Are Multimodal Large Language Models Pragmatically Competent Listeners in Simple Reference Resolution Tasks?

Simeon Junker Manar Ali Larissa Koch Sina Zarrieß Hendrik Buschmeier
CRC 1646 'Linguistic Creativity in Communication'
Bielefeld University, Bielefeld, Germany

Abstract

We investigate the linguistic abilities of *multimodal* large language models in reference resolution tasks featuring simple yet abstract visual stimuli, such as color patches and color grids. Although the task may not seem challenging for today's language models, being straightforward for human dyads, we consider it to be a highly relevant probe of the pragmatic capabilities of MLLMs. Our results and analyses indeed suggest that basic pragmatic capabilities, such as context-dependent interpretation of color descriptions, still constitute major challenges for state-of-the-art MLLMs.

1 Introduction

The advent of large language models (LLMs) and their expansion in scale, variety, and availability over the past decade has led to considerable interest – both in the research community as well as in the public at large – in understanding their capabilities and limitations. A lot of research focuses on their general abilities (reasoning, math, professional examinations, . . .) in order to answer questions about the level or category of 'intelligence' these models exhibit (e.g., Bubeck et al., 2023) or which practical tasks they might be suitable for. In contrast to this, research in (computational) linguistics is also interested in their basic linguistic abilities (Chang and Bergen, 2024; Millière, forthcoming).

In this paper, we investigate the linguistic abilities of *multimodal* large language models (MLLMs) on the level of language use, i.e., linguistic pragmatics, in the well-known reference resolution paradigm. More specifically, we examine whether off-the-shelf MLLMs (LLaVA-NeXT, Qwen2-VL, and Janus-Pro) are able to resolve references to abstract visual stimuli (color patches and color grids; Monroe et al., 2017; McDowell and Goodman, 2019; see Fig. 1) given in director-matcher-style dyadic reference games (Clark and Wilkes-Gibbs, 1986).

Although the task may not seem challenging for today's language models because it is straightforward and easy for human dyads, we consider it to be a highly relevant probe of the pragmatic capabilities of MLLMs. Being ubiquitous in all of language use, reference (generation and resolution) is highly context-dependent (here visual context). The same can be said about the color references at the center of the task (where color is a main distinguishing attribute between potential referents; Monroe et al., 2017). The complexity of the visual context varies between the two types of stimuli we consider, but can generally be considered simple. Their abstractness, however, poses demands on the basic visual perception capabilities of MLLMs.

Thus, we investigate how well MLLMs perform basic reference resolution tasks requiring contextualized pragmatic reasoning about color and simple spatial arrangements in two abstract visual domains.[1] Our results show that models with sufficient capacity achieve promising results for color patches. However, even the best-performing models struggle with the complex structure of color grids.

2 Background

The semantics and pragmatics of references to color have long played a special role in work on reference games (Pechmann, 1989; Baumgaertner et al., 2012; Koolen et al., 2013; Zarrieß and Schlangen, 2016). The categorization of colors is well-known to be subject to complex interactions between semantic and perceptual information (Mitterer and de Ruiter, 2008). The naming of colors in interactive games is also well-known to be subject to pragmatic reasoning and negotiation between interaction partners (Meo et al., 2014; McMahan and Stone, 2015; Monroe et al., 2017). Finally, color references have been

[1] Code and data of the study are available at https://doi.org/10.5281/zenodo.15553655 as well as https://github.com/clause-bielefeld/mllm-listeners.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24101–24109

director: *yellow green. the less green one*
model responses: 4× left, 4× **middle**, 0× right

(a) Color patch example (Monroe et al., 2017).

director: *first square is brown*
model responses: 5× left, 3× **middle**, 0× right

(b) Color grid example (McDowell and Goodman, 2019).

Figure 1: Example color patches and color grids stimuli for director-matcher-style dyadic reference games with the human director's description and reference resolution responses of eight different MLLMs. In both examples the target referent is the object in the middle and was correctly identified by the human matcher.

studied from the perspective of figurative language and creativity (Kawakami et al., 2016).

