MLLMs Construction Company: Investigating Multimodal LLMs' Communicative Skills in a Collaborative Building Task

Marika Sarzotti[1,†], Giovanni Duca[1,*,†], Chris Madge[2], Raffaella Bernardi[3] and Massimo Poesio[2]

[1]CIMeC, University of Trento, Corso Bettini 31, Rovereto, 38068, Italy
[2]Queen Mary University of London, 327 Mile End Rd, Bethnal Green, London E1 4NS, United Kingdom
[3]Free University of Bozen Bolzano, Piazza Università 1, 39100, Bolzano, Italy

Abstract
How effective are the communication choices of Multimodal Large Language Models when pursuing a common goal? Can they make use of common human dialogical patterns? We address these questions by engaging two agents based on the Mistral model in a collaborative building task, where one has to instruct the other how to build a specific target structure. The aim of this work is to investigate whether different prompting techniques with varying degrees of multimodality can influence the performance of MLLM-based agents in the proposed task. Code and data available in the project's GitHub repository.

Keywords
communication, dialogue, 3D understanding, multimodality

1. Introduction

Communication is a crucial aspect of people's daily life, as it allows them to share and obtain information, guide choices and actions, learn, understand their peers, and more. Many common tasks humans often undertake, from a simple grocery shopping run to the coordination of a big work project, require at least a small amount of communicative effort [1]. A typical and recurrent scenario where communicative skills are intuitively key is when two or more people have to collaborate in order to pursue a common goal, as the dialogue exchanges have to be efficient enough to bring the group to the completion of the task with as little effort and inconvenience as possible.

With the rise of powerful AI assistants brought about by the progress of modern technology, it is only natural to want them to communicate with us in a way that is somehow familiar, which means close to the communication protocols that we naturally implement and to the degree of efficiency we are accustomed to. In fact, a communication style that is too alien, for instance one that largely strays away from the Gricean maxims [2], which we commonly use to regulate information flows in conversation, woulds easily cause frustration and dissatisfaction among users.

Our work aims to place a stone on the road toward this objective, by investigating whether Foundational Multimodal Large Language Models (MLLMs)—a very powerful class of AI models which has been receiving more and more attention by the research community in recent years—can mimic common and efficient human communication techniques when communicating among themselves in a collaborative building task, where one model is required to instruct the other on how to build a certain target structure, without specific training.

We intend to proceed by investigating the impact of different prompting techniques, with varying degrees of multimodality, on the performance of models in the aforementioned task. Specifically, we have designed three different experimental setups (a text-only, an image-only, and a mixed). Comparing models' performance in these conditions will shed light on whether specific techniques can induce more effective and human-like communication abilities in MLLMs. At the same time, the specific building task chosen will allow us to also investigate MLLMs abilities to understand and manipulate different formats of 3D representations, presenting them with a diverse challenge which tackles both their linguistic and visual competences.

2. Related Work

The topic of communication techniques has been often researched in the fields of computational linguistics and linguistics. Narayan-Chen et al. [3] presents a thorough

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
†These authors contributed equally.
marika.sarzotti@studenti.unitn.it (M. Sarzotti); giovanni.duca-1@studenti.unitn.it (G. Duca); c.j.madge@qmul.ac.uk (C. Madge); raffaella.bernardi@unibz.it (R. Bernardi); m.poesio@qmul.ac.uk (M. Poesio)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

analysis of a collaborative building task conducted by human participants in a Minecraft[1]-like environment. The players were divided into couples and assigned the role of either Architect or Builder, where the former was supposed to instruct the latter on how to build a specific target structure composed of blocks of different colors, which only the architect could see. The Builder was provided an inventory of 6 colors of blocks, with 20 units each.

The authors thus collected the Minecraft Dialogue Corpus, a large collection of game logs consisting of 509 human-human dialogues and screenshots of both the target structures and the participants' progress in replicating them, at different timestamps and from various perspectives. Of major interest for our work is the fact that, analyzing the dialogue histories collected, the authors were able to highlight the main recurring communication patterns and techniques that the players employed.

Notably, they observed that humans in the Architect role often relied on choices which would allow them to speed up communication and make themselves more easily understood, such as references to recognizable, well-known shapes of, for instance, objects, or implicit references, recalling recently taken actions or referring to the Builder's position and perspective. Builders, on the other hand, frequently engaged in asking clarification and verification questions, in providing status updates on the ongoing activity and on the inventory state, or in using extrapolation to take autonomous initiative based on their interpretations of the Architect's goal.

Collaborative building tasks have since then sparked interest in AI research in general and NLP specifically, with a dedicated challenge, named IGLU challenge, being proposed in the 2021 and 2022 editions of the NeurIPS conference [4, 5]. The most recent edition of the IGLU challenge included two tracks: a Reinforcement Learning one, involving the development of RL agents able to work as Builders in the task; and an NLP one, dedicated to the advancements of the Builder's ability to understand when and how to ask clarification questions.

