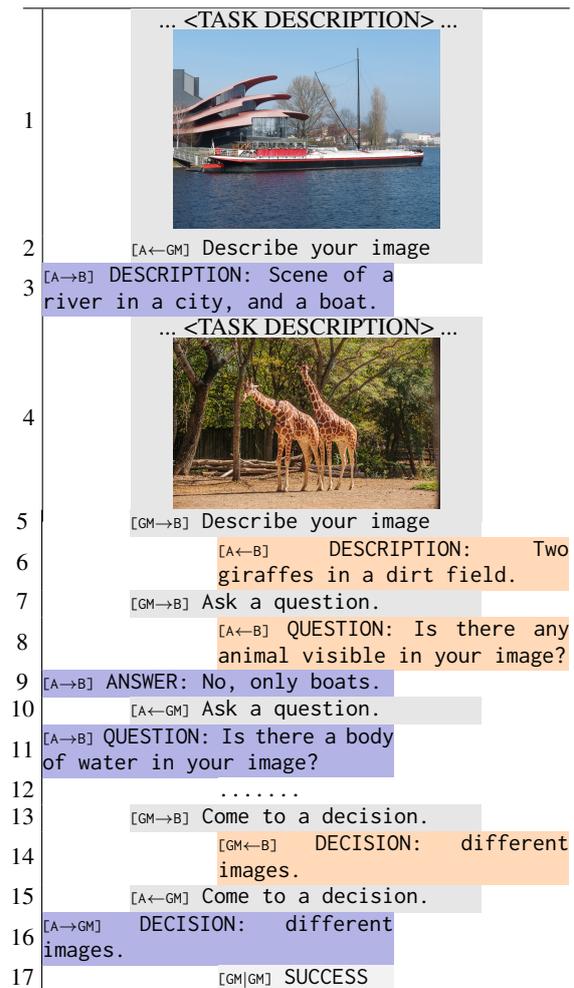


Figure 1: Example dialogue from *MatchIt* game, between players A (turns highlighted in purple) and B (orange), and a programmatic “game master” (grey). The task is to identify whether A and B were given the same image or not, via text interaction only. The game master scaffolds the game by prompting the players and relaying information.

models—*game-agency-based evaluation*—can be transferred to the evaluation of text and image models. Specifically, we selected one of the various frameworks that appeared last year for defining such games, *clmgame* (Chalamalasetti et al., 2023), and adapted it to evaluate multimodal mod-

els. We defined three dialogue games (on reference, image comparison, and navigation) that focus on the ability to build a model of a situation that is presented as an image and, in two of them, to align it with a partner. We make the following observations:

- Current LMMs are capable of conducting situated interactions if given enough scaffolding by an agent framework.
- There are significant differences in the degree of this ability, however, between commercial and open models (43 points on our 0–100 scale between the best of each kind), mirroring the situation that text-only models were in previously (Beyer et al., 2024).
- Much of the performance is driven by the excellent *deep captioning* abilities of the largest models; these break down on very detailed abstract images.
- Elementary capabilities for representing spatial configurations (or, more abstractly, graph structures) seem to be present in the larger models.

We made the source code for the implemented games and the extended framework publicly available at: <https://github.com/clembench/>. The leaderboard of evaluated multimodal LLMs is available here (tab *Multimodal*): <https://clembench.github.io/leaderboard.html>.

## 2 Related Work

**Evaluating LLMs** Following traditional practice in NLP, the first main paradigm for evaluation LLMs was what can be called *reference-based evaluation*, where a model response to a test item is compared to a known correct response. As a reaction to the rapidly increasing scores of the latest models and the saturation of existing benchmarks (Wang et al., 2019), meta-benchmarks have been set up, such as HELM (Liang et al., 2022) and BIGbench (Srivastava et al., 2022). While this method offers control over the tasks that are tested, a recently highlighted problem is ensuring train/test splits in the era of extremely large (and intransparent) training sets (Magar and Schwartz, 2022). Another popular method for evaluation falls under what could be called the *preference-based evaluation* paradigm. This is represented by Chatbot Arena (Chiang et al., 2024), which lets users

present queries to two models in parallel and then ask for preferences. This has the advantage of higher ecological validity as the human/system interaction is being evaluated. However, it comes with the cost of very little control over the tested distribution of tasks. Finally, a newly emerging paradigm is *game-agency-based evaluation* (Chalamalasetti et al., 2023; Chan et al., 2023; Qiao et al., 2023). In this paradigm, evaluation is framed as measuring the success of LLMs in conducting task-oriented interactions in simple conversational games. This has the advantage that it does not require user interaction (unlike the preference-based paradigm) while still keeping goal orientation and strategy in focus. In this paper, we want to explore this paradigm for the evaluation of LMMs.

**Evaluating LMMs** The evaluation of the newer field of Large Multimodal Models so far mostly remains within the *reference-base evaluation* paradigm,<sup>2</sup> with datasets such as *MME* (Fu et al., 2023), *MMBench* (Liu et al., 2023), *MMMU* (Yue et al., 2023), and *SEED-Bench* v1 (Li et al., 2023b) and v2 (Li et al., 2023a). These include image and text pairs as test instances for various tasks such as question answering, reasoning, answering scientific questions, etc. *VHELM* (visual HELM)<sup>3</sup> uses *MMMU* and two other visual question answering datasets (Gurari et al., 2018; Goyal et al., 2017) to extend the HELM framework to test multimodal LLMs. Our aim here is not to replace this kind of evaluation but rather to complement it with a focus on different capabilities, or at least differently challenged capabilities (see below).

Before we describe the general structure of our games, we will briefly also review literature relevant to each of them separately.

**Reference Games** The use of reference games where one player gets another to identify an item through verbal means, goes back to at least Krauss and Weinheimer (1964) and Lewis (1969) in linguistics and psycholinguistics, and has seen increased use in NLP in recent years as well (Shen et al., 2018; Haber et al., 2019; Sadler et al., 2024). Its attraction lies in the very clear way in which

<sup>2</sup>Although first attempts are underway to establish attempts in the *preference-based evaluation* paradigm as well, with the Multimodality Chatbot Arena <http://vlarena.opengvlab.com>. This however seems to be much less popular so far than its text-based counterpart.

<sup>3</sup>Accessed in May 2024. <https://crfm.stanford.edu/helm/vhelm/latest/>

it brings out context dependence (a good referring expression not only *describes* the target object, but also *excludes* distractor objects) and, especially in settings where there are repeated references (Sadovnik et al., 2012), partner-effects as well (precedence; (Brennan and Clark, 1996)).

**Image Comparison** The second game that we implement follows a suggestion by Schlangen (2019), who uses it to illustrate the concept of a *Grounded Agreement Game*. The idea here is to go beyond settings like Visual Dialog (Das et al., 2017) or Guesswhat?! (de Vries et al., 2017), which were quite popular in the language and vision field at the time. The criticism in this paper was that these settings while eliciting dialogue in the sense of sequential turns from different speakers, do not provide much purpose to the interaction. Grounded Agreement Games, on the other hand, by letting players share a common goal of reaching mutual understanding, provide a “*joint purpose*, a shared sense of semantic ownership of the interaction” (Schlangen, 2019). Also related, in incorporating visual information and cooperation between participants, is spot-the-difference (Lopes et al., 2018); this corpus, however, has only been used for linguistic analysis.

**Navigation and Exploration** Following natural language navigation instructions is a well-established task in the intersection of Computer Vision and NLP, and many special purpose models have been built in recent years (Gu et al., 2022). The task described below is related but posed to *generalist models*, and tasks the model only with exploration. More abstractly, what is tested is the ability to explore graph structures and testing spatial reasoning abilities (Shi et al., 2022; Rizvi et al., 2024). As a task for the assessment of models, something related has been used by Momennejad et al. (2023) in CogEval. Their results suggest that LLMs lack emergent cognitive map comprehension or planning competence, finding that LLMs can navigate simple graphs but struggle with spatial relationships and complex graphs due to looping, missing edges, and longer trajectories. The NLGraph benchmark (Wang et al., 2023) tested LLMs’ ability to perform explicit graph reasoning on eight tasks. The models perform basic graph reasoning on cycle and shortest path tasks but fail on the Hamilton Path, according to the NLGraph benchmark assessment. Bubeck et al. (2023) demonstrated anecdotally that GPT-4 seems

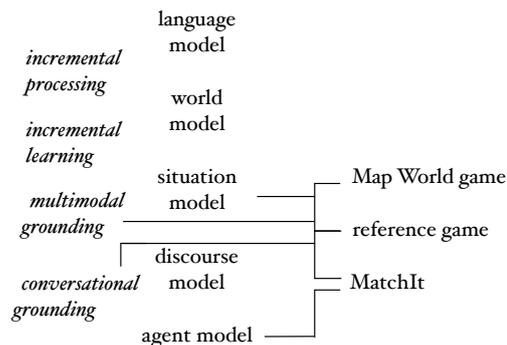


Figure 2: Relating the Dialogue Games used here to the construct model from Schlangen (2023b)

to have extensive spatial reasoning and map navigation abilities. This was criticized by (Liu and Wu, 2023), who showed that responses may become more error-prone when graph density exceeds a certain threshold, potentially causing hallucinations. Li et al. (2024) introduced another dataset for visual map navigation task with where commercial models (GPT-4V, Gemini) struggled with spatial reasoning sub-tasks.

### 3 Dialogue Games as Benchmarking Tool

**Construct Validity** We follow (Chalamalasetti et al., 2023) in striving for construct validity in our measurements and taking inspiration from the model of (Schlangen, 2023c,a). As can be seen in Figure 2 (modified from Chalamalasetti et al. (2023)), we link the games introduced here to the *situation model* (representing the situation in which the task at hand is to be performed), *multimodal grounding* (linking language and visual information), and, at least in simple forms, to *conversational grounding* and the *agent model*. See the original papers for an explanation of the model and the other components. How the individual games challenge these aspects will be explained below.

#### Scaffolding Game Play with a GameMaster

We decided to use clemgame/clembench (Chalamalasetti et al., 2023) as the framework to realise the idea of “self-play for evaluation”. The main idea of this framework is that dialogue games are specified through *prompt templates*, which explain the game goals to the players in natural language. The game goals include the task description and specific rules for formatting responses (so that they can be parsed). A programmatic GameMaster then realises the game play through the instantiation of the templates with specific game instances (e.g., in the game from Figure 1, an instance would be

defined by a given pair of images), and the turn-by-turn prompting of the *players* (which can be models or human players).

The resulting *episodes* are then scored through game-specific *scoring rules*. For each game, one scoring metric is determined as the *quality metric* (always ranging from 0 (worst) to 100 (best)). An overall score is computed by averaging this metric by game and then over games. Games where a player violates the parsing rules count as not played (until the end); the percentage of games played hence can serve as a metric for *formatting instruction following ability*, whereas the *quality metric* measures the ability to play the respective game successfully (only for those episodes that were played until the end). We aggregate these two scores to a single number, the *clemscore*, as the quality metric weighted by % played (scaled to the interval  $[0, 1]$ ).

## 4 Three Multimodal Games

In this section, we describe the three different games that we set up, with a focus on which capabilities exactly they are meant to target.

### 4.1 Reference: The Reference Game

**Game Description** Player A is presented with three images, and tasked with getting player B, who may see them in a different order, to identify the first of these. Player B is then presented with the three images, potentially in a different order, together with A’s reference, and is tasked to identify the referent. This is a single-turn game (Figure 9, 10, 6).

**Capability Tested** The idea is that this game challenges the referring model to go beyond simple descriptions of the image content towards *contrastive descriptions* that exclude the distractor images, and ideally also *efficient descriptions* that do so by concentrating on distinguishing features (Gatt and Kraemer, 2018).

**Scoring** Each episode is scored as 1 if successful (B picks out the intended referent), 0 otherwise.

**Instances** We created different sets of instances, with the hypothesis that they might challenge the models differently. First, we created grid-like pixel images (Figure 9), which we varied in terms of ‘compressability’: from simple-to-recognise (for humans) patterns to random placements. We created these stimuli in two different renderings: As character-based ‘images’ (hence suitable for text-only models, to allow for a comparison in perfor-

mance; filled cells are marked with the character “X”), as well as real images (converted from the text representations).

Second, we selected sets of photographs (Figure 10) (or photo-realistic renderings) of scenes or configurations of objects, to contrast handling of more naturalistic scenes with the set of grid-images. We included instances from three datasets: ADE20K (Zhou et al., 2017), DOCCI (Onoe et al., 2024), CLEVR (Johnson et al., 2017). We selected one target and two distractors chosen based on the similarity to the target (based on available metadata in each dataset; scene category information in ADE20K, the list of concepts in DOCCI, object categories in CLEVR).

Third, we created boards that include *pentomino* puzzle pieces (Figure 6) to analyse whether models are capable of handling unusual shapes and crowded scenes. We take code from Sadler et al. (2024) and generate a wide variety of scenes, in sets of images with very small differences. From this, we sample randomly. In total, there are 13 experiments corresponding to 390 instances.

### 4.2 Alignment: The MatchIt Game

**Game Description** Player A is presented with an image, as is Player B. The two images are either *identical*, or *different*. The task of the players is to find out which is the case. This game is heavily scaffolded by the GameMaster, which prompts the players to produce a description and ask a question of the other player (Figure 1). The dialogue continues with question and answering rounds (where both players ask and answer each other’s question) until players make a decision (SAME, DIFFERENT) about the given images (or GameMaster intervenes if maximum number of rounds is reached).

**Capability Tested** Our hypothesis is that good gameplay requires reasoning about what distinguishing features could be, the presence or absence of which would allow for making the same/different decision. This can then influence both the initial description that is produced and what questions are asked. In principle, allowing more rounds of mutual questioning should make the task easier.

**Scoring** Each episode is scored 1 (A and B both make the correct determination) or 0.

**Instances** Three difficulties were defined for the multimodal variant of MatchIt: both players get the same image, both players get similar images or completely different images, the hypothesis being



Figure 3: A pair of similar images for MatchIt.