Color perception and understanding tasks have also been used in recent work on probing language models. Loyola et al. (2023) explored the alignment between the perceptual color spaces of humans and text-based LLMs. Their findings indicate moderate alignment for basic color terms, which decreased as color descriptions become complex and subjective. Similarly, Abdou et al. (2021) found that alignment improves with model size. Jones et al. (2024) examined the sensitivity of MLLMs to sensorimotor features by testing their ability to identify images that match textual implied features, showing that the effect for color emerges only in the largest model. Rahmanzadehgervi et al. (2024)'s study on vision language models shows that even state-of-the-art models, such as GPT4-o and Gemini-1.5 Pro, perform surprisingly poorly at a range of low-level tasks, which humans are expected to complete with ease, e.g., determining whether two circles overlap.

Together, these studies warrant further investigations and analyses into multimodal language understanding tasks. In this paper, we investigate tasks for analyzing low-level language grounding of color stimuli in combination with pragmatic capabilities, focusing on referring expressions produced by human players of reference games.

3 Experimental Setting

3.1 Models

We investigate reference resolution in the following MLLMs: **LLaVA-NeXT** (Liu et al., 2024b) builds on LLaVA 1.5 (Liu et al., 2024a) and the original LLaVA (Liu et al., 2023), with CLIP-ViT-L as the vision encoder. For the LLM backbone, we use Vicuna at the 3b and 7b parameter scales, NousHermes2-Yi at the 34b scale, and Llama 3 at the 72b scale. **Qwen2-VL** (Wang et al., 2024) upgrades the Qwen-VL (Bai et al., 2023) models and uses ViT as a vision encoder and multimodal rotational position encoding (M-RoPE). It comes with Qwen2 as the LLM backbone and is available with 2b, 7b, and 72b parameters. Qwen2-VL-72B has been reported to perform similarly to state-of-the-art models such as GPT-4o (Wang et al., 2024). **Janus-Pro** (Chen et al., 2025) enhances the original Janus (Wu et al., 2024), both of which utilize the structure of decoupling visual encoding (SigLIP for understanding tasks, VQ tokenizer for generation) with an auto-regressive transformer LLM backbone. We use it at 1b and 7b parameter scales. See Table 5 (Appendix) for a more detailed model overview.

3.2 Data

We base our investigation on two human director-matcher-style reference game data sets, in which a director (speaker) describes a target to a matcher (listener) who must identify the described target (see Fig. 1). Each data point – one round of a game – consists of an abstract visual stimulus (e.g., three color patches in random order) and the textual utterances produced by the two participants. In both datasets, the stimuli were created with three degrees of visual complexity ('far', 'split', and 'close'), based on CIELAB color distances.

Color Patches The data set consists of 948 games of 50 rounds each (Monroe et al., 2017). Participants were assigned the role of either 'director' or 'matcher'. In each round of the game, both participants were presented with three color patches (Fig. 1a) shown in a different order. The director knows the 'target' color patch and has to describe it by text input, for the matcher to identify it. The matcher can clarify and has to select the target. With three colors being displayed, the description is thus not only shaped by the target color itself, but also by its context, the two 'distractor' colors from which it must be distinguished.

Model	Size (b)	Quant	Color Patches				Color Grids			
			Total	Far	Split	Close	Total	Far	Split	Close
Janus	1	—	36.1	38.1	35.4	34.8	33.3	33.2	33.4	33.3
	7	—	68.4	83.9	64.4	56.8	39.5	41.1	38.7	38.7
LLaVA	7	—	60.1	71.4	59.1	49.5	38.0	38.6	38.7	36.8
	13	—	59.4	70.0	57.8	50.4	37.7	38.3	37.9	36.8
	34	—	80.3	93.1	77.4	70.2	37.9	38.8	38.2	36.6
	72	8bit	62.3	75.8	59.9	51.2	39.9	42.0	40.2	37.6
Qwen	2	—	61.9	77.3	58.3	50.0	38.4	40.0	38.3	36.8
	7	—	83.0	94.1	81.1	73.8	45.2	47.7	44.4	43.4
	72	awq	**87.5**	**95.1**	**86.9**	**80.3**	**66.5**	**70.2**	**66.0**	**63.2**
human	—	—	90.0	97.0	89.7	83.3	92.7	96.0	92.4	89.8

Table 1: Accuracy scores (%) for all models and datasets. The highest score per column is highlighted in bold.

Color Grids The data set consists of 197 games of 60 rounds each (McDowell and Goodman, 2019). It follows the same procedure, but the stimuli consist of three 3×3 grids of color patches (Fig. 1b).