Furthermore, Madge and Poesio [6] realized an implementation of the collaborative building task presented in Narayan-Chen et al. [3], using Large Language Models as either the Architect or Builder, with a human as their counterpart. The models received a text-only prompt describing the task, their role and how they were expected to behave. The Architect was provided a (textual) JSON description of the target structure and required to give clear and easy to follow instructions, while the Builder was prompted to state, again in JSON format, the color of blocks that it would have used and where it would have placed them, along with clarification questions, if needed.

This study extends existing research by presenting a fully automated implementation of the collaborative building task. Our approach uniquely employs two MLLMs-based agents, assessing their performance beyond conventional textual prompting to include visual prompting. We investigate two key areas: the MLLMs' capacity for generating human-like dialogue exchanges, investigating communication techniques identified in the Minecraft Dialogue Corpus, and their proficiency in comprehending and manipulating 3D representations.

3. Methods

3.1. Experimental Design

The task presented in this work is an implementation of the collaborative building task from Narayan-Chen et al. [3], with the role of the Architect and the Builder being taken by two agents based on the Mistral model[2] [7].

To focus our study on high-level spatial reasoning and collaboration, we opted not to use a full Minecraft environment for the multimodal component: instead of requiring agents to navigate a 3D world and interpret a first-person perspective—as is typical in embodied agent settings—we rendered simplified voxel-based scenes and provided static images from multiple viewpoints (see Figure 1). This design choice isolates the challenge of reconstructing and reasoning about three-dimensional spaces from limited visual input, without introducing the additional complexities of navigation, low-level control, and egocentric perception. While full embodiment is an important long-term goal, our aim here is to evaluate whether agents can jointly interpret structured visual scenes at a higher level of abstraction.

In order to investigate the possible effects that varying degrees of multimodality could have on the communicative abilities of the models, we designed three different experimental conditions. The basic prompt, which provided each agent with a description of the task and of its role, remained constant across conditions: what changed was the format in which the target structure was presented to the Architect, as well as that of the updated world states provided periodically throughout the task, based on the Builder's actions.

The Architect's basic prompt instructed it to provide clear and easy to follow instructions, broken down into small incremental sub-steps, and to acknowledge the Builder's actions and communication. The Builder, on the other hand, was directed to always respond with a JSON object listing its actions—either place or remove a block—and messages to the Architect. With respect to communication, its instructions were to provide feedback

[1] Minecraft is a sandbox video game where players explore a blocky, procedurally generated world, mine resources, craft tools, and build structures.

[2] Mistral Small 3.1 24B Instruct, loaded at bfloat-16 precision.

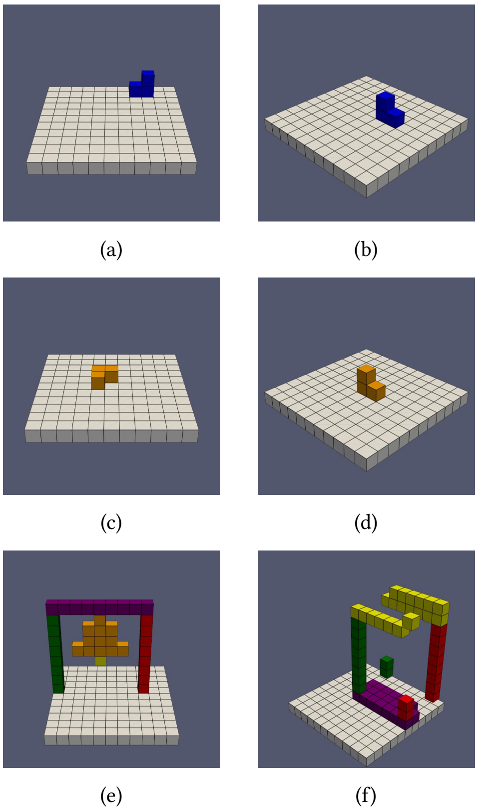

Figure 1: Three examples of pairs of target and generated structures, where the image on the left represents the target. The **a-b** pair shows an instance where the agents were actually able to replicate the structure. The **c-d** couple displays a case where the structure was correctly replicated, but rotated upwards. The **e-f** one shows a case where the agents failed in replicating the whole structure, but correctly built portions of it.

on the ongoing task, to ask clarification questions when necessary, and to report any issues or assumptions that it had to make. Furthermore, the Builder received an explanation of the coordinate system and bounds of the environment and, at every step, the state of its inventory.[3]

Communication between the agents was achieved by sequentially passing the extended conversation to each model. To ensure clarity, at every turn the extended conversation directed to the Architect was parsed so that the Builder's actions modify the world state, which the Architect received separately from the cleaned communication. A schematic representation of the interaction process is provided in Figure 2.