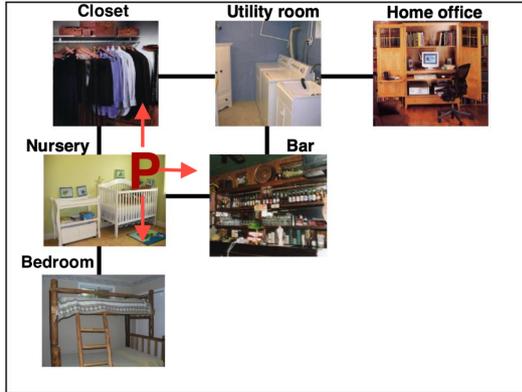


Figure 4: Environment for the text-only Map game. The player (denoted with **P**) is currently in the *Nursery* and has the option to move to one of the neighboring rooms (*Bar*, *Closet* or *Bedroom*). The player moves by choosing a cardinal direction: east, west, north, south

that different images are the easiest to recognize, followed by same and similar image pairs. The curation rationale for a similar picture was that both photos could be described with the same (short) sentence, but their difference should be striking enough that one (short) sentence should be enough. Figure 3 illustrates a similar image pair. The images used were taken from the *Visual Genome* dataset (Krishna et al., 2017) and sampled for the category of similar photos in a multi-step process via Jaccard similarity of sets of object annotations and their attributes and cosine similarity of *CLIP* image encoder embeddings (Radford et al., 2021). The detailed process is described in Appendix C. Ten instances of each difficulty (same, similar, different) were part of the final game play, for a total of 30 instances. Finally, we also sampled accordingly from the set of pentomino images described above in Section 4.1. In total, there are six experiments corresponding to 60 instances.

### 4.3 Navigation & Exploration: Map Game

**Game Description** This is a single-player game in which the player explores a network of connected rooms, it is based on the environment “MapWorld” of Ilinykh et al. (2019). At any point in the game, the player is in one of the rooms of the network (or

*map*). The player can move into adjacent rooms by issuing a navigational command (east, west, north, or south) (see Figure 4). Information about the room is relayed to the player by the GameMaster by giving the *image* of the room (the name of room, e.g. “Nursery”, is never revealed); information about the directions in which adjacent rooms can be found is always relayed via text (e.g. “From here you can go north, south, east.”). Within this general setting, we define several versions: *Go to X* ( $G2X$ ), in which the player is tasked to find a room of a specific category and indicate when they think the goal has been achieved. *Explore Exhaustively* ( $EE$ ), in which the player is tasked with visiting all rooms of the map and indicate when it thinks the goal has been achieved. In *graph reasoning* ( $EE-gr$ ), the player is prompted to generate the action along with the representation of the already explored graph explicitly.

**Capability Tested** Unlike in the previous two games, the situation relevant to the game is not observable in one go but rather must be explored actively. To perform well in this family of games, an internal representation of the map must be kept. Moreover, to be efficient, some spatial reasoning over this implicit structure is required to keep track of as yet unexplored rooms.

**Scoring** Scoring is more complex in this game. We define a metric for *efficiency*, which measures how many of the performed moves were necessary (see Appendix D.4 for the full definition); *question answering*, which measures the percentage of questions answered correctly (in the variant with questions); and *success*, which is 1 if the player ended the game in a success condition (indicated room found / all rooms explored), and 0 otherwise.

**Instances** Experiments on the  $EE$  version test the effect of map complexity by changing map size and connectedness. The maps can have 4, 6, or 8 rooms, whereas, in the 6 or 8-room case, we distinguish between maps with and without a cyclic path in them, yielding five experiments in total. We expect larger and more connected maps to be harder to explore. For our  $EE-gr$  version, we reuse the three experiments on map sizes from above. The goal is to have comparable results and measure the influence of explicit graph reasoning. On the  $G2X$  version, we experiment with distances from start to target, either starting on, close to, or far from the target. The hypothesis is that finding a target room nearby is easier than finding it far away. In total,

Model	MatchIt					Reference		Map Game	
	<i>clemscore</i>	<i>avg %p</i>	<i>avg ql</i>						
<b>Claude-3.5</b>	<b>80.77</b>	95.33	<b>84.73</b>	100.0	85.0	100.0	81.03	92.22	85.88
GPT-4o (Aug)	80.04	96.93	82.57	93.33	80.36	100.0	74.87	97.11	85.87
GPT-4-1106	73.55	97.79	75.21	100.0	80.0	98.97	68.39	96.67	75.89
GPT-4o (May)	69.56	87.73	79.29	100.0	78.33	100.0	75.38	79.56	80.91
Claude-3-opus	68.16	<b>99.33</b>	68.62	100.0	81.67	100.0	47.18	98.89	71.41
GPT-4o-mini	58.46	90.04	64.93	100.0	86.67	98.21	48.04	84.0	63.32
Gemini-1.5-flash	47.73	85.0	56.15	85.0	84.31	100.0	41.54	80.0	51.64
<b>InternVL2-26B</b>	<b>37.45</b>	<b>66.76</b>	56.09	100.0	93.33	85.13	34.34	49.56	50.93
InternVL2-76B	33.84	54.8	61.76	100.0	90.0	100.0	34.36	24.67	61.48
InternVL2-40B	32.23	56.27	57.28	96.67	79.31	100.0	36.15	28.22	56.97
Idefics-80B	29.55	58.29	50.7	88.14	55.77	100.0	33.59	34.44	54.71
Pixtral-12B	28.64	49.98	57.3	100.0	63.33	79.23	44.66	23.55	59.51
InternVL2-8B	23.17	46.61	49.7	100.0	68.33	86.41	37.09	15.55	0
Idefics3-8B	17.52	32.59	53.76	40.0	79.17	98.97	31.09	8.0	0
InternLM-XC	16.95	20.18	<b>83.98</b>	98.33	77.97	2.56	90.0	0.0	0
Phi-3.5-vision	15.64	40.67	38.46	100.0	0.0	100.0	15.38	1.11	0
Idefics-9B	12.29	38.0	32.34	100.0	33.33	90.0	31.34	0.0	0
Phi-3-vision	3.34	5.06	65.98	0.0	0	17.95	100.0	2.44	0

Table 1: The “clemscore” is calculated as  $(\text{avg } \%p * \text{avg } ql) / 100$  where *avg %p* (average played) is the average percentage of games played to completion, and *avg ql* (average quality score) is the measure of quality of the completed games. Results for Map Game are averaged over three variants of the game. The highest *clemscore*, *avg played* and *quality scores* for commercial and open-weight models and are highlighted in blue and teal, respectively.

there are five experiments with 50 instances for *EE*, three experiments with 30 instances for *EE-gr*, three experiments with 30 instances for *G2X*.

## 5 Results

### 5.1 Overall Results

**Models:** We selected models that i) support multi-turn dialogue and have been optimised to follow chat templates,<sup>4</sup> ii) encode multiple images in a single turn. We benchmarked both open-weight and commercial models. Of commercial models, we decided to evaluate Claude-3.5-Sonnet (June 2024), Claude-3-Opus (February 2024), GPT-4-vision (November 2023), GPT-4o (May & August 2024 versions), GPT-4o-mini (July 2024), and Gemini-1.5-Flash-001 (May 2024).<sup>5</sup> From the available open-weight models we selected InternVL2 (8B, 26B, 40B, 76B versions) (Chen et al., 2023), Idefics (9B, 80B versions) (Laurençon et al., 2023), Idefics-3 (8B-llama) (Laurençon et al., 2024), InternLM-XComposer-2.5 (Zhang et al., 2024), Phi-vision (3.0, 3.5 versions) (Abdin et al., 2024), Pixtral-12B (2409)<sup>6</sup>. We provide more details about models in Appendix A.

The benchmark results are given in Table 1. What first catches the eye is the significant difference in overall score (*clemscore*) between closed-

<sup>4</sup>[https://huggingface.co/docs/transformers/en/chat\\_templating](https://huggingface.co/docs/transformers/en/chat_templating)

<sup>5</sup>We excluded Gemini 1.5 Pro because querying the API backend resulted in many experiments being timed out, and Gemini 1.0 Pro was excluded since it does not support multi-turn dialogue.

<sup>6</sup><https://huggingface.co/mistralai/Pixtral-12B-2409>

weight / commercial and open-weight models, with the best open model trailing the worst commercial for 10 points and the best commercial one for 43 points. We can compare this to the situation with text-only games, where Beyer et al. (2024) report that the best/best distance was 55.25 points in June 2023, 41.18 five months later (November 2023), and in May 2024 was reduced to 24.94. This nicely reflects the somewhat less mature state of LMMs (large multimodal models) compared to LLMs.

What is also striking is that *% played (p)*, which measures the ability of the models to follow formatting instructions, is generally high; indicating that the scaffolding offered by the GameMaster was strong, but also perhaps that indeed these models are well tuned. We can also see that, in particular, the performance on the *Reference Game* seems to be a differentiator between models; while the commercial models are all in the same level on *MatchIt*, they differ more there (and to a lesser degree also on the *Map Navigation Games*). Overall, the Claude-3.5-sonnet and GPT-4o (Aug), which increased 10 points compared to the May 2024 version, are the best performing commercial models, and the InternVL2 models are the best performing open-weight models.

To investigate further, we turn to a more fine-grained analysis by implementing text-only variants of games.

### 5.2 Textual vs. Multimodal Performance

This section analyses the effect of moving from text-only LLMs to multimodal ones. We implemented text-only versions of three games by representing the tasks in ASCII characters. Each game

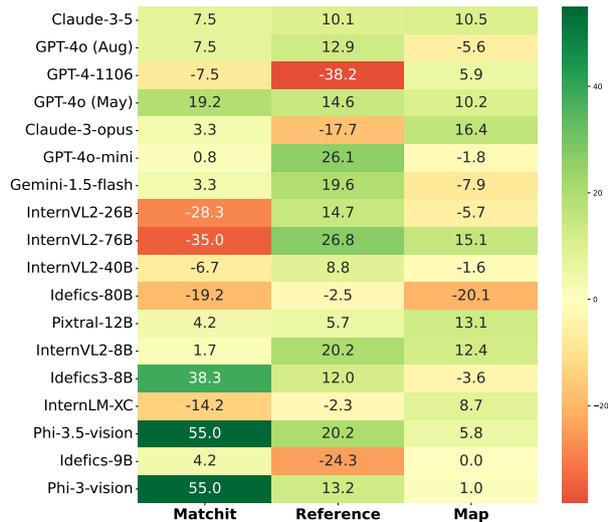


Figure 5: Performance difference in *clemscore* (textual - multimodal) across models and games. **Green** values indicate better textual and **red** values (negative) indicate better multimodal performance. The values closer to zero (in **yellow**) indicate that the performance of models are somewhat equal between modality variants.

has been implemented where inputs are represented in only text. For the Reference game, we ran the original ASCII character representation of grids (as in *clembench* (Chalamalasetti et al., 2023)). For the Matchit game, we used the same ASCII grids (Figure 12) to create similar/dissimilar experiments.

For the Map Navigation game, we implemented all three versions in text-only variants as follows: once the Player makes a move, the GameMaster provides information about the current room in text format, such as “You have entered Nursery. From here you can go north, south, east”. In the multimodal version, this information is given as the input image (e.g. Nursery) and then the text “From here you can go north, south, east” (without any information about the room in text form).

Next, we ran the benchmark on the models using textual versions of games and compared them against the multimodal results. Figure 5 shows the difference in *clemscore* between textual and multimodal scores, where we subtracted the multimodal value from textual one. Higher values (in green) indicate that models are better at textual, while lower values (in red) stand for better performance at multimodal games. In general, we can observe that most models are better at textual games; which perhaps can be explained by the dominance of text data in training datasets over other modalities (images in this case) (He et al., 2024). The commercial models such as GPT-4o (Aug’24) and Claude-3.5 (be-

ing the best two models in multimodal games) are also better at textual versions of the games while GPT-4-1106 is worse at the textual version of the Reference game. From the open-weight models, InternVL2-26B has the best score in multimodal games but clearly struggled with a textual version of the MatchIt game. We can also observe that InternVL2-40B is a better choice over 26B version (or any other open-weight model) as its performance is equally distributed across games for their textual and multimodal versions. The Map Navigation Game has steady performance of almost all models (except Idefics-80B) being better at textual variants than multimodal ones.

### 5.3 Zooming in on the Games

In this section, we discuss the individual findings across games by mentioning the hypothesis (H) and the finding (F).

#### 5.3.1 The MatchIt Game

The results breakdown in detail is in Appendix C.

**H:** Pairs of similar images pose the biggest challenge (shown for human players in a similar setting by Sagi et al. (2012)).

**F:** Figure 14 shows that this bears out for the image-based instances. This is less clear for the text-based instances.

**H:** Pairs of very different images will be the easiest to recognize (as being different) because the initial description might already make clear the incompatibility.

**F:** This has not been shown. Although for the text-based game variant, the highest scores have been achieved in with different grids, the difference to same grids is not significant. Through all versions of multimodal inputs, the highest scores are achieved in the “same image” case. This further indicates that the followed strategy relies on comparative reasoning to a lesser degree than anticipated.

#### 5.3.2 The Reference Game

See Appendix B for detailed analysis.

**H:** Due to naming difficulties, the task is harder for more abstract images (grids, pentomino pieces) than photos of common scenes (ADE20K, DOCCI).