3.3 Prompting Procedure

For generating multimodal prompts for the MLLMs, we render color patches and grid items in the same order as they were presented to the human matcher in the original data. We concatenate the three color patches or the three color grid items into a single image before feeding this combined image to the model. The verbal part of the prompt contains the full, original dialogue between director and matcher on a visual stimulus, transformed into a script-like format (*speaker:* . . . ¶ *listener:* . . . ¶ . . .). We prompt the model to locate the target referent described by the speaker in the dialogue (see Appendix B), generate the model response using greedy decoding, and extract the location information (i.e., left, middle, or right) from the model response using a regular expression.

4 Results

We evaluate the MLLMs' performance with accuracy, i.e., how often the models identify the visual target correctly. We compare results between (a) visual complexity condition (far/close/split, see Section 3.2) (b) humans and models; and (c) different models and model sizes (see Section 3.1). Table 1 shows the main results for color patches and color grids. In addition, we test our models on restricted subsets of the color patch and grid datasets, where we included only dialogues with a total length of ≤ 5 words, out of which $\geq 1/3$ have to be basic color terms (see Appendix E). We applied this restriction to reduce the linguistic complexity and to remove the meta-commentary that often appears in longer exchanges (Monroe et al., 2017).

4.1 Color Patches

Accuracies in the color patch task show a wide range of performance levels, across models and model sizes. The smallest Janus model barely performs above chance level (36 % accuracy), displaying a strong bias to predict *left* positions for targets (see Appendix F). In contrast, the largest Qwen model comes close to human accuracy (cf. Table 1). Qwen 7b also outperforms Janus and LLaVA variants of the same size, and even Qwen 2b achieves competitive results. Janus and Qwen generally improve with larger model sizes. For LLaVA, this is much less consistent, likely because the larger models use different LLM backbones (see Section 3.1, Table 5, Appendix). Here, the 7b, 13b, and 72b variants achieve very similar scores, with the largest variant outperforming the smallest model only by a 2-point increase in accuracy. LLaVA 34b behaves as an outlier and achieves notable improvements over all other LLaVA variants.

Model performance also depends substantially on the degree of visual complexity, with most models achieving the highest scores in the *far* condition, and the lowest accuracies for the more challenging *close* items. While increasing size in the LLaVA models (except 34b) does not incur any improvement on the *close* condition, we find a substantial improvement for Qwen 72b over 7b on the *close* condition.

On the subset of items with limited description complexity, all models show only small improvements (cf. Table 3, Appendix E). Reducing linguistic complexity in the dialogues does not make it substantially easier for the models to resolve the visual targets. This suggests that models face challenges

in linguistically simple expressions consisting of a few basic color terms only. Our inspection of examples (Section 4.3) supports this. Overall, the results show near-human scores from some models, but also demonstrate that even supposedly simple color patches can pose problems for current MLLMs.

4.2 Color Grids

Accuracy scores for color grids (Table 1) are considerably lower than for color patches, with the best-performing model (Qwen 72b) falling short of the total human accuracy by more than 25 points. The lowest performing model (Janus 1b) does not even perform above chance level. We again observe that larger models achieve higher accuracies, but only Qwen 72b achieves a somewhat satisfactory accuracy of 66%, surpassing Qwen 7b by more than 20 points. This indicates that all models did not learn certain aspects of multimodal language grounding required in this task, failing to reason about the more challenging abstract grid shapes.

Overall, differences between the visual complexity conditions are less pronounced than for color patches, supporting the impression that models generally struggle with grounding references in the grid shapes. Notably, reducing linguistic complexity leads to further decreases in accuracies for the *far* and *split* color conditions (Table 3), indicating that this task exhibits pragmatic complexities beyond simple color understanding.

Part of this is the more complex spatial arrangement of colors in the grid, resulting in frequent spatial descriptions in the dialogues. Qualitative inspections of examples led us to suspect that models adopt a simple strategy of spotting spatial keywords and using these in their responses. To test this, we partition the grid data into smaller sets of dialogues containing exactly one of the position labels "left", "middle" or "right", and calculate the frequency in which our models generate the respective label in the responses. The results in Table 4 (Appendix F) show that all of our models are biased towards predicting labels which are also mentioned in the dialogues. This indicates that all tested models struggle with this nested structure of grids and do not differentiate enough between position descriptions within grids or between potential referents.