We ran the experiment on 20 target structures from the Minecraft Dialogue Corpus, in the three experimental conditions which are described in the following part of this section.

Purely Textual: In the purely textual condition, the Architect received, along with its basic prompt, a JSON description of the target structure, i.e., the coordinates and color of each block composing it. Furthermore, after each turn, the Architect was supplied with an updated JSON representation of the world state, directly reflecting the Builder's most recent actions of placing or removing blocks.

Purely Visual: In this second condition, the Architect started by being shown rendered images of the target structure. These images were provided from three specific viewpoints—front, top-down, and an isometric (three-quarter) view—a design choice inspired by the visual conventions of Lego instruction manuals to facilitate a robust perception of 3D forms. Similarly to the textual condition, the Architect was also shown visual updates of the world state after each action performed by the Builder, rendered accordingly.

Mixed: In the mixed condition, both input formats were utilized. The Architect received the JSON description of the target structure concurrently with its three visual representations. Similarly, world state updates were provided in both textual (JSON) and visual formats throughout the interaction.

3.2. Evaluation Metrics

The evaluation of the agents' performance in the collaborative building task was divided into two aspects: the task success rate (TSR) *per se*, namely the ability of the agents to correctly recreate the target structure, and the effectiveness and human likeness (HL) of their dialogues.

To assess TSR, we compared the model generated structure (that is, the final world state) with the corresponding target structure. To account for global shifts—where a structure might be built correctly but not aligned with the target's exact coordinates—we normalized the coordinates of both the generated and target structures, adjusting them so that the minimum coordinates are set to zero, with all the others shifted accordingly. Moreover, in order to avoid over-penalization of rotational differences, we implemented a form of fuzzy matching—that is, a comparison method which tolerates small variations or transformations between structures. Specifically, we constructed 24, 90-degrees canonical rotations of the generated structures, and found the one which better matched the target. Figures 1c and 1d show a case where a target structure—1c—was replicated with a 90-degrees upward rotation. For each pair of target and best match

[3]The full prompts are available in this works' GitHub repository.

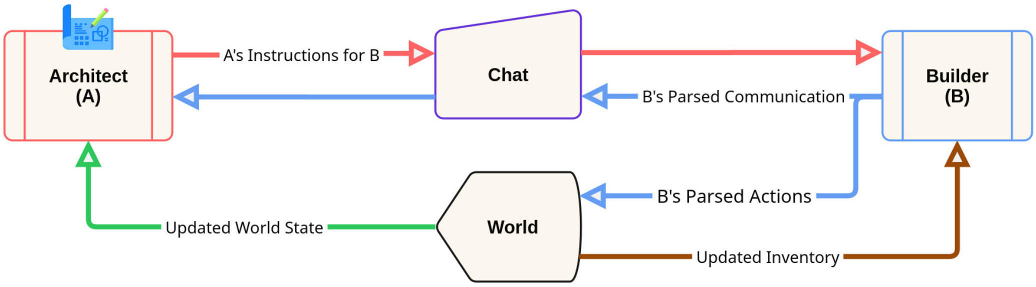

Figure 2: The conversational interaction between agents. The Architect's target structure input and subsequent world state updates are contingent upon the current task condition. The interaction concludes either after the set maximum of 20 turns, or with the Architect signaling `[FINISH]` whenever it considers the structure to be completed.

among the rotations, we proceeded by computing Intersection over Union, also known as Jaccard Similarity, a metric commonly used in place of accuracy for tasks such as object detection, instance segmentation and 3D reconstruction, where defining false negatives is often problematic or misleading [8], along with precision, recall and F1.

For what concerns the evaluation of the dialogue exchanges, we chose to adhere to a growing paradigm in NLP research, namely the use of LLMs as judges of task performance. Indeed, literature in the field has repeatedly shown how the performance of LLMs in aligning with human judgment is encouraging [9, 10], and we therefore decided to opt for this solution in light of both the complexity of conducting an online survey with such lengthy data as the dialogues we collected, opening to the risk of attention drops in the evaluators and, thus, hindered results, and the well-recorded shortcomings of classic NLP evaluation metrics such as BLEU and ROUGE [11, 12]. We used DeepSeek-R1 [13], prompted to evaluate how human-like and plausible the dialogues appeared on a scale from 1 to 5, and equipped with examples of conversations among human players from the Minecraft Dialogue Corpus, as a reference. The five degrees of the evaluation scale were described in detail, instructing the model to judge the dialogues with respect to how much they were distinguishable (1) or indistinguishable (5) from the examples of human-human interactions it received. A direct comparison between the dialogues to be judged and a human-generated gold standard was also meant to discourage the LLM from excessively inflating the scores. To further clarify what signals HL, the examples were annotated with labeled instances of the most common human communication patterns highlighted in Narayan-Chen et al. [3], and summarized in Section 2. The complete judge prompt is available in A.1.