**F:** Table 4 includes detailed results for each individual experiment. GPT models & Claude-3.5 (but not Gemini) get much higher scores on ADE, DOCCI, and CLEVR experiments than on grid and

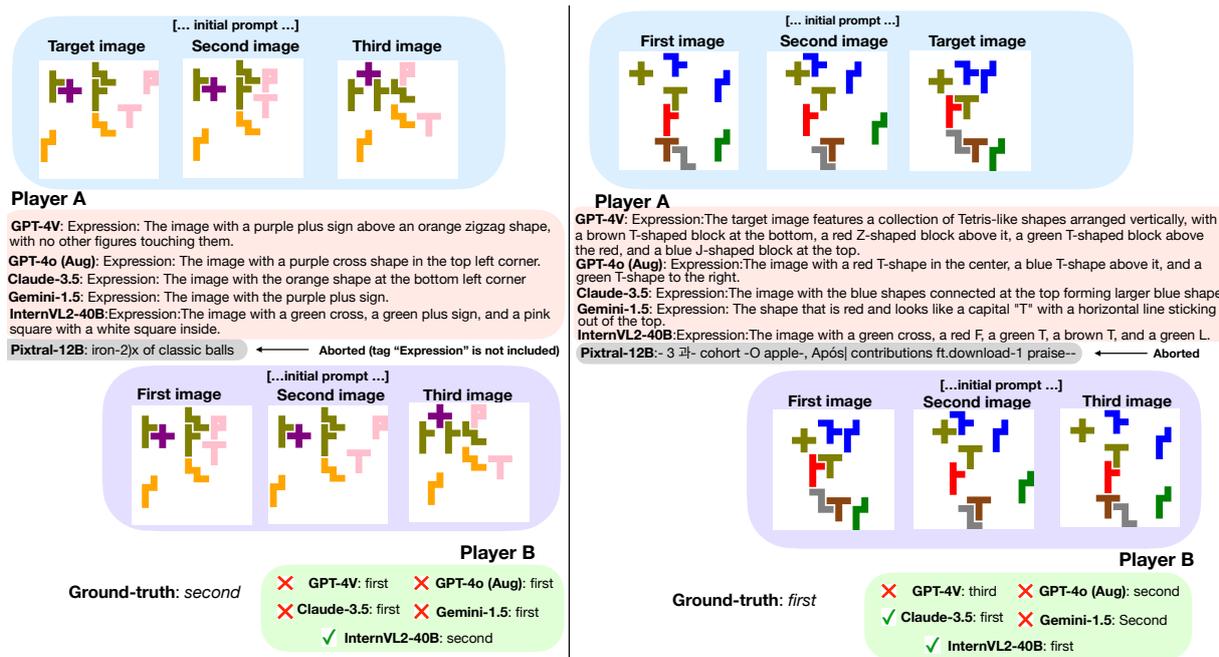


Figure 6: Sample outputs generated by the models for two Pentomino experiments. The example on the left has the target in the *first* position for *Player A* while the right example has in the *third* position. The order of images for the *Player B* is shuffled.

pentomino experiments. The pentomino experiment results (bad across the board) show that the task is far from being solved. We speculate that this set might touch the limits of the vision encoder and its ability to distinguish objects (of usual kinds). See Figures 6 for sample outputs for the Pentomino experiment. (see Figure 9 and 10 for ADE, CLEVER experiments).

**H:** Random images are more challenging to describe than patterns and objects.

**F:** Indeed, the results in Table 4 show that most models struggled with random grids for both textual and multimodal variants of the game. The same difference can be observed for photo images vs. random collections of pentomino pieces.

**H:** Given that the base models from which the models were trained are text models, even the resulting models perform better on the text-only renderings of grid experiments than the image ones.

**F:** Results are mixed. Some models are better at ASCII representations (Gemini-1.5, GPT-4o), while others are better at multimodal representations (Claude-3, GPT-4V). See Figures 9 and 11 for sample outputs.

**H:** To reach high scores in this game, player A needs to do Referring Expression Generation (REG; Gatt and Kraemer (2018)), as opposed to captioning.

**F:** We were initially surprised by the high scores achieved, in particular by the GPT-4s. On inspecting the transcripts, it became clear that the model achieves its high performance in parts through its exceptional ability to produce detailed descriptions (especially for the photo sets), thereby reaching a level of detail where a description of the target itself is enough to single it out (see also Appendix B.4 for an ablation on the (missing) effect of the distractors). This is also evident from the average number of generated tokens, which is 27 for GPT-4V and 20 for GPT-4o, as compared to 14 for Gemini and Claude. We also find little evidence for the use of negations (“the one without cars”), which can be an efficient REG strategy (although Claude does produce this occasionally). As mentioned above, performance breaks down for the pentomino dataset. Overall, this suggests that the game, as currently defined, leans more towards evaluating deep captioning (where there is still room to grow).

### 5.3.3 The Map Navigation Games

As can be seen in Table 1 above, of the three games, performance is lowest on the Map Navigation Games, showing an especially pronounced gap between the commercial and the open-weight models, which are only able to finish a much smaller percentage of games (% played). A detailed results breakdown is in Appendix D. Samples are

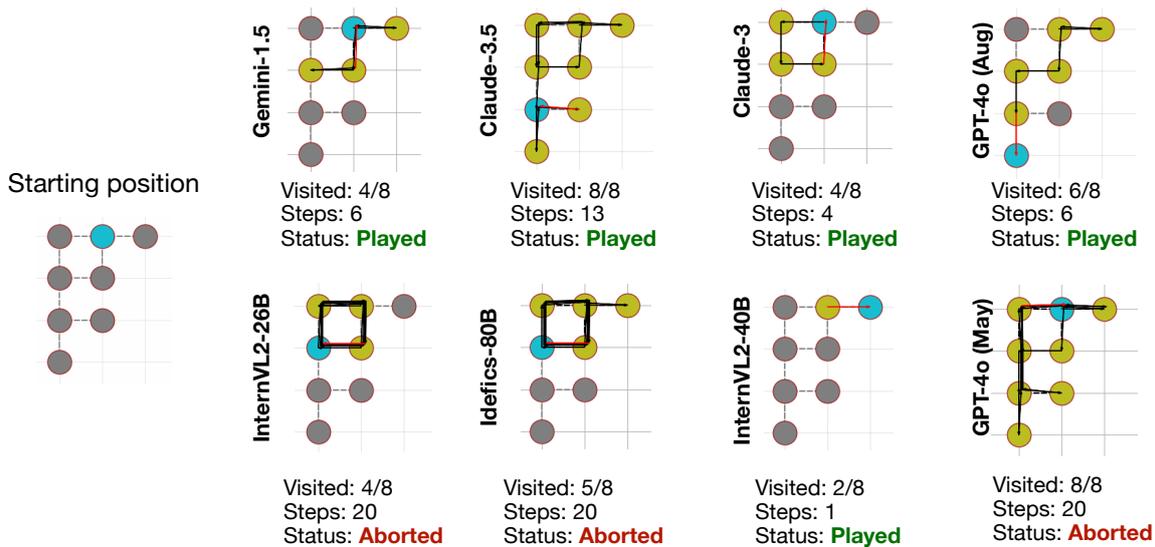


Figure 7: Map navigation samples by selected models for the experiment “Large with cycle”. The currently visited room is marked as cyan, rooms that have been visited are olive colored and the gray rooms have not been visited yet. The game is *Played* when a model decides to stop on its own by generating “DONE”. The game is *Aborted* if a model generates an output that does not comply with formatting or if the number of turns reaches the maximum limit of 20. The number of visited nodes and total steps are also given for each model.

also given in Figure 7, which shows how open-access models (InternVL2-26B, Idefics-80B) reach the maximum turn limit because they enter a loop.

**H:** A larger map makes exhaustive exploration more difficult. There is a higher chance of missing something if there are more things to discover.

**F:** Holds true for smaller models. Looking at the results in Table 11, a slight downward trend is noticeable. However, some models appear to be behaving differently, showing better results on medium or larger maps compared to small maps. This is likely due to their thoroughness when it comes to exploration. Models make more steps than the number of nodes in the map, e.g. GPTs tend to take more redundant steps. The ratio of redundant exploration to useful exploration decreases with larger map sizes, leading to higher scores.

**H:** A more complex map layout (w/ cycles) is harder to navigate.

**F:** Table 11 seem to indicate that this holds true. While there is only a marginal difference between maps of medium size with and without cycles, the difference becomes more apparent with large maps.

**H:** In *G2X* (go to specific room), the further away from the starting position the target is, the harder it is to identify it, as more exploration is needed and distractor categories might be encountered.

**F:** The results in Table 14 show a clear correlation between distance and success. Not a single model could accurately find every target room at a

distance of three or more.

Overall, we take these findings as an indication that the game posed a significant challenge to the models, and that successful completion requires sophisticated representational and spatial reasoning abilities.

## 6 Conclusions

We have transferred a recent evaluation paradigm—game-based evaluation—from the text-only domain to the evaluation of multimodal models. We have set up a set of games that challenge, in different ways, mostly a model’s capability to represent (and describe) a *situation*. We have systematically varied the complexity of these situations, as well as how they are given to the model (where we have included, for comparison, purely text-based renderings). We argue that the results on the benchmark are a valid measurement of (aspects of) specific underlying capabilities, which static benchmarks do not address. We observe a large difference in performance between the largest commercial models and the smaller open-weight models, albeit to a smaller degree than other researchers have observed in the early stages of text-only models. The benchmarks indicate that there is room to grow both for the closed and the open models, while there already is a basis for the development of new kinds of situated interactive systems.

## Acknowledgments

The work reported here has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grants 423217434 (“RECOLAGE”) and 317633480 (SFB 1287); and by Bundesministerium für Bildung und Forschung (BMBF, German Federal Ministry of Research), project “COCOBOTS” (01IS21102A). We thank the anonymous reviewers for their helpful feedback.

## Limitations

The first and biggest limitation is that the prompts that define the games are only given in, and hence the results are restricted to, English, even though several of the tested models are listed as being able to process other languages as well. While we have yet to do this, translating the prompts and measuring their impact should be straightforward; we plan to do this in future work.

As discussed in the text above, some of the findings are limited to certain respect by the fact that excellent capabilities of providing image captions open up simpler strategies than what we initially wanted to challenge. While this doesn’t impact the significance of the measurements—there is still room to grow, clearly so for the open weight models, but also for the closed one—it should again be straightforward to modify the games so that interactional phenomena (such as valuing efficiency in producing referring expressions in the *Reference Game*, and putting weight on the questioning in the *MatchIt Game*) are further emphasised. Similarly, the amount of scaffolding provided by the GameMaster is quite high (e.g., in the *MatchIt* game, it determines much of the strategy), which limits the amount to which we gain insight into the strategic abilities of the models. But again, reducing it in future versions of the game should be straightforward.

## Ethics Statement

Using paid proprietary APIs with underlying models about which little is known (training data, model architecture) in academic research is less than ideal. At the moment, the models tested here seem to be the only ones that are able to follow the structure of the games. It is our hope that open models will catch up soon on multimodal tasks, and proper research can be done with them.

## References

- Marah I Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, and et al. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *CoRR*, abs/2404.14219.
- Anne Beyer, Kranti Chalamalasetti, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2024. [clembench<sub>2024</sub>: A challenging, dynamic, complementary, multilingual benchmark and underlying flexible framework for llms as multi-action agents](#). *Preprint*, arXiv:2405.20859.
- Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with GPT-4](#). *CoRR*, abs/2303.12712.
- Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. [clembench: Using game play to evaluate chat-optimized language models as conversational agents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. [Chateval: Towards better llm-based evaluators through multi-agent debate](#). *CoRR*, abs/2308.07201.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. [Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). *arXiv preprint arXiv:2312.14238*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#). *Preprint*, arXiv:2403.04132.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. [Visual dialog](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. [Guesswhat?! visual object discovery through multi-modal dialogue](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4466–4475. IEEE Computer Society.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). *CoRR*, abs/2306.13394.
- Albert Gatt and Emiel Kraemer. 2018. [Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation](#). *Journal of Artificial Intelligence Research*, 61:65–170.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.
- Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. [Vision-and-language navigation: A survey of tasks, methods, and future directions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland. Association for Computational Linguistics.
- Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. [Vizwiz grand challenge: Answering visual questions from blind people](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.
- Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. [The PhotoBook dataset: Building common ground through visually-grounded dialogue](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.
- Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, Jifeng Dai, Yong Zhang, Wei Xue, Qifeng Liu, Yike Guo, and Qifeng Chen. 2024. [LLMs meet multimodal generation and editing: A survey](#). *CoRR*, abs/2405.19334.
- Nikolai Illykh, Sina Zarriëß, and David Schlangen. 2019. [Meet up! a corpus of joint activity dialogues in a visual environment](#). In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. [CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.
- Robert M. Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:266–278.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *Int. J. Comput. Vis.*, 123(1):32–73.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. [Obelics: An open web-scale filtered dataset of interleaved image-text documents](#). *Preprint*, arXiv:2306.16527.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. [What matters when building vision-language models?](#) *Preprint*, arXiv:2405.02246.
- David Lewis. 1969. *Convention*. Harvard University Press.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. [Seed-bench-2: Benchmarking multimodal large language models](#). *CoRR*, abs/2311.17092.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. [Seed-bench: Benchmarking multimodal llms with generative comprehension](#). *CoRR*, abs/2307.16125.
- Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulic. 2024. [Topviewr: Vision-language models as top-view spatial reasoners](#). *CoRR*, abs/2406.02537.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, and *alia*. 2022. [Holistic evaluation of language models](#). *CoRR*, abs/2211.09110.
- Chang Liu and Bo Wu. 2023. [Evaluating large language models on graphs: Performance insights and comparative analysis](#). *Preprint*, arXiv:2308.11224.

- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. [Mmbench: Is your multi-modal model an all-around player?](#) *CoRR*, abs/2307.06281.
- José Lopes, Nils Hemmingsson, and Oliver Åstrand. 2018. [The spot the difference corpus: a multi-modal corpus of spontaneous task oriented spoken interactions](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.
- Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frureri, Hiteshi Sharma, Microsoft Research Redmond, sharma Robert Osazuwa, Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. 2023. [Evaluating cognitive maps in large language models with cogeval: No emergent planning](#). In *Proceedings of NeurIPS 2023*.
- Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. [Docci: Descriptions of connected and contrasting images](#). *Preprint*, arXiv:2404.19753.
- Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. [Gameeval: Evaluating llms on conversational games](#). *CoRR*, abs/2308.10032.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Md Imbesat Hassan Rizvi, Xiaodan Zhu, and Iryna Gurevych. 2024. [Sparc and sparp: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models](#). *CoRR*, abs/2406.04566.
- Philipp Sadler, Sherzod Hakimov, and David Schlangen. 2024. [Sharing the cost of success: A game for evaluating and learning collaborative multi-agent instruction giving and following policies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14770–14783, Torino, Italia. ELRA and ICCL.
- Amir Sadovnik, Yi-I Chiu, Noah Snaveley, Shimon Edelman, and Tsuhan Chen. 2012. [Image description with a goal: Building efficient discriminating expressions for images](#). In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 2791–2798. IEEE Computer Society.
- Eyal Sagi, Dedre Gentner, and Andrew M. Lovett. 2012. [What difference reveals about similarity](#). *Cogn. Sci.*, 36(6):1019–1050.
- David Schlangen. 2019. [Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings](#). *CoRR*, abs/1908.11279.
- David Schlangen. 2023a. [Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy](#). *CoRR*, abs/2304.07007.
- David Schlangen. 2023b. [On general language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8818–8825, Singapore. Association for Computational Linguistics.
- David Schlangen. 2023c. [What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis](#). *CoRR*, abs/2302.08590.
- Judy Hanwen Shen, Matthias Hofer, Bjarke Felbo, and Roger Levy. 2018. [Comparing models of associative meaning: An empirical investigation of reference in simple language games](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 292–301, Brussels, Belgium. Association for Computational Linguistics.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. [Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11321–11329. AAAI Press.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, and et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *CoRR*, abs/2206.04615.
- Alessandro Suglia, Ioannis Konstas, and Oliver Lemon. 2024. [Visually Grounded Language Learning: a review of language games, datasets, tasks, and models](#). *Journal of Artificial Intelligence Research*, 79:173–239.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy,

and Samuel R. Bowman. 2019. [SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems](#). In *NeurIPS*, July, pages 1–30.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. [Can language models solve graph problems in natural language?](#) *Preprint*, arXiv:2305.10037.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, and et al. 2023. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). *CoRR*, abs/2311.16502.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, and et al. 2024. [Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output](#). *arXiv preprint arXiv:2407.03320*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. [Scene parsing through ADE20K dataset](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5122–5130. IEEE Computer Society.