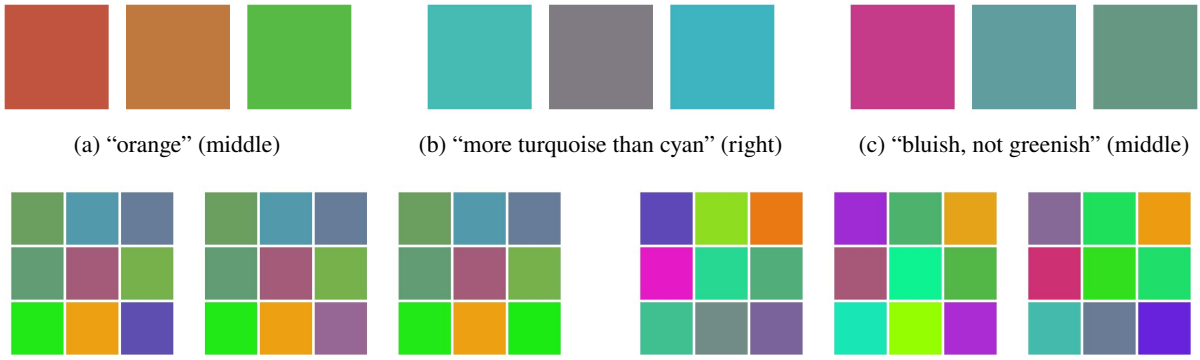
4.3 Examples

A qualitative inspection of reference resolution failures in the color patch data (Examples 1a–1c and Table 2) suggests that models struggle with different semantic and pragmatic phenomena. In Example 1a, the basic color term "orange" also applies to a distractor, which, however, would rather be called "red" in this context. Some models show problems with this ambiguity, including all LLaVA variants. Similarly, in Example 1b, all models struggle with a basic but highly context-dependent comparative color description. Example 1c illustrates a case where negation and graded color terms challenge many of the models. In the color grid data (Examples 1d and 1e, Table 2), these issues combine with challenges in understanding the basic grid layouts. For instance, in Example 1d, models fail to locate the graded description of a shade of *purple* within the respective grid and seem to simply focus on the locative adjective *right*. Models also fail on complex grid examples like Example 1e, where descriptions of multiple rows or cells within a grid need to be composed, for a correct resolution of the reference.

4.4 Discussion

In general, our results do not warrant a conclusive answer to the title question. While the large variants of the Qwen model achieve close to human performance on the simple color patch task and show promising accuracies in resolving references to more complex color grids, the LLaVA and Janus models do not show robust competencies in either setting. The great difference between the large LLaVA and Qwen models is particularly striking in this regard, suggesting that architectural decisions in MLLMs and fine-tuning protocols can have substantial effects on fundamental capabilities in situated language understanding – compared to LLaVA, Qwen uses a different vision encoder and multimodal rotational position encoding, and was subject to multiple pretraining stages (see comparison in Table 5, Appendix). Future work should explore such variations more systematically, potentially with smaller LMs, to gain a deeper understanding of the factors that lead to implicit pragmatic competencies in LMs.

In addition, different reasoning skills can be necessary for correctly resolving references. While descriptions often only apply to single items (e.g., "yellow green" in Figure 1a), they are more ambiguous in other cases and require reasoning about how the director might have referred to alternative targets (e.g., "orange" in Example 1a, see Section 4.3). The distinction between these cases is not always clear, which is why our results do not differentiate



(a) “orange” (middle) (b) “more turquoise than cyan” (right) (c) “bluish, not greenish” (middle)

(d) “the right bottom is a bright purplish blue” (left) (e) “middle tile green (not aqua), lower right purple” (right)

Example 1: Examples for color patches (a–c) and color grids (d, e). The caption of each example shows the description of the target given by the director, and the target (left, middle, right). See Table 2 for human and model responses.

				Janus		LLaVA				Qwen	
Ex.	**Condition**	**Correct**	**Human**	**1b**	**7b**	**7b**	**13b**	**34b**	**72b**	**2b**	**72b**
1a	split	middle	**middle**	left	**middle**	left	left	left	left	**middle**	**middle**
1b	split	right	**right**	left	left	left	left	left	left	left	left
1c	split	middle	**middle**	left	**middle**	right	right	**middle**	right	left	**middle**
1d	split	left	**left**	right	right	right	right	right	right	right	right
1e	close	right	**right**	left	middle	middle	middle	middle	middle	middle	middle

Table 2: Human reference resolution and model responses for the target color patch examples 1a–1c and color grid examples 1d and 1e. Correctly resolved references are highlighted in boldface.

between different forms of contextual or pragmatic reasoning.