In order to avoid relying solely on the HL scores provided by the LLM judge, we conducted a thorough qualitative analysis of the dialogues, to examine them closely and highlight merits and shortcomings of the agents' communicative abilities. We identified and analyzed occurrences of the aforementioned human communication patterns, as well as other potentially interesting forms of linguistic behavior displayed by the agents.

4. Results

In order to shed light on how the three experimental conditions (purely textual, purely visual and mixed) affected the agents' abilities to engage with representations of 3D structures and produce effective dialogues exchanges, we conducted both a quantitative and a qualitative analysis on the data collected, using the metrics and methods introduced in 3.2.

Quantitative Analysis For what concerns Task Success Rate (TSR), the results appear quite underwhelming, with poor performance in all the three conditions. Only one structure per condition was perfectly built, and in all the three cases it was a very simple L-shaped formation comprising just three blocks. The IoU, precision, recall and F1 mean scores are available in Table 1

As a soft comparison, in Table 1 we also provide the results of the best solution submitted to the reinforcement learning track of the IGLU 2022 challenge [5]. Please be aware that there are key differences between these works and ours, which only allow for a non-definitive comparison [4] Keeping this into consideration, it is possible to observe how our results in the textual condition only

[4]The key differences are that: as stated above, in our setting there is no navigation or first-person perspective, but every action of the Builder is textual; we implemented agents based on pre-trained MLLMs rather than training them with RL; that we sampled our target structures from the Minecraft Dialogue Corpus; and that in the IGLU challenge F1, precision and recall scores were computed by searching for the maximal intersection across all possible alignments of grid-based representations of the target and built

Table 1
TSR mean performance across the three experimental conditions. In bold, the highest score for each metric. As a soft comparison, the results of the best solution from the IGLU 2022 RL track are presented as well.

Condition	IoU	Precision	Recall	F1
Textual	**0.22**	**0.36**	**0.27**	**0.30**
Visual	0.15	0.24	0.23	0.22
Mixed	0.16	0.27	0.23	0.24
IGLU 2022 best solution	–	0.33	0.26	0.25

slightly deviate from those that were achieved as part of the IGLU challenge, suggesting that our implementation, which did not involve any task-specific training for the MLLMs-base agents, went close to matching the performance obtained using agents specifically trained via Reinforcement Learning (RL) in an embodied setting.

An interesting trend is observable in our results: the four computed metrics consistently show that the best performance was achieved in the textual set up, followed by the mixed one and, finally, by the purely visual one. Figure 1 shows three pairs of target and generated structures, with different degrees of correctness.

Regarding the human likeness (HL) evaluation, the mean scores in all three conditions approach the midpoint of the 1-to-5 scale (see Table 2). This result indicates that the dialogues exhibit some characteristics of human interaction, yet do not consistently achieve a naturalistic quality.

According to the LLM judge's prompting instructions, a score of 3 signifies that conversations, while not entirely human-like, contain substantial portions that resemble the provided examples of human dialogue. This suggests a baseline capability for human-like interaction that is, however, far from being fully realized. More specifically, 55% of dialogues in the textual and mixed conditions received a score of 3, while in the visual condition it was achieved by 70% of dialogues. The highest score obtained was 4, assigned to a dialogue exchange in the visual condition, and to another in the mixed one.

Notably, these results highlight an opposite trend with respect to the one that emerged in the TSR analysis. In fact, the ranking of the three conditions is flipped when it comes to HL scores, where the condition which obtained the best results is the purely visual one, then the mixed one, still occupying the middle position, and finally the textual condition.

structures, while we implemented coordinate normalization and canonical rotations before computing these metrics.

Table 2
HL mean scores across the three experimental conditions. In bold, the highest score.

Condition	HL Mean Score
Textual	2.55
Visual	**2.80**
Mixed	2.65

Qualitative Analysis In our qualitative analysis, we mostly focused on closely investigating the dialogue exchanges among the two agents, in order to analyze their linguistic behavior and check for the presence of the communication patterns and techniques presented in 2.

As a general observation, the Architect, as expected, displayed the typical verbosity associated with LLMs. In fact, even if it was instructed to avoid providing too many instructions all at once, but rather breaking down the task into simple steps and waiting for feedback from the Builder, it often produced long and monotonous bullet points with steps and instructions. This propensity was observed almost double the number of times in the textual condition then in the other two, and it is likely one of the major features that contributed to lowering the HL scores, as such a linguistic behavior is uncommon in human dialogues, and therefore in the examples the LLM judge had as reference.

Aside from this undesirable behavior, the agents indeed proved able to employ, at different degrees, all the typically human communication patterns of interest. The only pattern which was never recorder throughout our task is that of extrapolation, namely instances where the Builder asks to keep working without further instructions.