## A Model Evaluation Details

Model	Base Language Model		Base Image Processor	
	Name	Parameters	Name	Parameters
Idefics-80B	LLaMA	65B	laion/CLIP-ViT	630M
Idefics-9B	LLaMA	7B	laion/CLIP-ViT	630M
InternVL2-76B	Hermes2-Llama3	70B	InternViT-V1-5	6B
InternVL2-40B	Hermes2-Yi	34B	InternViT-V1-5	6B
InternVL2-26B	Internlm2	20B	InternViT-V1-5	6B
InternVL2-8B	Internlm2_5	7B	InternViT-448px	300M
Idefics3-8B	Llama-3.1	8B	SigLIP	400M
Internlm-XC	InternLM2	7B	ViT-L/14	428M
Phi-3.5-vision	Phi-3.5	3.8B	Phi3VProcessor	400M
Phi-3-vision	Phi-3-mini	3.8B	Phi3VProcessor	400M
Pixtral-12B	Nemo-12B	12B	-	400M

Table 2: Internal details for open-weight models

This section provides a detailed view of the models present in our primary results, along with their utilization. As outlined in the main text, seven commercial (Claude-3.5, Claude3, Gemini-1.5-Flash, GPT-4o (Aug’24), GPT-4o (May’24), GPT-4o-mini, GPT-4-1106) and 11 open-weight models (Idefics-9B, Idefics-80B, Idefics-3-8B, InternVL2-76B, InternVL2-40B, InternVL2-26B, InternVL2-8B, Internlm-XC, Phi-3.5-vision, Phi-3-vision, Pixtral-12B) are included in the primary results. Detailed information about these models can be found in Table 3.

### A.1 Image resolution limits

Image resolution in Table 3 defines the maximum scaled-down resolution during image processing. For Claude-3,<sup>7</sup> though the maximum image resolution is  $1568 \times 1568$  pixels, it does allow images up to  $8k \times 8k$  which are then scaled down to 1568. If any edge of the image exceeds 8k pixels, the model rejects the image. Another constraint specified by Anthropic is the token limit, capped at 1600 tokens per image. Other commercial models only specify the maximum scaled-down resolution without specifying explicit pixel or token constraints. Considering open-weight models, these models specify just one image resolution, and if any image exceeds this, it will be scaled down.

### A.2 Compute Details

The commercial models were used by integrating their APIs with the *clembench* backend,<sup>8</sup>. For open-weight models, inferencing was conducted on a local cluster, which comprises four Nvidia A100 (80GB) GPUs. The model weights were distributed

<sup>7</sup><https://docs.anthropic.com/en/docs/vision>

<sup>8</sup><https://github.com/clp-research/clembench/tree/main/backends>

Model Name	Parameters	Context length	Image resolution (pixels)	Release date	Commercial	Backend	Training data	Cut-off date
Claude-3.5-sonnet	-	200K	1568x1568	Jun 2024	✓	anthropic	Aug 2023	
Claude-3-opus	2T*	200K	1568x1568	Mar 2024	✓	anthropic	Aug 2023	
Gemini-1.5-flash	-	1048K	-	Apr 2024	✓	google	Nov 2023	
Gpt-4o (May)	1.76T*	128K	768x2000	May 2024	✓	openai	Oct 2023	
Gpt-4o (August)	1.76T*	128K	768x2000	Aug 2024	✓	openai	Oct 2023	
Gpt-4o-mini	-	128K	768x2000	Jul 2024	✓	openai	Oct 2023	
Gpt-4V-1106	1.76T*	128K	768x2000	Nov 2023	✓	openai	Apr 2023	
Idefics-80B	80B	2K	224x224	Aug 2023	×	huggingface	Feb 2023	
Idefics-9B	9B	2K	224x224	Aug 2023	×	huggingface	Feb 2023	
Pixtral-12B	12B	128K	1024x1024	Sep 2024	×	huggingface	-	
InternVL2-76B	76B	8K	448x448	Aug 2023	×	huggingface	Feb 2023	
InternVL2-40B	40B	8K	448x448	Aug 2023	×	huggingface	Feb 2023	
InternVL2-26B	26B	32K	448x448	Aug 2023	×	huggingface	Feb 2023	
InternVL2-8B	8B	32K	448x448	Aug 2023	×	huggingface	Feb 2023	
Idefics3-8B-llama	8B	128K	384x384	Aug 2024	×	huggingface	-	
Internlm-XC	7B	24K	224x224	Jul 2024	×	huggingface	-	
Phi-3.5-vision	4B	128K	1344x1344	Aug 2024	×	huggingface	Aug 2024	
Phi-3-vision	4B	128K	1344x1344	May 2024	×	huggingface	Apr 2024	

Table 3: Commercial and open-weight model details. Image resolution - indicates the maximum resolution of the image including any scaling. Supports multiple images - indicates whether the model can process multiple images in a single turn, such as in multimodal reference game. Backend - specifies the underlying script used to access the model for gameplay. Dashes (-) denote information that is not publicly available. \* denotes estimated parameter size.

evenly on each GPU, and the models used their default precision values. No weights were offloaded to the CPU during inference. The open-weight models considered here are their HuggingFace-compatible versions, loaded via Auto Classes methods to maintain generalizability, which leads to straightforward integration of additional models with minimal changes required<sup>9</sup>.

### A.3 Internal Details of Models

The internal details of open-weight models are described in this section. The internal information about the base language model and base image processor is available in Table 2. To train the Idefics models, the authors developed their own dataset - Obelics (Laurençon et al., 2023). This dataset is based on multimodal web documents, so a single sample of this dataset contains multiple images, making these models suitable for multimodal reference game runs.

## B A Picture Reference Game

### B.1 Prompt Templates

The prompt template for both players of the Reference Game is given in Figure 8.

### B.2 Overall Results

Table 4 displays the detailed results for the different Reference Game experiments.

### B.3 Qualitative Samples

Here, we provide example instances and the respective model responses for the pentomino experiments (Figure 6), the CLEVR and ADE experiments (Figure 10), and the multimodal (Figure 9) and the textual (Figure 11) grid experiments.

### B.4 Static Target Image

In order to understand the effect of the distractors on the generated expression, we created another set of instances from the ADE, CLEVR, and DOCCI datasets where the target image is kept the same for all instances, and the distractors are chosen from similar images in the respective datasets. The distractors from the ADE dataset were chosen from the same scene category “bedroom”. For the DOCCI instances, we used the concept category “dog” to select distractor images from the same category. The instances from the CLEVR dataset were chosen based on the object annotations (large metal green sphere, small rubber blue cube, etc.). The images with the most objects in common with the target were taken as distractors.

The results are given in Table 5. We can observe that both GPT-4 models are not susceptible to the change in instances while being the best performing models for all datasets. Comparing the *Claude-3* results in Table 4 and Table 5, we can observe that the performance drops for DOCCI and ADE instances with static target.

We were initially hoping that we can automatically test for context-sensitivity (and hence, the output really being a referring expression more than

<sup>9</sup>[https://huggingface.co/docs/transformers/model\\_doc/auto](https://huggingface.co/docs/transformers/model_doc/auto)

**TEMPLATE B.1.1**

You are given three images, one is called target and the other two are distractors. Your task is to generate a referring expression that best describes the target image while distinguishing it from the two other distractor images. The first image is <IMAGE\_POSITION>, the second image is <IMAGE\_POSITION>, and the third image is <IMAGE\_POSITION>.

Instruction: Describe the target image. Generate the referring expression starting with the tag "Expression: " for the given target image. Omit any other text.

Target image: <IMAGE\_PATH>  
 Second image: <IMAGE\_PATH>  
 Third image: <IMAGE\_PATH>

**TEMPLATE B.1.3**

You are given three images. You are also given a referring expression that describes one of the given images. Your task is to select the image that matches the given referring expression. Generate only the number (in text) of the image that matches the given expression by selecting first, second, or third.

TARGET\_EXPRESSION  
 Question: Which image does the expression refer to? Start with the tag "Answer: ", followed by your selection. Omit any other text.

First image: <IMAGE\_PATH>  
 Second image: <IMAGE\_PATH>  
 Third image: <IMAGE\_PATH>

**TEMPLATE B.1.2**  
 Expression: \$EXPRESSION\$

**TEMPLATE B.1.4**  
 Answer: \$ANSWER\$

(a) Prompt template for Player A (Instruction Giver) in the Reference Game.

(b) Prompt template for Player B (Instruction Follower) in the Reference Game

Figure 8: Reference game prompt templates for both players

**Player A**

**Player B**

**Multimodal Row**  
 Ground-truth: first

- ✓ GPT-4V: first
- ✓ GPT-4o: first
- ✓ Claude-3: First
- ✓ Gemini-1.5: first
- ✗ Idefics-80B: third
- ✗ Idefics-9B: third

**Multimodal Random Grid**  
 Ground-truth: second

- ✗ GPT-4V: first
- ✗ GPT-4o: second
- ✗ Claude-3: third
- ✗ Gemini-1.5: first
- ✓ Idefics-80B: 2
- ✗ Idefics-9B: 1

Figure 9: Sample outputs generated by the models for multimodal row and random grid experiments

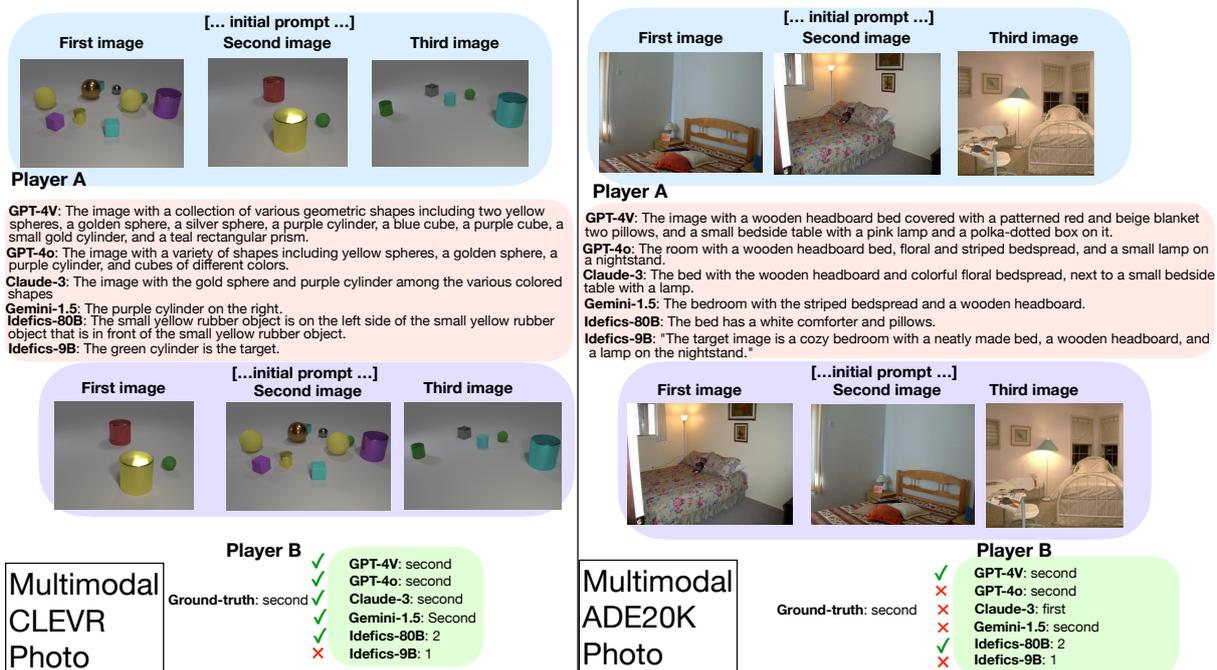


Figure 10: Sample outputs generated by the models for the CLEVR and ADE experiments.

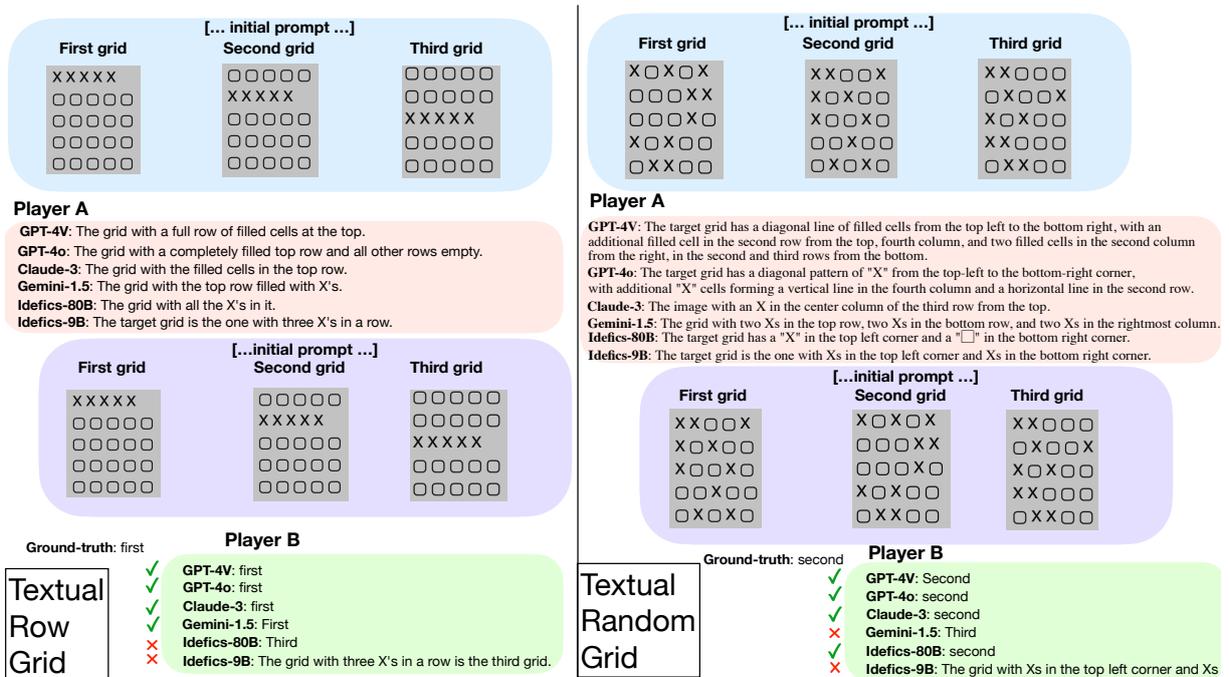


Figure 11: Sample outputs generated by the models for textual row and random grid experiments.