Finally, one advantage that the human matchers had in the reference games is that they can collaborate with the director, e.g., by requesting clarification or more information. In contrast to that, the MLLMs in our investigation are merely ‘overhearers’ (Schober and Clark, 1989) and had to rely on the information specified in the prompt. Future work should explore more interactive settings where the agent can negotiate common ground with the director.

5 Conclusion

In this paper, we investigated how well MLLMs perform basic reference resolution tasks that require contextualized pragmatic reasoning about color in two abstract visual domains of different complexity. We found that models with sufficient capacity achieve promising results in the simpler domain of color patches, but that even the best-performing models struggle with the more complex structure of color grids.

Limitations

The study presented in this paper has a number of limitations. First, we have focused on models whose weights are available and can be run locally. The performance of commercial models such as ChatGPT or Gemini may be different. Second, we presented the stimuli as a single image to the MLLMs. Changing the format of the visual input so that each of the three object is provided as an individual image could enable the models to use the full potential of their visual encoders for each stimulus object. This could be particularly beneficial in the color grid domain with more complex objects. Third, here we did not systematically study the dialogue structure of some of the human interactions. It would be very interesting to investigate if there are differences between cases where the required information is contained in a single utterances in contrast to cases where it is built up by interlocutors turn-by-turn. Finally, the order of location labels was fixed (left, middle, right) in our prompts to the models and there is a tendency, at least in some of the models, to prefer the label first mentioned (left). Changing the order in the prompt may mitigate this issue.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Bert Baumgaertner, Raquel Fernández, and Matthew Stone. 2012. Towards a flexible semantics: Colour terms in collaborative reference tasks. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 80–84, Montréal, Canada. Association for Computational Linguistics.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms. Their Universality and Evolution*. University of California Press, Berkeley, CA, USA.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *Preprint*, arXiv:2303.12712.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50:293–350.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Cameron R. Jones, Benjamin Bergen, and Sean Trott. 2024. Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50:1415–1440.

Kazuya Kawakami, Chris Dyer, Bryan Routledge, and Noah A. Smith. 2016. Character sequence models for colorful words. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1954, Austin, Texas. Association for Computational Linguistics.

Paul Kay and Chad K. McDaniel. 1978. The linguistic significance of the meanings of basic color terms. *Language*, 54:610–646.

Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. 2013. The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37:395–411.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Pablo Loyola, Edison Marrese-Taylor, and Andres Hoyos-Idrobo. 2023. Perceptual structure in the absence of grounding: The impact of abstractedness and subjectivity in color language for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1536–1542, Singapore. Association for Computational Linguistics.

Bill McDowell and Noah Goodman. 2019. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy. Association for Computational Linguistics.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, pages 107–115, Edinburgh, UK.

Raphaël Millière. forthcoming. Language models as models of language. In Ryan M. Nefdt, Gabe Dupre, and Kate Stanton, editors, *Oxford Handbook of the Philosophy of Linguistics*. Oxford University Press, Oxford, UK.

Holger Mitterer and Jan Peter de Ruiter. 2008. Recalibrating color categories using world knowledge. *Psychological Science*, 19:629–634.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language

understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 18–34, Hanoi, Vietnam.

Michael F. Schober and Herbert H. Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21:211–232.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

Sina Zarrieß and David Schlangen. 2016. Towards generating colour terms for referents in photographs: Prefer the expected or the unexpected? In *Proceedings of the 9th International Natural Language Generation conference*, pages 246–255, Edinburgh, UK. Association for Computational Linguistics.

A Risks and Ethical Considerations

We do not believe that there are significant risks associated with this work, as we analyze existing models using data with limited scale without contents that might be perceived as hurtful. No ethics review was required.