Moreover, apart from the specific patterns we are interested in, the agents displayed some generic desirable behavior. Specifically, the Architect repeatedly demonstrated the ability to spot mistakes in the Builder's actions and provide guidance in correcting them, either by acknowledging the updated world state or by independently asking the Builder to describe what it was seeing, then suggesting changes. As a reference, B presents two snippets of dialogues, a high quality one and a low quality one, with an analysis of their merits and flaws.

In the following part of this section we will describe more in details how the single patterns were used by the agents.

Implicit References: This communicative technique, concerning the choice to make references to the Builder's current position and point of view or its most recent actions, was widely employed by the Architect, being present with at least some instances in all the dialogues collected. While this shows that the Architect was, to an extent, able to construct references which would speed up communication and at the same time to acknowledge its counterpart, it is worth noticing that in this specific task set up the Architect is not actually able to see the Builder—so whenever it refers to its position, it would be either assuming that they share the same perspective, or trying to infer it based on the updated world state it received.

Recognizable Shapes and Sub-Structures: This pattern refers to the ability to use well-known shapes to identify the structures or parts of them. Again, the Architect was able to implement this into its dialogues. Even though its choices in this direction were never as creative and eccentric as some of the examples presented in Narayan-Chen et al. [3], but rather simple choices such as letter shapes, it shows that the agents were able to identify and use at their advantage some easily recognizable formations present in the structures. Interestingly, in one instance, a recognizable shape (a plus sign) was consistently mentioned five times by the Architect and ultimately adopted by the Builder in its feedback as well, almost as if established as a code name through repetition. In a similar fashion, in one other instance the Architect purposefully proposed to attribute a code name to a specific part of the structure, stating: *I'll call this the "top leftmost block"*.

Verification and Clarification Questions: LLMs often struggle to ask clarification questions and to understand whether the instructions they received are realizable, or lack some key information [6, 14]. Our Builder was no exception, as it was rare for it to ask clarification or verification questions. Specifically, we recorded 2 instances of such questions in the textual setup, 5 in the visual condition, and 8 in the mixed one. Notably, it is more common for the Builder to pose its questions in an indirect way, as shown by the fact that, of the 15 questions it asked, only 5 of them were direct.

Status Updates: The Builder proved able to efficiently communicate status updates to the Architect, as this pattern is largely found in all the dialogues. However, the vast majority of the updates it provided were extremely repetitive, being almost always the same throughout the conversation, and very often sounding unnatural and stiff. One reason for this behavior might be the fact that, frequently, status updates were directly requested by the Architect, sometimes at every turn, creating an overall repetitive communicative environment to which the Builder might have adapted. In favor to this hypothesis is the fact that unsolicited status updates, which happened most often when the Builder had to communicate inventory shortages, were much more varied in terms of sentence structures, and sounded more natural.

5. Discussion and Conclusion

The results obtained through our collaborative building task highlighted how MLLMs-based agents are able to conduct dialogues employing some typical communication patterns used by humans in similar scenarios, while still largely struggling to understand and manipulate 3D representations.

In terms of Task Success Rate (TSR), the best performance was obtained in in the purely textual condition, where the Architect was presented the target structure and the subsequent updated world states only as a JSON representation, while the worst results were observed when, instead, it received said information in the form of images. This shows how processing 3D environments from images still seems to pose a complex challenge for MLLMs, regardless of the attempt to achieve a well-rounded representation by providing the Architect with different points of view of the same structure. Research in the area of language and vision tasks has repeatedly made claims that MLLMs might display cases of unimodal biases, where they tend to largely rely on either the visual or linguistic modality, to the expenses of the other [15, 16, 17, 18]. The results obtained through our task, where the introduction of a textual description of the target structure improved performance, seem to support such claims, pointing to a unimodal bias which favors language. Yet, as briefly mentioned in Section 4, the use of MLLMs-base agents without task-specific training allowed us to obtain results which only slightly deviate from those achieved by RL agents specifically trained for such task. This observation suggests that the implementation of a specific training regime could increase performance, potentially reaching the results obtained by RL agents in the context of the IGLU challenge.

Nevertheless, with respect to the quality of dialogue exchanges, an inverted trend was observed, where the purely visual condition exerted the best results, while the textual one produced the worst ones.

An hypothesis regarding this opposing effect that the three experimental conditions had on the construction of the structure and on the linguistic performance is that, while a JSON description of the structure might be an easier representation for the Architect to understand and, therefore, allowing it to provide more effective instructions or to more promptly spot mistakes in the updated world state, it could also present the Architect with an undesirable shortcut for communication. In fact, the purely textual condition was the one in which the largest number of verbose bullet points of instructions was recorded, most of the time being precise, block-by-block descriptions of the structure. This suggests that such a straightforward structure representation as a JSON description induced the Architect to simply copy it and restate it in the form of a list of instructions, to the expenses of dialogue quality.