Text-only Reference Game						
Models	Row	Column	Diagonal	Letter	Shape	Random
InternVL2-26B	53.3	46.7	56.7	30.0	40.0	36.7
InternVL2-40B	56.7	33.3	60.0	46.7	40.0	33.3
InternVL2-76B	76.7	86.7	43.3	50.0	60.0	50.0
Phi-3-vision	40.0	30.0	26.7	20.0	33.3	36.7
Phi-3.5-vision	40.0	10.0	46.7	36.7	53.3	26.7
Pixtral-12B	30.0	36.7	66.7	43.3	43.3	26.7
Claude-3-5	<b>100.0</b>	<b>100.0</b>	<b>93.3</b>	<b>86.7</b>	90.0	76.7
Claude-3-opus	23.3	26.7	33.3	36.7	33.3	23.3
Gemini-1.5	76.7	76.7	63.3	43.3	60.0	46.7
GPT-4-1106	26.7	33.3	33.3	30.0	33.3	20.0
GPT-4o (May)	96.7	<b>100.0</b>	90.0	70.0	<b>93.3</b>	<b>90.0</b>
GPT-4o (Aug)	90.0	<b>100.0</b>	<b>93.3</b>	76.7	80.0	86.7
GPT-4o-mini	76.7	90.0	80.0	63.3	73.3	56.7
Idefics-80B	33.3	33.3	33.3	26.7	23.3	36.7
Idefics-9B	16.7	0.0	0.0	0.0	6.7	0.0
InternLM-XC	0.0	0.0	0.0	0.0	0.0	0.0

Multimodal Reference Game										
	Row	Column	Diagonal	Letter	Shape	Random	ADE	DOCCI	CLEVR	Pent.
InternVL2-26B	33.3	33.3	40.0	33.3	33.3	33.3	33.3	30.0	3.3	33.3
InternVL2-40B	30.0	26.7	36.7	30.0	40.0	30.0	33.3	30.0	60.0	<b>43.3</b>
InternVL2-76B	33.3	30.0	33.3	30.0	33.3	30.0	40.0	30.0	33.3	20.0
Phi-3-vision	0.0	0.0	0.0	0.0	0.0	0.0	33.3	33.3	33.3	33.3
Phi-3.5-vision	33.3	33.3	33.3	33.3	33.3	33.3	0.0	0.0	0.0	0.0
Pixtral-12B	30.0	43.3	26.7	43.3	23.3	30.0	73.3	3.3	63.3	0.0
Claude-3-5	<b>76.7</b>	<b>90.0</b>	<b>93.3</b>	56.7	<b>80.0</b>	<b>76.7</b>	83.3	96.7	<b>100.0</b>	33.3
Claude-3-opus	63.3	36.7	60.0	46.7	33.3	40.0	50.0	70.0	66.7	30.0
Gemini-1.5	56.7	53.3	56.7	30.0	43.3	36.7	40.0	53.3	33.3	30.0
GPT-4-1106	43.3	43.3	66.7	33.3	46.7	40.0	93.3	96.7	90.0	33.3
GPT-4o (May)	56.7	63.3	76.7	<b>63.3</b>	50.0	50.0	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	30.0
GPT-4o (Aug)	63.3	53.3	76.7	60.0	53.3	46.7	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	26.7
GPT-4o-mini	46.7	23.3	60.0	33.3	36.7	40.0	56.7	43.3	83.3	23.3
Idefics-80B	36.7	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3	36.7
Idefics-9B	30.0	30.0	33.3	33.3	33.3	33.3	26.7	20.0	33.3	13.3
InternLM-XC	0.0	0.0	0.0	0.0	0.0	0.0	6.7	3.3	0.0	0.0

Table 4: Average success scores across experiments for textual and multimodal reference games. Random performance is at 33.3 (as there are three images to choose from). Note: Only the grid experiments are used in the text-only version, as the others are actual images and cannot easily be represented in ASCII format. *N/A*: not available means that the experiment was not played due to the model not following the formal instructions. The best results for each experiment are highlighted in bold. **Pent**: Pentomino pieces

a description) by pairing the target image in different distractors and seeing whether the generation changes. However, it turns out that the observed context sensitivity is of the wrong kind, as the generated expression also changes if the target image and the same distractors simply are presented in different orders. Hence, we have no grounds to assume that the generated expressions are more than very detailed descriptions.

## C An Agreement Game: MatchIt

### C.1 Text-only (ASCII) version

The base grids used for the text based ASCII variant of MatchIt consist of the *diagonal*, *letter* and *shape* grids used in the Reference game. Besides pairs of same and different grids, we used two sets of instances with similar grid pairs. One set of instances

was created by either mirroring the grids vertically or horizontally or by turning them 90 degrees and another set of instances consisted of grids paired with new grids with edit distance of two, where the symbol was inverted at two random positions of the grid. Figure 12 shows the modifications for similar grid pairs. For each of the difficulties (same, similar with transformed motive, similar with edit distance of two and different grids), 10 instances were part of the final game play, for a total of 40 instances.

### C.2 Experimental Setup

**Game rules** Importantly, each utterance has to start with the right flag such as ‘DESCRIPTION’, ‘QUESTION’, ‘ANSWER’ and ‘DECISION’ or the game will be aborted. Also, explanations for

Model	ADE	DOCCI	CLEVR
InternVL2-26B	36.7	36.7	0.0
InternVL2-40B	56.7	16.7	36.7
InternVL2-76B	66.7	33.3	33.3
Phi-3-vision	33.3	33.3	33.3
Phi-3.5-vision	0.0	0.0	0.0
Pixtral-12B	73.3	0.0	50.0
Claude-3-5	80.0	86.7	<b>100.0</b>
Claude-3-opus	30.0	20.0	66.7
Gemini-1.5-flash	23.3	36.7	46.7
GPT-4-1106	<b>100.0</b>	<b>100.0</b>	93.3
GPT-4o (May)	<b>100.0</b>	<b>100.0</b>	90.0
GPT-4o (Aug)	<b>100.0</b>	<b>100.0</b>	93.3
GPT-4o-mini	50.0	36.7	80.0
Idefics-80B	33.3	33.3	30.0
Idefics-9B	30.0	16.7	33.3
InternLM-XC	13.3	6.7	0.0

Table 5: Average success scores across ADE, DOCCI and CLEVR instances where the target image is kept the same for all instances and the set of distractors varies.

Original Grid					Similar Grid (1)					
X	X	X	X	X	X	X	X	X	X	
X	□	□	□	□	□	□	□	□	X	
X	X	X	□	□	□	□	X	X	X	
X	□	□	□	□	□	□	□	□	X	
X	X	X	X	X	X	X	X	X	X	
Similar Grid (2)										
	X	X	X	X	X					
	X	□	X	□	□					
	X	X	X	□	□					
	X	□	□	□	□					
	X	□	X	X	X					

Figure 12: Example grids for the matchit\_ascii version of the game

final decisions are not allowed. An example of schematic game play is shown in Figure 1.

### C.3 Prompts

All prompt templates used for the game are displayed in Figure 13

### C.4 Instances

Three groups of instance types were used for MatchIt: photographs, images of pentomino boards and grids of ASCII grids. Those instances were grouped into three (to four) difficulties; both players get the same image, both players get similar images (two different types for ASCII grids) or completely different images. The process of producing similar and different image pairs is described below. The curation rationale for a similar picture was that both pictures could be described with the same (short) sentence, but their difference should be striking enough that also one (short) sentence should be enough. All images for this multimodal variant are taken from the Visual Genome Dataset (Krishna et al., 2017), which has rich annotations

for every object in the image including attributes and relations to other objects. For each sampled image pair, the Jaccard index between all object labels and their respective attributes for each image was calculated. The lowest scoring pairs up to a threshold of 0.05 were chosen as “different” pairs. This ensures that there are no to very little shared semantic contents between the pictures. In order to get similar pairs, the cosine similarity of the image embeddings of the CLIP model (Radford et al., 2021) of image pairs with a Jaccard index above 0.22 was calculated and pairs with a cosine similarity above 0.8 chosen. Both thresholds were chosen considering quality of pairs as well as sufficient numbers of pairs. From those, final pairs were selected manually following the curation rationale mentioned above filtering out instances that were not enough or too similar.

### C.5 Results

The average main scores depending on difficulty are presented in Tables 6 and 7.

### C.6 Detailed Analysis

**Model type** There is a big jump in performance from the open source models to commercial models with the GPT and Claude models having slightly higher played and main scores than Gemini-1.5. Possible reasons for this difference are carried out in the following paragraphs.

**Modality** While the initial photo descriptions are mostly correct, the grid and pentomino board descriptions often are incorrect and/or incomplete. As seen for the reference game, the overall performance with pentomino board instances is worse than with photos, with open models sometimes not even remotely describing a picture of abstract shapes and rather trying to find a concrete meaning where there is none (e.g. describing a pentomino board as people playing soccer, people playing tic-tac-toe or as pixel art of a tree). There are two forms of grid descriptions, either in form of a row-by-row (or by column or borders) description of the placements of X’s and squares or by repeating the string of characters verbatim - a behavior we did not predict and therefore not prohibit. No instance was found where a general motive was described instead of the single parts of the grid. This is interesting since other games using grids have shown that correct grid descriptions can be made. The wording of of the prompts was adopted completely

TEMPLATE C.3.1

You are participating in a collaborative guessing game. The goal is to find out whether this picture and another picture only I can see, are the same. Please describe your image first. Then, I will provide my description and we can ask each other questions about the images to figure out whether they are the same. Now start your short image description with "DESCRIPTION:" followed by the description. Do not add anything else.

TEMPLATE C.3.3

Now ask a question in order to find out new aspects of my image that may be different to your image. Start with "QUESTION:" and do not add anything else.

TEMPLATE C.3.2

DESCRIPTION: \$DESCRIPTION\$

TEMPLATE C.3.4

QUESTION: \$QUESTION\$

(a) Initial prompt.

(b) Eliciting questions from the players.

TEMPLATE C.3.5

Start your answer with "ANSWER" and do not add anything else.

TEMPLATE C.3.6

ANSWER: \$ANSWER\$

(c) Eliciting answers from the players.

TEMPLATE C.3.7

Now come to a decision. What do you think: are your picture and the other picture described the same picture? Write "DECISION: same images" if you think they are the same picture or "DECISION: different images" if you think they are different pictures. Do not add anything else.

TEMPLATE C.3.8

DECISION: \$DECISION\$

(d) Eliciting decisions from the players.

Figure 13: Prompt templates for MatchIt.

<b>matchit (multimodal) - Photo input</b>			
	0	1	2
InternVL2-26B	90.0	90.0	<b>100.0</b>
InternVL2-40B	<b>100.0</b>	70.0	<b>100.0</b>
InternVL2-76B	<b>100.0</b>	80.0	<b>100.0</b>
Phi-3-vision	-	-	-
Phi-3.5-vision	0.0	0.0	0.0
Pixtral-12B	40.0	80.0	<b>100.0</b>
Claude-3.5	<b>100.0</b>	90.0	<b>100.0</b>
Claude-3	<b>100.0</b>	70.0	<b>100.0</b>
Gemini-1.5	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
GPT-4-1106	80.0	90.0	<b>100.0</b>
GPT-4o (May)	90.0	<b>100.0</b>	<b>100.0</b>
GPT-4o (August)	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
GPT-4o-mini	70.0	<b>100.0</b>	<b>100.0</b>
Idefics-80B	88.9	12.5	62.5
Idefics-9B	<b>100.0</b>	0.0	0.0
InternLM-XC	<b>100.0</b>	60.0	<b>100.0</b>

<b>matchit (multimodal)- Pentomino input</b>			
	0	1	2
InternVL2-26B	<b>100.0</b>	80.0	<b>100.0</b>
InternVL2-40B	<b>100.0</b>	10.0	<b>100.0</b>
InternVL2-76B	<b>100.0</b>	60.0	<b>100.0</b>
Phi-3-vision	-	-	-
Phi-3.5-vision	0.0	0.0	0.0
Pixtral-12B	0.0	60.0	<b>100.0</b>
Claude-3.5	90.0	30.0	<b>100.0</b>
Claude-3	90.0	30.0	<b>100.0</b>
Gemini-1.5	83.3	30.0	<b>100.0</b>
GPT-4-1106	40.0	70.0	<b>100.0</b>
GPT-4o-August	88.9	0.0	<b>100.0</b>
GPT-4o-May	70.0	10.0	<b>100.0</b>
GPT-4o-mini	60.0	<b>90.0</b>	<b>100.0</b>
Idefics-80B	50.0	77.8	40.0
Idefics-9B	<b>100.0</b>	0.0	0.0
InternLM-XC	60.0	60.0	90.0

Table 6: Mean main score for each model, grouped by difficulty level. 0: same image/pentomino board, 1: similar image/pentomino board, 2: different image/pentomino board

from the multimodal variant except for necessary exchanges such as image  $\leftrightarrow$  grid, so a more tailored prompt introducing the grids as such could

<b>matchit (text-only variant)</b>				
	0	1_1	1_2	2
InternVL2-26B	90.0	50.0	30.0	90.0
InternVL2-40B	<b>100.0</b>	40.0	60.0	80.0
InternVL2-76B	<b>100.0</b>	20.0	30.0	70.0
Phi-3-vision	<b>100.0</b>	85.7	42.9	<b>100.0</b>
Phi-3.5-vision	80.0	30.0	50.0	60.0
Pixtral-12B	22.2	80.0	80.0	90.0
Claude-3-5	<b>100.0</b>	80.0	90.0	<b>100.0</b>
Claude-3-opus	80.0	80.0	80.0	<b>100.0</b>
Gemini-1.5-flash	77.8	60.0	70.0	<b>100.0</b>
GPT-4-1106	70.0	70.0	50.0	<b>100.0</b>
GPT-4o (May)	90.0	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
GPT-4o (Aug)	60.0	80.0	90.0	<b>100.0</b>
GPT-4o-mini	50.0	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
Idefics-80B	<b>100.0</b>	12.5	11.1	0.0
Idefics-9B	40.0	30.0	40.0	40.0
InternLM-XC	80.0	40.0	60.0	70.0

Table 7: Mean main score for each model, grouped by difficulty level. 0: same grid, 1\_1: similar grid (motive transformation), 1\_2: similar grid (edit distance of two), 2: different grid

have maybe mitigated that. Comparing the main scores of the different modalities, for multimodal game play, larger models have an apparent advantage over their smaller counterparts that disappears when no visual input is given.