B Prompts

Our model prompts consist of three parts: (i) a general task instruction, (ii) the formatted utterances for the current item in question, and (iii) a repetition of the set of possible output labels. The prompts are constructed as in the following example from the color grid domain:

> *In this image you can see three color grids. In the following dialogue, the speaker will describe exactly one of the grids. Please indicate to me whether he refers to the left, middle or right grid.*
>
> *speaker: first square is brown*
>
> *Is it the left, middle or right grid?*

In case a chat dialogue between director (speaker) and matcher (listener) occurs, it is included in part (ii) of the prompt as follows:

> *speaker: CENTER BOX is DULL purple*
> *listener: with bright green on left middle or dull green*
> *speaker: BOTTOM RIGHT CORNER is green*

C Implementation Details

For our experiments we rely on models from huggingface. In detail, we used the following models:

- deepseek-ai/Janus-Pro-1B
- deepseek-ai/Janus-Pro-7B
- llava-hf/llava-v1.6-vicuna-7b-hf
- llava-hf/llava-v1.6-vicuna-13b-hf
- llava-hf/llava-v1.6-34b-hf
- llava-hf/llava-next-72b-hf (quantized using the *bitsandbytes* library)
- Qwen/Qwen2-VL-2B-Instruct
- Qwen/Qwen2-VL-7B-Instruct
- Qwen/Qwen2-VL-72B-Instruct-AWQ

To generate responses with our models, we used Python 3.9.20 with the following libraries:

- torch (2.5.1)
- transformers (4.46.2)
- autoawq (0.2.8)
- bitsandbytes (0.44.1)

We used three NVIDIA RTX A6000 GPUs for inference with the 72b models, two GPUs of the same type for LLaVA 34b and a single GPU of the same type for the remaining models. Depending on model size, generating responses took between 10 h and 46 h for the color patch data and between 2.5 h and 13 h for the color grid data.

D Scientific Artifacts

In our work, we mainly used scientific artifacts in the form of publicly available datasets and model implementations (MIT, Apache 2.0, and Llama 2 licenses), as well as Python frameworks and modules

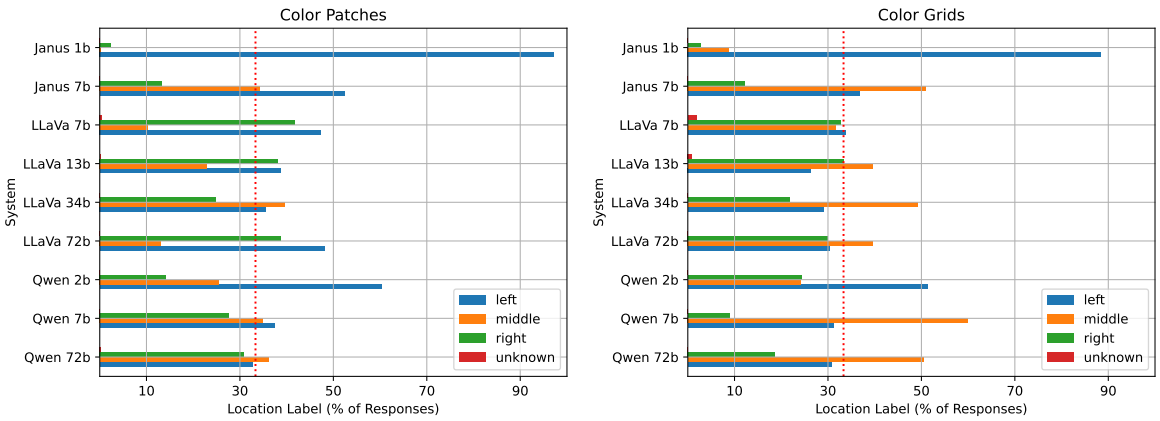


Figure 2: Location biases in model responses for color patches (left) and color grids (right). The vertical dotted red lines denote the approximately equal distribution of target locations in the data.

(cf. Appendix C). The color patch dataset can be downloaded from cocolab.stanford.edu, the color grid dataset is available on GitHub (MIT License). In all cases, we are confident that our work is consistent with their intended use.

Data and code of this study are available (Apache License 2.0) at:

- https://doi.org/10.5281/zenodo.15553655
- https://github.com/clause-bielefeld/mllm-listeners

E Results for Simplified Dialogues

To test effects of linguistic complexity, we test our models on restricted subsets of the color patch and grid datasets, where we only include dialogues with a total length of ≤ 5 words, out of which $\geq 1/3$ have to be included in the following *basic color terms* (cf. Berlin and Kay 1969; Kay and McDaniel 1978): *black*, *white*, *red*, *green*, *yellow*, *blue*, *brown*, *orange*, *pink*, *purple*, *gray/grey*. See Table 3 for detailed results.