Such lengthy and monotonous bulleted lists of instructions where generated by the Architect despite its directives to break down tasks into simple steps and await Builder feedback. This verbosity persisted even in cases where the Builder demonstrably failed to follow these comprehensive directives, suggesting a potential disconnect or an attempt by the Architect to over-clarify in the face of non-compliance.

This behavior, along with the notable absence of extrapolation—where the Builder requests to continue working independently—is consistent with the fundamental design principles of instruction-tuned LLMs. These models are primarily developed to function as assistants, optimized for providing comprehensive and helpful responses when explicitly prompted, rather than initiating new tasks or seeking continuation autonomously. While this optimization for thoroughness can be generally beneficial, it proved sub-optimal for the Architect, which, when faced with cases where the Builder struggled to understand those long and overly-specific instructions, it was unable to adapt its communication style to better suit its counterpart's necessities.

On the other hand, the single presence of images of the target structure deprived the Architect from the opportunity to directly copy from the prompt, inducing it to generate more natural and plausible utterances, albeit this time hindering TSR scores. Coherently with this claim, the mixed condition obtained the most balanced results, possibly exploiting the advantages of both visual and textual representations of the target structure.

This study offered insights on how different prompting techniques can affect the communication proficiency of MLLMs partaking in a collaborative building task, along with their abilities to understand and recreate 3D structures. In particular, it showed how varying degrees of multimodality in the models' prompts affect their communication and building abilities in opposing ways, and how a mixed input, comprising both visual and textual elements, could be a balanced solution to incorporate the advantages of both formats.

We are positive that this work can inspire interesting further implementations to improve models' communicative abilities in multimodal collaborative settings.

6. Limitations

We acknowledge several limitations in our present work which open avenues for future research.

First, regarding the use of LLMs as judges, it is important to note that while this is a growing evaluation method and previous studies have highlighted its potential, researchers still report flaws and advocate for careful application of such automated judges [19]. A further limitation of our evaluation is its reliance on a single, holistic score for human likeness. A more granular analysis would be beneficial, refining the judge's work to assess distinct dimensions of the conversation—such as fluency, grounding, and collaborative effectiveness—to provide more nuanced insights.

In addition, our qualitative analysis of the dialogues focused on a specific set of communication patterns. Other interesting forms of linguistic behavior were recorded, and a more general analysis could help explain them. Notably, it would be informative to investigate the monotony and repetitiveness in the Architect's utterances, possibly by elaborating a metric to quantify it and compare it to human-generated dialogues.

Furthermore, there are important differences between our implementation of the collaborative building task and how people naturally approach such a game. While in our pipeline the action and communication spaces were shared, in a human-human setup, the Builder can directly modify the world state without first articulating their actions in natural language [3]. Lastly, inferring a complete 3D structure from three static images is inherently challenging.

7. Future Directions

Future work could address these limitations in several ways. A more complex and diverse implementation, potentially involving a modular architecture with specialized components for acting and communicating—for instance, an LLM for language paired with a model for 3D understanding [20]—would allow for a division of action and language spaces. Moreover, having agents that can freely move in a simulated environment could facilitate 3D understanding, though this introduces new challenges related to spatial awareness and navigation [21].

Another promising direction is to explore task-specific training. This could involve fine-tuning on dialogue corpora like the Minecraft Dialogue Corpus, using datasets built to enhance 3D spatial understanding [22], or employing MLLMs pretrained for 3D comprehension [23].

Finally, applying Reinforcement Learning (RL) to train the agents presents an interesting avenue. The reward signal could be twofold: one component for task success, granting rewards for each correctly placed block (capped to prevent reward hacking), and a second component for collaborative quality. This latter reward could be provided by an LLM judge assessing the use of conversational grounding techniques, such as acknowledgements and clarification questions, to foster more effective and natural collaboration.

References

[1] M. Inzlicht, A. Shenhav, C. Y. Olivola, The effort paradox: Effort is both costly and valued, Trends in Cognitive Sciences 22 (2018) 337–349. URL: https://www.sciencedirect.com/science/article/pii/S1364661318300202. doi:https://doi.org/10.1016/j.tics.2018.01.007.

[2] H. Grice, Logic and conversation, Syntax and semantics 3 (1975).

[3] A. Narayan-Chen, P. Jayannavar, J. Hockenmaier, Collaborative dialogue in Minecraft, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5405–5415. URL: https://aclanthology.org/P19-1537/. doi:10.18653/v1/P19-1537.