**Instance difficulties** There are significant differences between levels of difficulty. Similar images (for both modalities) get classified wrong most times, showing that the level of detail of the asked questions is not high enough to determine that there are differences in the images. The two types of similarity of the grid inputs do not seem to differ regarding the performance although the kinds of similarities between grids are quite distinct.

**General game play** Following are a few observations about the general game play based on a sample of played instances for each model and difficulty level. A large difference between smaller open models and larger as well as commercial modes are the types of questions asked. The dialogue of the latter contains more open questions that even includes some details about the player’s own image and leaves the other player room to elaborate whereas smaller models usually produced simple yes/no questions. Similar to that, some answers of the larger and commercial models are not simply a yes/no answer but explanations and com-

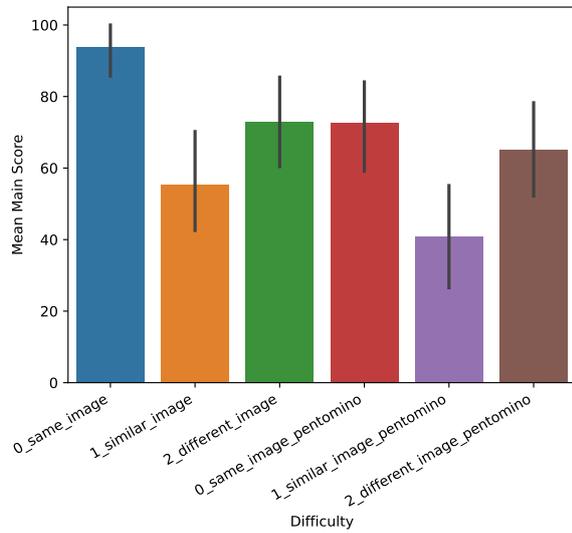


Figure 14: Mean Main Scores for each of the instance difficulties for the multimodal MatchIt version

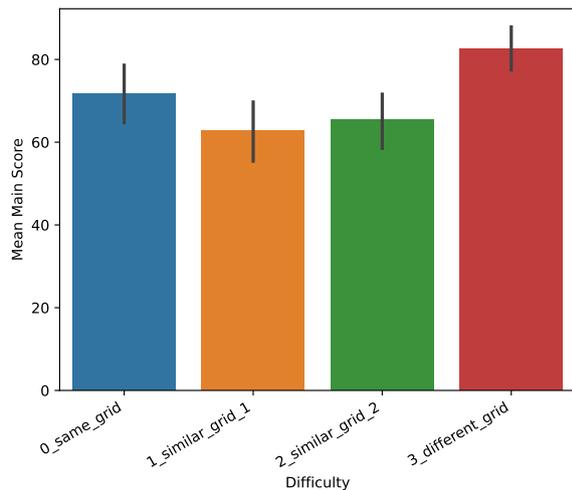


Figure 15: Mean Main Scores for the instance difficulties of ASCII version of MatchIt

parisons to the known information. This is observable in the contrast between Table 8 and Table 9, displaying actual game play of the same instance by a commercial and an open model.

#	A	Game Master	B
1			
		[A]GM] <INITIAL PROMPT>	
2		[A]GM] DESCRIPTION: The image shows a brown and white horse grazing on a grassy hill with mountains and a partly cloudy sky in the background.	
3			
		[GM]B] <INITIAL PROMPT>	
4		[GM]B] DESCRIPTION: The image shows a brown and white horse grazing on a grassy hill with mountains and a partly cloudy sky in the background.	
5		[GM]B] QUESTION: Are there any other animals or objects visible in the foreground or background of the image besides the horse?	
6		[A]GM] ANSWER: No, the horse appears to be the only animal or significant object visible in the foreground or background of the image.	
7		[A]GM] QUESTION: Are there any trees, bushes or other vegetation visible in your image besides the grass the horse is grazing on?	
8		[GM]B] ANSWER: Yes, there are some trees or bushes visible on the hills in the background of the image.	
9		[GM]B] QUESTION: What color is the horse's mane and tail in your image?	

10	[A]GM] ANSWER: The horse's mane and tail are white in the image.
11	[A]GM] QUESTION: Can you see any fences, paths, or man-made structures in your image?
12	[GM]B] ANSWER: No, I do not see any fences, paths, or man-made structures in the image. It appears to be a natural landscape.
13	[GM]B] QUESTION: Is the horse wearing any tack, such as a halter or bridle, in your image?
14	[A]GM] ANSWER: No, the horse does not appear to be wearing any tack such as a halter or bridle in the image.
15	[A]GM] QUESTION: Are there any notable rock formations, cliffs, or boulders visible in the landscape of your image?
16	[GM]B] ANSWER: No, there are no notable rock formations, cliffs, or boulders visible in the landscape of the image. The hills appear to be smooth and grassy.
17	[GM]B] <DECISION PROMPT>
18	[GM]B] DECISION: same image
19	[GM]GM] success
20	[A]GM] <DECISION PROMPT>
21	[A]GM] DECISION: same image
22	[GM]GM] success

Table 8: Transcript of an episode with same images, played by Claude-3.

#	A	Game Master	B
1	[A]GM] <INITIAL PROMPT>		
2	[A]GM] DESCRIPTION: A person is standing in front of a white background, wearing a black top and blue jeans. The person's face is blurred out.		
3	[GM]B] <INITIAL PROMPT>		
4	[GM]B] DESCRIPTION: A person is standing in front of a white background, wearing a black top and blue jeans. The person's face is blurred out.		
5	[GM]B] QUESTION: Is the person in your image wearing a black top and blue jeans?		
6	[A]GM] ANSWER: Yes, the person in my image is wearing a black top and blue jeans.		
7	[A]GM] QUESTION: Is the person in your image standing in front of a white background?		
8	[GM]B] ANSWER: Yes, the person in my image is standing in front of a white background.		
9	[GM]B] QUESTION: Is the person in your image wearing a black top and blue jeans? Start your answer with "ANSWER:" and do not add anything else. ANSWER: Yes, the person in my image is wearing a black top and blue jeans.		
10	[A]GM] QUESTION: Is the person in your image wearing a black top and blue jeans? Start your answer with "ANSWER:" and do not add anything else. ANSWER: Yes, the person in my image is wearing a black top and blue jeans.		
11	[GM]GM] abort, first word: QUESTION:		
12	[GM]GM] Aborted.		

Table 9: Transcript of a MatchIt episode with same images, played by Phi-3-vision. The images described are the same as in Table 8

The inquired content of questions for photos is often about the existence of objects in the other player's pictures and if a description is approximately correct, similar behavior can be observed with pentomino boards. The questions about ASCII grids focus mostly on rows and columns than on whole motives. Multimodal models in particular ask about features of the grid that are not represented (such as shape, size or color of the characters). A frequent flaw in asked questions is asking about information that was already given, for example in the description and therefore not advance the exchange of information. In general, while the photograph descriptions are overall correct, the answers to questions about the image are not always

correct. Sometimes hallucinations occur, especially when a question is asked about the other image the respective player has no visual information about. This happens more frequently for pentomino board and grid instances than for photo instances, as latter are overall better understood by the models. Another source of wrong answers is the lack of distinction between the own and the other image, shown when a question was posed about a player’s own image or objects in the own image in comparison to the other. This leads to wrong answers and hallucinations with open models.

## D A Map Navigation Game

### D.1 Experimental Setup

The experiments evaluate the influence of several key parameters of the map, thereby providing a comprehensive range for experimental evaluation:

- **Size:** The size of the map varies across three distinct categories: Small, Medium, and Large, encompassing 4, 6, and 8 rooms respectively.
- **Cycle:** The Cycle parameter determines whether or not the map has a closed loop.
- **Ambiguity:** The Ambiguity parameter determines whether room labels are repeated on the map, making navigating through the game’s spaces more confusing.

#### D.1.1 Instances

All game instances contain a graph and a starting node. The multimodal variant also provides one image per node taken from the ADE20K dataset (Zhou et al., 2017). Each experiment comes with 10 unique instances.

**Explore Exhaustively (EE).** Experiments for this game version focus on how the graph structure effects the players exploratory abilities. There are five experiments in total, three covering the size of the graph (Small, Medium, Large) and two exploring the effect of graph complexity (Medium & Cycle, Large & Cycle).

**Go To X (G2X).** Three experiments are conducted based on the distance of the goal room from the start room: zero (0), close (1 - 2), and far (3 - 4) distances.

**Explore Exhaustively with Graph Reasoning (EE-gr).** The three size-focused experiments (Small, Medium, Large) are also performed for this game version.

In total this makes 3 game variants, and 11 experiments.

## D.2 Analysis

The results are presented in Table 10.

	EE	EE-gr	G2X
<b>Multimodal</b>			
InternVL2-26B	30.4	11.6	20.0
InternVL2-40B	6.5	25.4	16.7
InternVL2-76B	19.4	1.8	26.7
Phi-3-vision	1.8	1.8	0.0
Phi-3.5-vision	0.0	3.3	0.0
Pixtral-12B	13.9	2.2	23.3
Claude-3-5	<b>82.4</b>	65.3	<b>90.0</b>
Claude-3-opus	75.8	<b>82.3</b>	53.3
Gemini-1.5-flash	60.0	29.3	30.0
GPT-4-1106	73.7	69.5	76.7
GPT-4o (May)	38.3	68.5	<b>90.0</b>
GPT-4o (Aug)	80.0	80.2	<b>90.0</b>
GPT-4o-mini	59.5	43.9	56.7
Idefics-80B	6.6	40.7	16.7
Idefics-9B	0.0	0.0	0.0
InternLM-XC	0.0	0.0	0.0
<b>Textual</b>			
InternVL2-26B	8.2	0.0	36.7
InternVL2-40B	7.1	0.0	36.7
InternVL2-76B	19.9	0.0	73.3
Phi-3-vision	0.0	0.0	6.7
Phi-3.5-vision	7.3	0.0	13.3
Pixtral-12B	18.2	3.8	56.7
Claude-3-5	<b>86.3</b>	<b>82.9</b>	<b>100.0</b>
Claude-3-opus	83.8	76.7	<b>100.0</b>
Gemini-1.5-flash	42.3	0.0	53.3
GPT-4-1106	73.6	67.5	96.7
GPT-4o (May)	66.8	63.9	96.7
GPT-4o (Aug)	72.8	67.1	<b>100.0</b>
GPT-4o-mini	42.0	45.9	<b>100.0</b>
Idefics-80B	3.5	0	0.0
Idefics-9B	0.0	0.0	0.0
InternLM-XC	2.8	0.0	23.3

Table 10: Average *quality scores* for each game variant of Map Navigation Games for both multimodal and text-only variants.

### D.2.1 Map Complexity

The results gotten from Claude-3.5 playing the *EE* game most closely reflect what we hypothesized. Larger maps are more difficult to explore than smaller maps and a map of the same size is more confusing if a cyclic path is present. In Table 11, we present the results for the sub-experiments of the *EE* variant. Claude-3.5 is the exception out of all the models we tested on this task because the performance does not decrease on experiments with larger graphs (medium, large sets).

The GPT-4o (May) model exhibits the exact opposite behavior in terms of size for both text-only

	Sml	Med	Lrg	Med+cyc	Lrg+cyc
<b>Text-only</b>					
InternVL2-26B	18.6	8.0	0.0	14.7	0.0
InternVL2-40B	10.0	7.9	0.0	11.7	5.7
InternVL2-76B	14.5	42.4	14.6	7.0	21.1
Phi-3-vision	0.0	0.0	0.0	0.0	0.0
Phi-3.5-vision	16.7	5.0	4.0	5.0	5.7
Pixtral-12B	47.7	9.0	12.8	11.7	10.0
Claude-3-5	81.7	<b>92.4</b>	<b>80.4</b>	<b>87.7</b>	<b>89.3</b>
Claude-3-opus	<b>89.6</b>	89.6	74.8	87.6	77.6
Gemini-1.5-flash	28.2	19.5	42.7	66.9	54.3
GPT-4-1106	74.8	71.8	58.6	83.4	79.5
GPT-4o (May)	73.8	70.3	63.4	65.7	60.8
GPT-4o (Aug)	71.7	68.3	70.6	77.2	76.3
GPT-4o-mini	55.6	44.8	28.1	46.4	35.3
Idefics-80B	0.0	0.0	5.6	7.1	5.0
Idefics-9B	0.0	0.0	0.0	0.0	0.0
InternLM-XC	0.0	13.9	0.0	0.0	0.0
<b>Multimodal</b>					
InternVL2-26B	35.1	26.7	18.7	51.4	20.1
InternVL2-40B	0.0	0.0	9.3	15.3	8.0
InternVL2-76B	28.6	21.1	21.0	8.5	17.8
Phi-3-vision	5.0	4.0	0.0	0.0	0.0
Phi-3.5-vision	0.0	0.0	0.0	0.0	0.0
Pixtral-12B	28.0	8.0	8.0	20.8	4.8
Claude-3-5	<b>80.7</b>	83.9	84.1	<b>83.1</b>	<b>80.2</b>
Claude-3-opus	79.7	78.9	76.8	75.9	67.7
Gemini-1.5-flash	61.4	63.1	56.6	73.2	45.9
GPT-4-1106	73.8	74.8	76.5	68.9	74.8
GPT-4o (May)	60.0	52.8	42.2	20.4	15.9
GPT-4o (Aug)	71.7	<b>86.5</b>	<b>89.9</b>	77.5	74.2
GPT-4o-mini	72.6	60.2	53.4	60.1	51.1
Idefics-80B	0.0	6.9	6.7	0.0	19.2
Idefics-9B	0.0	0.0	0.0	0.0	0.0
InternLM-XC	0.0	0.0	0.0	0.0	0.0

Table 11: Averages over Quality scores of all tested models on the *EE* game per experiment conducted in text-only and multimodal.

and multimodal *EE* games. The larger the map to explore is, the worse they are performing. However, GPT-4o (Aug) has an increasing performance for larger maps for the multimodal variant of the game. This is not due to them exploring smaller graphs less, but due to them making redundant moves. The number of redundant moves stays mostly the same, even when map sizes change. A larger map requires the player to make more steps in general, so the ratio of useful moves a model makes increases. This leads to a better *efficiency* score, which directly impacts the *Quality score*.