F Location Biases in Model Responses

For color grids we test if location terms in input dialogues introduce biases in model responses by defining subsets with utterances that contain exactly one of the labels "left", "middle" or "right". For each of the location labels, we report the proportion of cases (%) where the model predicts the respective label. The results in Table 4 show that all models are affected by location descriptions in input dialogues, albeit to varying degrees. Figure 2 illustrates model biases for target location predictions with respect to the full datasets.

Model	Size (b)	Quant	Color Patches				Color Grids			
			Total	Far	Split	Close	Total	Far	Split	Close
Janus	1	—	**36.9**	**39.1**	**36.1**	**34.8**	33.1	32.7	**33.6**	33.3
	7	—	**75.2**	**89.4**	**68.9**	**61.5**	**41.5**	**41.7**	**39.4**	**43.6**
LLaVA	7	—	**64.5**	**74.2**	**62.2**	**52.9**	37.6	36.7	37.3	**39.7**
	13	—	**62.0**	**70.1**	**59.5**	**52.7**	36.0	35.5	35.1	**38.1**
	34	—	**84.9**	**95.2**	**80.8**	**74.6**	37.5	36.9	36.9	**39.7**
	72	8bit	**65.2**	**77.0**	**60.9**	**52.6**	**40.5**	40.1	**40.5**	**41.1**
Qwen	2	—	**69.9**	**83.7**	**64.4**	**55.7**	**41.0**	**40.8**	**39.9**	**42.8**
	7	—	**87.3**	**95.9**	**84.3**	**77.9**	44.0	45.2	41.7	**44.7**
	72	awq	**90.1**	**96.4**	**88.6**	**82.5**	65.9	66.9	64.9	**65.1**
human	—	—	**91.6**	**97.7**	**90.7**	**83.6**	**94.5**	**96.8**	**93.6**	**90.9**

Table 3: Accuracy scores (%) for all models and datasets, restricted to items with limited description complexity (max. 5 tokens per item, from which min. 1/3 have to be basic color terms). Scores which surpass the full dataset results are highlighted in bold. Note that *close* annotations tend to be longer than *far* and *split* annotations, i.e., total scores are skewed towards easier items.

Model	Size (b)	Quant	Left		Middle		Right	
			Predicted	Accuracy	Predicted	Accuracy	Predicted	Accuracy
Janus	1	—	100.0	33.1	32.2	36.2	13.1	33.7
	7	—	94.4	34.7	95.6	34.7	53.8	41.9
LLaVA	7	—	90.9	37.6	89.2	37.0	94.4	32.2
	13	—	88.7	37.6	95.4	34.8	95.8	32.5
	34	—	92.5	34.3	99.5	33.7	90.5	32.6
	72	8bit	88.1	38.4	87.1	37.2	91.1	32.5
Qwen	2	—	95.9	35.0	67.0	38.3	73.2	38.7
	7	—	76.1	46.4	92.8	37.2	30.7	47.4
	72	awq	47.4	71.0	69.0	55.8	35.9	74.9

Table 4: Location biases for all models in the *color grid* task. For each of the location labels "left", "middle", and "right", we report the proportion of cases (%) where the model predicts the respective label, if it is the only location label in the annotations (33.3 % is the expected result). Accuracy reports the accuracy of predictions (%) in these cases.

	Qwen2-VL	LLaVA-NeXT	Janus-Pro
Vision Encoder	ViT and 2D-RoPE	CLIP-ViT-L, higher-res grids	SigLIP-Large-Patch16-384
Cross-modal Connection	MLP	MLP	MLP
LLM Backbone	Qwen2	Vicuna (3b, 7b), NousHermes2-Yi (34b), and Llama 3 (72b)	Autoregressive Transformer
Training	Pretraining on image-text pairs (ViT), full parameter training and instruction fine-tuning (ViT frozen)	Pretraining cross-modal connection and instruction fine-tuning	Pretraining cross-modal connection, image heads, unified pretraining, and instruction fine-tuning
Data	Extensive and diverse datasets	High-quality visual and multimodal data	Expanded dataset and synthetic data
Special Features	Naive Dynamic Resolution support and Multimodal Rotary Position Embedding (M-ROPE)	Response prompting, dynamic high resolution, and data-efficient	Decoupled Visual Encoding

Table 5: Comparison of aspects of the multimodal large language models Qwen2-VL, LLaVA-NeXT, and Janus-Pro used in this study, based on information available in the publications on these models (see Section 3.1).