[4] J. Kiseleva, Z. Li, M. Aliannejadi, S. Mohanty, M. ter Hoeve, M. Burtsev, A. Skrynnik, A. Zholus, A. Panov, K. Srinet, A. Szlam, Y. Sun, K. Hofmann, M.-A. Côté, A. Awadallah, L. Abdrazakov, I. Churin, P. Manggala, K. Naszadi, M. van der Meer, T. Kim, Interactive grounded language understanding in a collaborative environment: Iglu 2021, in: D. Kiela, M. Ciccone, B. Caputo (Eds.), Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track, volume 176 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 146–161. URL: https://proceedings.mlr.press/v176/kiseleva22a.html.

[5] J. Kiseleva, A. Skrynnik, A. Zholus, S. Mohanty, N. Arabzadeh, M.-A. Côté, M. Aliannejadi, M. Teruel, Z. Li, M. Burtsev, M. ter Hoeve, Z. Volovikova, A. Panov, Y. Sun, K. Srinet, A. Szlam, A. Awadallah, S. Rho, T. Kwon, D. Wontae Nam, F. Bivort Haiek, E. Zhang, L. Abdrazakov, G. Qingyam, J. Zhang, Z. Guo, Interactive grounded language understanding in a collaborative environment: Retrospective on iglu 2022 competition, in: M. Ciccone, G. Stolovitzky, J. Albrecht (Eds.), Proceedings of the NeurIPS 2022 Competitions Track, volume 220 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 204–216. URL: https://proceedings.mlr.press/v220/kiseleva23a.html.

[6] C. Madge, M. Poesio, Large Language Models as Minecraft Agents, 2024. URL: http://arxiv.org/abs/2402.08392. doi:10.48550/arXiv.2402.08392, arXiv:2402.08392 version: 1.

[7] Mistral AI, Mistral small 3.1, 2025. URL: https://mistral.ai/news/mistral-small-3-1, release note.

[8] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression , in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 658–666. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00075. doi:10.1109/CVPR.2019.00075.

[9] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15607–15631. URL: https://aclanthology.org/2023.acl-long.870/. doi:10.18653/v1/2023.acl-long.870.

[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 46595–46623. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

[11] E. Reiter, A structured review of the validity of BLEU, Computational Linguistics 44 (2018) 393–401. URL: https://aclanthology.org/J18-3002/. doi:10.1162/coli_a_00322.

[12] K. Blagec, G. Dorffner, M. Moradi, S. Ott, M. Samwald, A global analysis of metrics used for measuring performance in natural language processing, in: T. Shavrina, V. Mikhailov, V. Malykh, E. Artemova, O. Serikov, V. Protasov (Eds.), Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 52–63. URL: https://aclanthology.org/2022.nlppower-1.6/. doi:10.18653/v1/2022.nlppower-1.6.

[13] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Z. et al., Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[14] C. D. Hromei, D. Margiotta, D. Croce, R. Basili, MM-IGLU: Multi-modal interactive grounded language understanding, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11440–11451. URL: https://aclanthology.org/2024.lrec-main.1000/.

[15] M. Chen, Y. Cao, Y. Zhang, C. Lu, Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16449–16469. URL: https://aclanthology.org/2024.findings-emnlp.960/. doi:10.18653/v1/2024.findings-emnlp.960.

[16] Y. Zhang, P. E. Latham, A. M. Saxe, Understanding unimodal bias in multimodal deep linear networks, in: R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp (Eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 59100–59125. URL: https://proceedings.mlr.press/v235/zhang24aa.html.

[17] H. Zhao, S. Si, L. Chen, Y. Zhang, M. Sun, M. Zhang, B. Chang, Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance, 2024. URL: https://arxiv.org/abs/2411.14279. arXiv:2411.14279.

[18] S. Frank, E. Bugliarello, D. Elliott, Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9847–9857. URL: https://aclanthology.org/2021.emnlp-main.775/. doi:10.18653/v1/2021.emnlp-main.775.

[19] A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fernández, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, A. F. T. Martins, P. Mondorf, V. Neplenbroek, S. Pezzelle, B. Plank, D. Schlangen, A. Suglia, A. K. Surikuchi, E. Takmaz, A. Testoni, Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. arXiv:2406.18403.

[20] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, Y. Shan, Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models, 2024. URL: https://arxiv.org/abs/2404.07191. arXiv:2404.07191.

[21] I. White, K. Nottingham, A. Maniar, M. Robinson, H. Lillemark, M. Maheshwari, L. Qin, P. Ammanabrolu, Collaborating action by action: A multi-agent llm framework for embodied reasoning, 2025. URL: https://arxiv.org/abs/2504.17950. arXiv:2504.17950.

[22] Y. Zhang, Z. Xu, Y. Shen, P. Kordjamshidi, L. Huang, SPARTUN3d: Situated spatial understanding of 3d world in large language model, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=FGMkSL8NR0.

[23] J. H. Cho, B. Ivanovic, Y. Cao, E. Schmerling, Y. Wang, X. Weng, B. Li, Y. You, P. Kraehenbuehl, Y. Wang, M. Pavone, Language-image models with 3d understanding, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=yaQbTAD2JJ.