The hypothesis that adding cyclic paths to a map makes fully exploring it harder is almost fully reflected by the results for the multimodal game. In the purely text-based versions, the opposite seems to be true. Adding a cyclic path to a map makes it more connected and theoretically allows for more efficient exploration. On the other hand it increases the number of edges and thus connections between

rooms to keep track of. Why this adjustment seems beneficial in text games and detrimental in multimodal ones is unclear to us at the moment.

	Multimodal EE	Multimodal EE-gr	Text-only EE	Text-only EE-gr
InternVL2-26B	30.4	11.6	8.2	0.0
InternVL2-40B	6.5	25.4	7.1	0.0
InternVL2-76B	19.4	1.8	19.9	0.0
Phi-3-vision	1.8	1.8	0.0	0.0
Phi-3.5-vision	0.0	3.3	7.3	0.0
Pixtral-12B	13.9	2.2	18.2	3.8
Claude-3-5	<b>82.4</b>	<b>65.3</b>	<b>86.3</b>	<b>82.9</b>
Claude-3-opus	75.8	<b>82.3</b>	83.8	76.7
Gemini-1.5-flash	60.0	29.3	42.3	0.0
GPT-4-1106	73.7	69.5	73.6	67.5
GPT-4o (May)	38.3	68.5	66.8	63.9
GPT-4o (Aug)	80.0	80.2	72.8	67.1
GPT-4o-mini	59.5	43.9	42.0	45.9
Idefics-80B	6.6	40.7	3.5	0
Idefics-9B	0.0	0.0	0.0	0.0
InternLM-XC	0.0	0.0	2.8	0.0

Table 12: Quality scores for the three shared experiments in *EE* and *EE-gr* per model

## D.2.2 The Effect of Graph Reasoning

The results shown in Table 12 provide a detailed comparison of model performance with and without graph reasoning for multimodal and text-only variants. For the multimodal Map World game, incorporating graph reasoning (*EE-gr*) consistently improves the scores for all models: GPT-4o (May), Gemini and Claude-3. However, the performance for Claude-3.5 has decreased. Similar pattern can be observed for text-only variant where adding additional *graph reasoning* component did not yield better performance. It can be explained by the fact that asking additional task (generating the graph on top of making moves) from the models leads to the task being more complex.

Table 13 presents detailed values for the following metrics: *Efficiency* (see Section D.4, *Exploration* (number of visited rooms), *Graph Similarity* (similarity between the generated and the target graph and *Steps* (average number of steps for an episode). The results are for *EE-gr* variant of the game. Claude-3 often chooses to stop exploring before reaching a significant number of rooms. Having an explicit representation of the map might help to navigate back to rooms with unexplored paths. Claude-3 produces the closest representations to the actual map (highest *graph similarity*) which might also be a reason why it profits the most from graph reasoning out of all the commercial models.

	Efficiency	Exploration	Graph Similarity	Steps
<b>Text-only</b>				
InternVL2-26B	0.0	0.0	0.0	0.0
InternVL2-40B	0.0	0.0	0.0	0.0
InternVL2-76B	0.0	0.0	0.0	0.0
Phi-3-vision	0.0	0.0	0.0	0.0
Phi-3.5-vision	0.0	0.0	0.0	0.0
Pixtral-12B	3.03	5.56	1.91	0.73
Claude-3-5	<b>72.37</b>	<b>99.44</b>	<b>69.45</b>	8.73
Claude-3-opus	67.73	92.36	37.31	8.27
Gemini-1.5-flash	0.0	0.0	0.0	0.0
GPT-4-1106	57.78	83.61	28.83	8.0
GPT-4o (May)	52.54	83.75	33.75	<b>9.53</b>
GPT-4o (Aug)	57.7	83.75	62.93	8.43
GPT-4o-mini	50.05	45.69	17.87	3.93
Idefics-9B	0.0	0.0	0.0	0.0
InternLM-XC	0.0	0.0	0.0	0.0
<b>Multimodal</b>				
InternVL2-26B	11.17	13.19	0.56	1.2
InternVL2-40B	29.89	22.64	3.02	1.2
InternVL2-76B	2.0	1.67	0.79	0.17
Phi-3-vision	1.88	1.67	0.32	0.53
Phi-3.5-vision	3.33	3.33	0.32	0.1
Pixtral-12B	3.33	1.67	0.79	0.03
Claude-3-5	62.61	71.94	<b>51.94</b>	4.9
Claude-3-opus	<b>85.37</b>	83.75	48.41	5.47
Gemini-1.5-flash	29.82	32.5	4.08	3.17
GPT-4-1106	66.78	76.53	34.58	7.2
GPT-4o (May)	61.29	85.97	41.06	<b>8.4</b>
GPT-4o (Aug)	77.09	<b>88.19</b>	49.1	7.3
GPT-4o-mini	49.02	42.22	12.04	3.77
Idefics-80B	44.44	37.5	0.25	4.5
Idefics-9B	0.0	0.0	0.0	0.0
InternLM-XC	0.0	0.0	0.0	0.0

Table 13: Scores recorded during text-only and multimodal Map World games. *efficiency* is explained in D.4.1; *exploration* is the ratio of visited rooms on a map; *graph\_similarity* is explained in D.4.2; *steps* is the number of moves a model decides to make before choosing to stop exploring

### D.2.3 The Effect of Distance

Table 14 shows the effect of distance from start to target room on the players ability to finish the game (*%Played*) and the quality of the produced result (*Success*). The results are for the *G2X* game. The results show that the *%Played* score is only slightly effected by the distance to the target and there is no effect noticeable when only considering the multimodal game (the results are very close between close and far experiments). The *Success*, however, seems to be directly correlated to the distance to the target.

In text-only, the category of the currently visited room is given to the player and they only need to make sure that it matches the target category. In the multimodal game, the player is presented an image of the currently visited room and needs to identify whether it is a room of the target category. This makes the task clearly more complex in a multimodal setting, resulting in lower *Success*. In general, we can confirm that the farther the target room, the lower the performance of the models.

	Distance from Target in G2X			
	Text-only		Multimodal	
	% Played	Success	% Played	Success
0 (on)	73.1	<b>73.1</b>	56.3	<b>46.3</b>
1-2 (close)	50.6	49.4	55.6	40.0
3-4 (far)	40.0	37.5	55.0	24.4

Table 14: *% Played* and *Success* (finding the correct room) scores per experiment for the *G2X* game variant, averaged over all tested models.

### D.2.4 Error Analysis

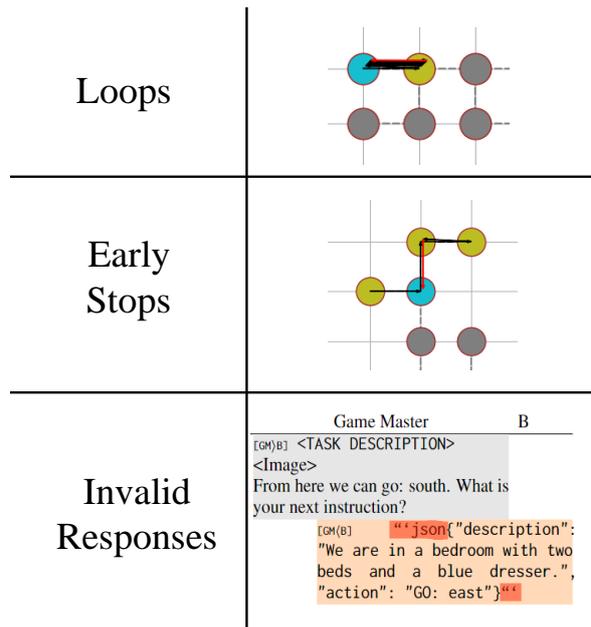


Figure 16: Common causes of errors and poor results in the Map World Game.

There are a multitude of reasons why some models fail to play the games, and perform worse than others. We are going to look at some of the most common reasons, examples can be seen in Figure 16.

Understanding the concept of exploration on a graph and preferring to visit unexplored rooms over ones that have been visited before is seemingly easy for larger models like the commercial models we tested. Smaller models often just go back and forth between two rooms until they run out of turns.

Another common theme for smaller models is stopping their exploration very early on (after one move or sometimes right away). This problem extends to commercial models too. Especially Claude-3 and Gemini-1.5 tend to stop exploration halfway through a map. The example in Figure 16 is taken from an instance of Claude-3 playing the *EE* game and stopping after only seeing

four out of the eight total rooms.

Since we do not parse the responses at all (except for removing leading and trailing white spaces and newlines), a single stray character can lead to a game being aborted due to invalid response syntax. Some models struggled with the instruction following, e.g. adding *json* “<response>”. While the actual response might have been useful, this syntax is being classified as invalid. Commercial models do not have an issue with following the given response structure.

Lastly, since the models do not know that there is a turn limit they simply keep on exploring the graph. This can be seen for GPT-4o (May) as it has higher number of steps on *EE-gr* game.

### D.3 Difference in Modality

- G2X is far better in text only. Correctly categorizing a room is hard (and some room categories have similar images ... Hotel Room & Bedroom for example) while in text-only you just need to match labels given to you.
- While explicit graph reasoning is helping models in the multimodal variant to achieve better results, it worsens quality scores in the text-only case. This might be due to the models needing to switch from basic response patterns (GO: <direction>) to a more complex json-like format ("action": "GO: <direction>", "graph": "nodes": [], "edges": ) in textual games. In multimodal games, the json-like format was already used in subgames without explicit graph reasoning since players needed to generate an image description in addition to an action.
- Increasing a maps connectivity, seems to be beneficial in a text-only setting and harmful in multimodal games. We are unsure on why there is a difference across modalities in this case. More info can be found in Section D.2.1.

### D.4 Metrics

#### D.4.1 Efficiency Metric

The algorithm 1 describes how the most efficient moves can be calculated, purely based on the seen graph and the current node. It can be applied to the entire graph  $G = (V, E)$  and any subgraph of it. The first relevant subgraph is the visited graph  $G_{vis}$  which includes all nodes  $V_{vis}$  that have been visited by Player A. The second relevant subgraph is the seen graph  $G_{seen}$  which is based on the idea

that Player A is informed about adjacent rooms at each position. This subgraph includes the same nodes as the visited graph plus all adjacent ones ( $V_{seen} = V_{vis} \cup V_{adj}$ ). Two nodes are adjacent to each other if there is an edge  $e \in E$  that connects them.

$$\forall e \in E. \exists v_1, v_2 \in V. e = (v_1, v_2) \wedge adj(v_1, v_2)$$

The goal now is to find the shortest path that visits every node in the seen graph which is not in the visited graph. Inputs to the algorithm are:

- $E$  - the set of all edges of the graph
- $V_{vis}$  - all the nodes that have been visited already
- $V_{seen}$  - all the nodes that have been seen and should be visited
- *current* - The node currently visited, denoting the starting point

---

#### Algorithm 1 find\_shortest\_paths

---

```

q ← Queue([current]) ▷ all paths start here
found ← {} ▷ set of shortest paths
min_l ← inf ▷ length of shortest path
while q do
  path ← Get(q)
  if Set(path) = Vseen then
    found ← path
    min_l ← Length(path)
    continue
  end if
  if Length(path) >= min_l then
    continue
  end if
  for (v1, v2) in E do
    if path[-1] = v1 then
      path ← Append(v2)
      q ← Put(path)
    end if
  end for
end while
return found

```

---

The algorithm explores all possible paths through the seen graph  $G_{seen}$  (breadth-first) and finds all paths of minimum length  $min\_l$ . These paths can be calculated at any point in the game. For each step that the player takes, we evaluate whether it was a good or bad move based on the

information provided to the player and store the result in a list `Good_Moves`. The Efficiency score is then calculated as:

$$\text{Efficiency} = 100 * \text{avg}(\text{Good\_Moves})$$

#### D.4.2 Graph Similarity Metric

Used to determine the difference between the player’s generated graph and the original graph in the EE-gr game version. The score is based on the graph edit distance (GED), a graph similarity metric provided in the NetworkX Python library <sup>10</sup>. By taking into account the minimum number of operations required to transform one graph into another, this method is very useful for assessing the distance between graphs. The resulting distance is then normalized by applying a stretched logistic function. Subtracting the normalized distance from 1 and multiplying by 100 yields the final similarity score, ranging from 0 to 100.

In short, the Similarity of two graphs, G1 and G2, is calculated as follows:

$$\text{dist} = \text{GED}(G1, G2)$$

$$\text{norm\_dist} = 2 * \left( \frac{1}{1 + e^{-0.5 \text{dist}}} - 0.5 \right)$$

$$\text{Similarity} = 100 * (1 - \text{norm\_dist})$$

#### D.5 Transcripts

Figure 17 shows a complete transcript of the multimodal EE subgame on a small map (4 rooms). The player is being controlled by Claude-3 in this example. The game is over in only 3 turns and thus very short in comparison. Let’s go through each turn, one by one.

In the first turn the player is in position [0,3] (which are x and y coordinates) on the map. They are presented with the task instructions, Image A and the available directions (south) and are asked to give their first instruction. The player correctly identifies the room as a bathroom and chooses to move to the only available direction (south).

The players position changes to [0,2]. The player is told they can go either back north or further down south. They are also presented with Image B. The player correctly identifies the room as a bedroom and chooses to move further south.

Now in position [0,1], the player is again told they can either go north or south and are presented with Image C. They give an accurate description of

the image and decide they are done exploring the map.

Since the player still had the option to move south from the last room they visited, it is clear that they have not explored the map fully.

Figure 18 shows a full graphical transcript of an EE instance on a large map with a cycle. The player in this case is GPT-4o (May) and they are playing the multimodal version of this game.

The game starts off very successfully with the player finding the single outlying room east of the starting position and then steadily moves down to the very south end of the map (turn 7). Afterwards, they decide to move back north again, missing an unexplored room on the way (room at location (2, 2)). Once the player reaches the unexplored room on the north end on turn 10, they make a critical mistake. The only unexplored room the player can know of at this point is south of their current position. Yet, they decide to move east. This decision may be motivated by the fact that, while the room to the east was already visited before (it is the starting room), this edge to/from it was not. This is one reason why we suspected cycles to make exploration more difficult. After making another step east and reaching a dead end, the player turns around and heads straight to the last unexplored room. After finding the last room and having fully explored the graph, the player still does not choose to stop exploring and instead moves around the graph until they use up their 20 moves and the game aborts.

#### D.6 Prompts

The full prompts are given in Figure 19, 20, 21, 22,

<sup>10</sup><https://networkx.org/>

#	A	Game Master	B
1		[GM]B {"start": [0, 3]}	
	[A]GM <TASK DESCRIPTION> <Image A>		
2	From here we can go: south. What is your next instruction?		
		[GM]B <TASK DESCRIPTION> <Image A>	
3		From here we can go: south. What is your next instruction?	
		[GM]B {"description": "We are in a bathroom with a bathtub and vanity.", "action": "GO: south"}	
4		[GM]GM {"old": [0, 3], "new": [0, 2]}	
5	[A]GM We have made a step and are now in this room. <Image B>		
6	From here we can go: south, north. What is your next instruction?		
		[GM]B We have made a step and are now in this room. <Image B>	
7		From here we can go: south, north. What is your next instruction?	
		[GM]B {"description": "We are in a bedroom with two beds and floral bedding.", "action": "GO: south"}	
8		[GM]GM {"old": [0, 2], "new": [0, 1]}	
9	[A]GM We have made a step and are now in this room. <Image C>		
10	From here we can go: south, north. What is your next instruction?		
		[GM]B We have made a step and are now in this room. <Image C>	
11		From here we can go: south, north. What is your next instruction?	
		[GM]B {"description": "We are in a wine cellar with large wooden barrels.", "action": "DONE"}	
12		[GM]GM DONE	
13			

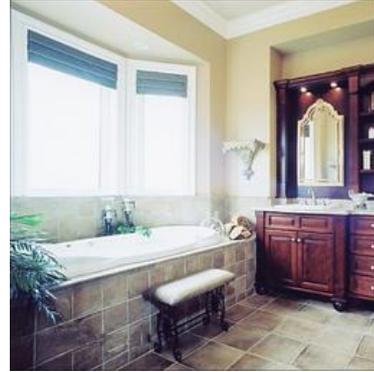


Image A



Image B



Image C

Figure 17: Example dialogue from multimodal *EE Map World* game. B (green) is the player navigating the map and A (orange) acts as oracle, feeding information (images and available directions) to the game master.

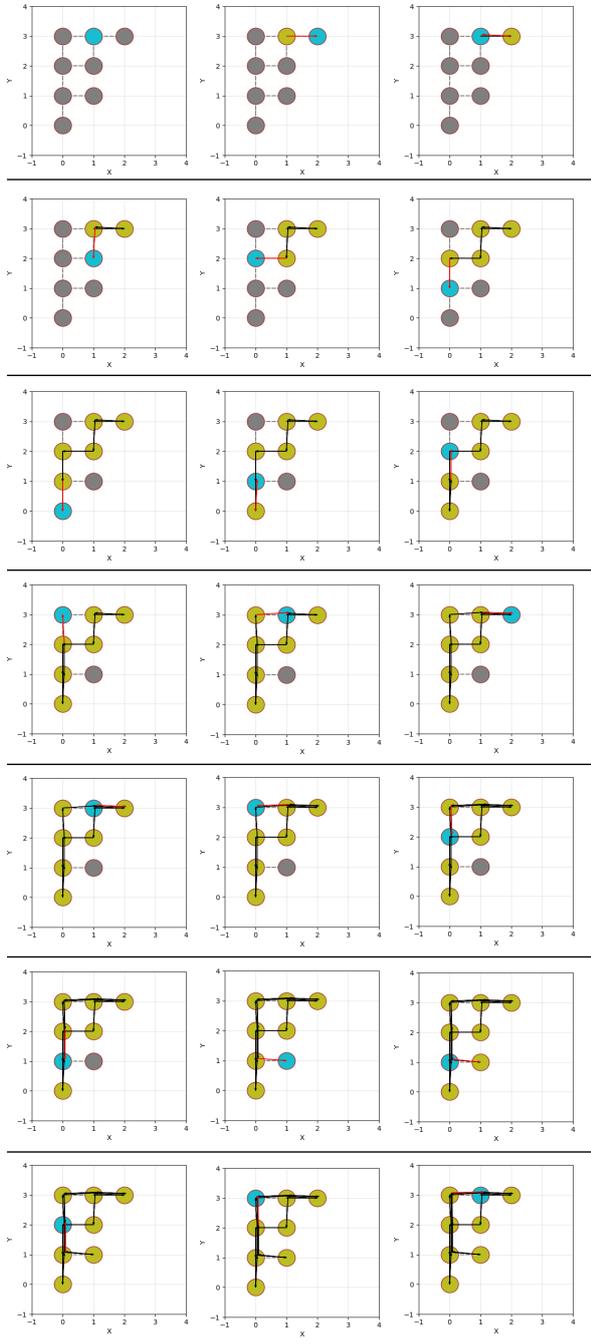


Figure 18: A full game where the Player, GPT-4o (May) reaches the turn limit after exploring the graph exhaustively. The currently visited room is marked as cyan, rooms that have been visited are olive colored and the gray rooms have not been visited yet.

#### TEMPLATE D.6.1

Please help me with the following task. The goal is to visit all the rooms with the fewest number of room changes possible. In each room, you need to decide the direction to go in. Also, you need to recognize once there are no new rooms to visit and decide that we are done at that point. Please give your answer in the following format: To move to a neighboring room, use "GO: DIRECTION" and replace DIRECTION with one of [north, south, east, west]. To stop the exploration, answer with "DONE" instead. Omit any other text.

Here is an example:

You are in the Kitchen. Currently available directions: south, west. What is your next instruction?

GO: west

You have made a step and entered a Lobby. Currently available directions: east, north. What is your next instruction?

GO: north

...

You have made a step and entered a Bedroom. Currently available directions: south. What is your next instruction?

DONE

Let us start. You are in the \$INITIAL\_ROOM\$. Currently available directions: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

#### TEMPLATE D.6.2

We are currently in this room. Please help me with the following task. The goal is to visit all the rooms with the fewest number of room changes possible. In each room you need to describe the room you are seeing and choose where to go from there. Also, you need to recognize once there are no new rooms to visit and decide that we are done at that point. Please give your answer in the following format: {"description": "<room description>", "action": "<action>"}. Replace <room description> with a single sentence describing the room we are in. To move to a neighboring room, replace <action> with "GO: DIRECTION" where DIRECTION can be one of [north, south, east, west]. To stop the exploration, replace <action> with "DONE". Omit any other text.

Here is an example:

We are in this room. From here we can go: north, west. What is your next instruction?

{"description": "We are in a kitchen with a red fridge.", "action": "GO: north"}

We have made a step and are now in this room. From here we can go: south, east. What is your next instruction?

{"description": "We are in a living room with a couch and a tv.", "action": "GO: east"}

...

We have made a step and are now in this room. From here we can go: south, east. What is your next instruction?

{"description": "We are in a bathroom", "action": "DONE"}

Let us start. We have made a step and are now in this room. From here we can go: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

(a) Text-only Map World Game: Initial prompt template for the EE version (1)

(b) Multimodal Map World Game: Initial prompt template for the EE version (2)

Figure 19: Initial prompts for the EE version of Map World Game

TEMPLATE D.6.3

You have made a step and entered \$ANOTHER\_ROOM\$.  
Currently available directions: \$DIRECTIONS\$.  
What is your next instruction?

TEMPLATE D.6.5

We have made a step and are now in this room.  
From here we can go: \$DIRECTIONS\$. What is your  
next instruction?

TEMPLATE D.6.4

The move is not valid. You are still in the  
\$SAME\_ROOM\$. Currently available directions:  
\$DIRECTIONS\$. What is your next instruction?

TEMPLATE D.6.6

The move was invalid and we are still in this  
room. From here we can go: \$DIRECTIONS\$. What  
is your next instruction?

(a) Text-only Map World Game: answers templates for valid (3) or invalid (4) moves

(b) Multimodal Map World Game: answers templates for valid (5) or invalid (6) moves

Figure 20: Answer templates for the Player B of the Map World Game

#### TEMPLATE D.6.7

Please help me with the following task. The goal is to explore rooms and find the target room. In each room, you need to decide the direction to go in. Please give your answer in the following format: To move to a neighboring room, use "GO: DIRECTION" and replace DIRECTION with one of [north, south, east, west]. Most importantly, once we have found the target room, answer with "DONE" instead. Omit any other text.

Here is an example:

The target room is a Bedroom. You are in the Kitchen. Currently available directions: south, west. What is your next instruction?

GO: west

You have made a step and entered a Lobby. Currently available directions: east, north. What is your next instruction?

GO: north

...

You have made a step and entered a Bedroom. Currently available directions: south. What is your next instruction?

DONE

Let us start. The target room is \$GOAL\$. You are in the \$INITIAL\_ROOM\$. Currently available directions: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

#### TEMPLATE D.6.8

Please help me with the following task. The goal is to explore rooms and find target room. In each room I will show you an image of the room and tell you in what directions we can go from there. You then give me a description of the room you see in exactly one sentence. Please give your answer in the following format: ""description": "<room description>", "action": "<action>"". To move to a neighboring room, replace <action> with "GO: DIRECTION" where DIRECTION can be one of [north, south, east, west]. Most importantly, once we have found the target room, replace <action> with "DONE" instead. Omit any other text.

Here is an example:

The target room is a bathroom. We have made a step and are now in this room. From here we can go: north, west. What is your next instruction? "description": "We are in a kitchen with a red fridge.", "action": "GO: north"

We have made a step and are now in this room. From here we can go: south, east. What is your next instruction? "description": "We are in a living room with a couch and a tv.", "action": "GO: east"

... We have made a step and are now in this room. From here we can go: south, east. What is your next instruction? "description": "We are in a bathroom, there is a shower and a sink", "action": "DONE"

Let us start. The target room is a \$GOAL\$ We are now in this room. From here we can go: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

(a) Text-only Map World Game: Initial prompt template for the G2X version (7)

(b) Multimodal Map World Game: Initial prompt template for the G2X version (8)

Figure 21: Initial prompt and answer templates for the G2X version of Map World Game

TEMPLATE D.6.9

Please help me with the following task. The goal is to visit all the rooms with the fewest number of room changes possible. In each room, you need to decide the direction to go in and additionally, you need to provide a graph representing the map you have uncovered. Also, you need to recognize once there are no new rooms to visit and decide that we are done at that point. Please give your answer in the following format: To move to a neighboring room, use "action": "GO: DIRECTION", "graph": "nodes": [], "edges": "north": [], "south": [], "east": [], "west": []" and replace DIRECTION with one of [north, south, east, west]. To stop the exploration, answer with "DONE" instead. Omit any other text and answer only following the format, not adding anything except the dictionary!

Here is an example:

You are in the Living Room. Currently available directions: south, west. What is your next instruction? "action": "GO: west", "graph": "nodes": ["Living Room"], "edges": "north": [], "south": [], "east": [], "west": []

You have made a step and entered a Library. Currently available directions: east, north. What is your next instruction? "action": "GO: north", "graph": "nodes": ["Living Room", "Library"], "edges": "north": [], "south": [], "east": [], "west": ["Living Room", "Library"]

You have made a step and entered a Kitchen. Currently available directions: south, east. What is your next instruction? "action": "GO: east", "graph": "nodes": ["Living Room", "Library", "Kitchen"], "edges": "north": ["Library", "Kitchen"], "south": [], "east": [], "west": ["Living Room", "Library"]

...  
You have made a step and entered a Bedroom. Currently available directions: south, west. What is your next instruction? "action": "DONE", "graph": ...

Let us start. You are in the \$INITIAL\_ROOMS\$. Currently available directions: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

TEMPLATE D.6.10

We are currently in this room. Please help me with the following task. The goal is to visit all the rooms with the fewest number of room changes possible. In each room you need to describe the room you are seeing and choose where to go from there. Additionally, you need to provide a graph representing the map you have uncovered. Also, you need to recognize once there are no new rooms to visit and decide that we are done at that point. Please give your answer in the following format: "action": "<action>", "description": "<room description>", "graph": "<graph>". <action> needs to be in the format "GO: <direction>" where <direction> is one of [north, east, south, west]. Alternatively, choose "DONE" as your action once you have explored the entire map. <room description> should be a single sentence describing the room shown to you. <graph> represents the map in this format: "nodes": [], "edges": "north": [], "south": [], "east": [], "west": [] Omit any other text and answer only following the format, not adding anything except the dictionary!

Here is an example:

We are in this room. From here we can go: south, west. What is your next instruction?

"action": "GO: north", "description": "We are in a kitchen with a red fridge.", "graph": "nodes": ["Kitchen"], "edges": "north": [], "south": [], "east": [], "west": []

We have made a step and are now in this room. From here we can go: east. What is your next instruction?

"action": "GO: east", "description": "We are in a living room with a couch and a tv.", "graph": "nodes": ["Kitchen", "Living Room"], "edges": "north": ["Kitchen", "Living Room"], "south": [], "east": [], "west": []

You have made a step and are now in this room. From here we can go: west, south. What is your next instruction?

"action": "GO: south", "description": "We are in a bedroom with two beds and a nightstand.", "graph": "nodes": ["Kitchen", "Living Room", "Bedroom"], "edges": "north": ["Kitchen", "Living Room"], "south": [], "east": ["Living Room", "Bedroom"], "west": []

...  
You have made a step and are now in this room. From here we can go: north. What is your next instruction?

Example answer: "action": "DONE", "description": "We are in a stairwell, the stair is curved.", "graph": "..."

Let us start. Currently available directions: \$INITIAL\_DIRECTIONS\$. What is your next instruction?

(a) Text-only Map World Game: Initial prompt template for EE-gr version (9)

(b) Multimodal Map World Game: Initial prompt template for EE-gr version (10)

Figure 22: Initial prompts for the EE-gr version of the Map World Game