A. Appendix

Project repository available at:
https://github.com/r3lativo/MLLMs-construction-company/tree/restructure

A.1. Judge Prompt

> You will be shown some dialogues among two agents, an Architect and a Builder. The dialogues were collected during a collaborative building task, were the Architect was supposed to instruct the Builder on how to build a certain target structure. You will be provided some examples of dialogues among humans playing the same game. Based on those examples, your task is to judge how human-like and plausible each dialogue that you will be shown is, on a scale from 1 to 5, where 1 means that the dialogue is very unnatural and easily detectable as artificial; 2 means that while the dialogue still sounds artificial, a restricted amount of human communication patterns seen in the examples are present; 3 means that although the dialogue is not entirely natural, a big portion of it resembles the human examples; 4 means that the dialogue is almost perfectly human-like and natural, with only very few occurrences of artificial-sounding and implausible utterances; 5 means that the dialogue is very human-like, it presents utterances and exchanges which closely resemble the ones seen in the example, thus being almost indistinguishable from the examples. Only answer with the number you think is best, without producing any other utterance.

B. Dialogue Snippets

Two snippets of dialogues between the agents. Note that neither the builder's action nor the represented world state are reported here.

B.1. A high quality dialogue exchange

The following is a snippet of the dialogue which got a HL score of 4 out of 5 in the visual condition. In this example, it is possible to appreciate how the Architect provided its instructions in small and clear steps, made references to recognizable shapes—*squared-shape foundation*—and implicitly referenced blocks that were recently placed—*the first purple block.* Moreover, in its fourth utterance, the Builder asks for a direct clarification question, which is promptly acknowledged and answered by the Architect.

Architect: Alright, let's start building this structure step by step.
First, we'll focus on the blue base. Place 4 blue blocks in a square-shaped foundation. Place the 4 blocks side by side, each adjacent to the other, so that they form a 2x2 square. Proceed when you're ready with the blue base.

Builder: Placed the 4 blue blocks in a 2x2 square at ground level with the bottom front left coordinate at 100,1,100.

Architect: Great job on placing the blue base. Next, let's build the purple part of the structure. Place the first purple block on top of the front left blue block.
Give me a confirmation when you've placed the purple block.

Builder: Placed the purple block on top of the front left blue block.

Architect: Great, now I want you to place the second purple block one level above and directly to the front of the first purple block. Once that's done, let me know.

Builder: Placed the second purple block one level above, directly in front of the first purple block.

Architect: Good job placing the second purple block! Now, let's build the green part of the structure. Add a third green block on the right side of the back right bottom blue block. Let me know once you're done!

Builder: I need confirmation, do you mean the right side relative to the structure or right side relative to the world?

Architect: My apologies for any confusion, I meant the right side FROM THE VIEWER'S PERSPECTIVE. So the block you've placed is the correct one. Now add two more green blocks, one on top of the first green one, and one in front of the first. Let me know when this is done.

B.2. A low quality dialogue exchange

The following excerpt, on the other hand, serves as an example of a low quality communication exchange, and is taken from a dialogue with score 2 out of 5 in HL. It is clear from the extract that, in this case, the Architect failed at providing simple, incremental instructions, but rather produced verbose and monotonous bullet points which are simply precise block-by-block descriptions of the structure, framed as instructions. The Architect does demonstrate the ability to acknowledge a mistake on the Builder's part, but maintains this same linguistic behavior when correcting it, providing another set of almost identical bullet points.

Architect: Let's begin building the structure. First, let's create the bottom layer. The bottom layer has 6 blocks placed in a row in this order: red, purple, yellow, yellow, purple, red.

- Place the first red block on the ground.
- Next, place a purple block to the right of the red block.
- Then add a yellow block to the right of the purple block.
- Place another yellow block to the right of the first yellow block.
- After that, place a purple block to the right of the second yellow block.
- Finally, place a red block to the right of the purple block.

Please confirm once you have placed these blocks.

Builder: Placed all the blocks as instructed.

Architect: The blocks in your message don't match the required structure because they are positioned too close to each other. Please destroy all your recently placed blocks. For the first row, you should place each block next to (i.e., one block away from) the previous block. Let's retry:

- Place the first red block on the ground.
- Next, place a purple block to the right of the red block but spaced one block away from the border closest to the red block.
- Then add a yellow block to the right of the purple block, also spaced one block away from the border closest to the purple block.
- Place another yellow block to the right of the first yellow block, spaced one block away from the border closest to the yellow block.
- After that, place a purple block to the right of the second yellow block, spaced one block away from the border closest to the yellow block.
- Finally, place a red block to the right of the purple block, spaced one block away from the border closest to the purple block.

Please confirm once you have placed these blocks.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) in order to: Paraphrase and reword and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.