Dialogue with LLaVA: does it "understand" the pragmatics of the MeetUp task?

Nikolai Ilinykh and Simon Dobnik
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
name.surname@gu.se

Introduction Recently large language-and-vision models like BLIP-2 (Li et al., 2023) have achieved good performance on various multi-modal tasks. These models are initially pre-trained on a large number of image-text pairs to capture general multi-modal understanding and then fine-tuned for specific downstream tasks. Models like LLaVA (Liu et al., 2023) are fine-tuned with *prompts* that *instruct* the model to perform a task. Using these models for different tasks requires rephrasing the tasks into the specific language and format that the models understand. Visual dialogue is a very challenging multi-modal task ,and MeetUp (Ilinykh et al., 2019) is an example of such a task and dataset. In this task two players are placed in a virtual house environment represented as 2D images and must find each other. They can use a chat interface to communicate and execute commands to change images, i.e., move from one room to another. The collected dataset introduces many challenges for multi-modal models as they must consider both visual and textual history at each moment in the dialogue. In this paper we examine the performance of the large multi-modal model LLaVA (Liu et al., 2023) for the task of next utterance generation in MeetUp. This task was initially introduced and described as part of the Visual Dialogue Generation Challenge (Ilinykh and Dobnik, 2023). We prompt LLaVA with two different prompts which are structurally similar but vary in the game-relevant information that they include. By evaluating the quality of generated utterances and the model's perplexity in predicting human-generated utterances, we draw conclusions about LLaVA's ability to be used in the context of such visual dialogue task as MeetUp.

Prompting LLaVA for turn generation LLaVA (Liu et al., 2023) is a large transformer-based model designed to assist humans in completing various visual tasks. The model consists of two large pre-trained transformers: CLIP (Radford et al., 2021) and Vicuna (Chiang et al., 2023), allowing it to use their pre-trained knowledge. One of LLaVA's strengths is its strong performance in tasks involving conversation, detailed image descriptions, and complex reasoning. The model has been fine-tuned on automatically generated instruction-following data based on the image-text pairs from the MS COCO dataset (Lin et al., 2014). For the MeetUp task we design two prompts (shown in the Appendix) following the style of the instructions used to fine-tune LLaVA. Prompt **A** describes the input to the model along with extra information about the game, such as "the players are trying to find each other". Prompt **B** describes *only* the input to the model with minimal information about the game. We use either of these prompts to generate each next utterance in the MeetUp dialogues. The model receives an image showing the rooms visited by the players up to the current timestep as visual input. On the textual side the model is provided with the current chat history. We tested LLaVA on predicting 5695 extracted turns that contain utterances. We did not explore other types of turns such as those with navigation commands or private chat messages to the game bot. We measure the quality of model-generated messages by comparing them with the ground-truth messages using several classic n-gram based metrics (e.g., BLEU, ROUGE, and METEOR, for a survey see Sai et al. (2022)) as well as BERTScore (Zhang et al., 2020).

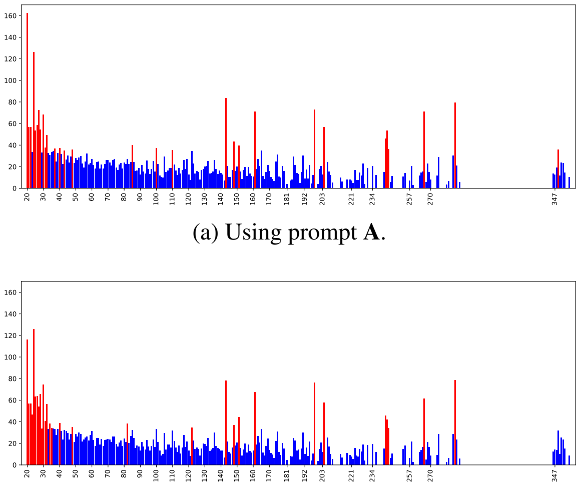
As shown in Table 1 prompt **A** leads to slightly higher scores in automatic evaluation than prompt **B**. All scores are low and close to each other, indicating that the generated utterances are very distant from the human ground-truth. Higher BERTScore values, which are closer to 1.0 (indicating the highest cosine similarity), show that the generated descriptions are very similar to those generated by humans in terms of their semantics. One possible explanation for this, which requires additional

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

Prompt type	BLEU-1	BLEU-4	ROUGE	METEOR	BERTScore
Prompt **A**	6.46	0.12	8.10	15.65	0.78
Prompt **B**	6.40	0.08	7.96	16.46	0.79

Table 1: Automatic evaluation of quality of generated turns given different prompts. The scores are averaged.

testing, is that the generated descriptions are thematically within the domain of the dataset (e.g., describing images), but it is unclear how contextual and correct these descriptions are.



(a) Using prompt **A**.

(b) Using prompt **B**.

Figure 1: PPL score (vertical axis) against messages per turn (horizontal axis, visualised as turn id). The scores per turn are averaged across all dialogues. 20 turns with highest PPL score are highlighted in red.

We calculate the perplexity of the model against human-generated utterances. Perplexity (PPL) is defined as $\exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(w_i \mid w_1,\ldots,w_{i-1})\right)$, where N is the number of words in the sequence and $p_\theta(w_i \mid w_1,\ldots,w_{i-1})$ is the probability of the i-th word given the preceding words according to the model parameterised by θ (LLaVA). According to Figure 1, the model is most uncertain about predicting ground-truth utterances at the beginning of dialogues, with occasional high uncertainty towards the end of the dialogue. An excerpt from the MeetUp corpus in Example (1) with specified turns shows that these parts typically include greetings ("What's up" in t-21), negotiation of game strategies, and planning ("Oh k let me look for it" in t-29, other examples in t-107 and t-120), while the middle of the dialogues includes more descriptions of visual content ("a stand alone sink on the left" and others in t-36–t-50). MeetUp dialogues also contain a larger number of utterances describing visual content, as indicated by Figure 2, which shows that most of the turns with text messages appear between turns 30 and 50. The data indicates that the model cannot understand such parts in MeetUp dialogues, which are not directly related to its visual content. One way to improve the model's understanding of the game's information and instructions is through prompt engineering. However, there is no clear evidence that changing the prompt affects the perplexity of the model, as both prompts result in similar average perplexity across turns (22.87 for prompt **A**, 22.51 for prompt **B**). Future work should explore other prompts.

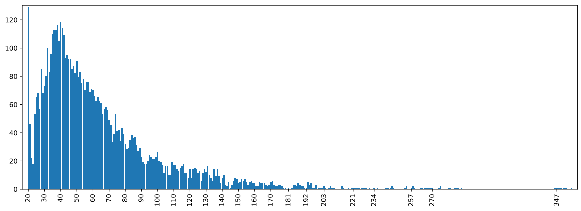


Figure 2: The number of turns with utterances. Turn ids are shown on the horizontal axis, the number of instances of each turn is displayed on the vertical axis.

Conclusion We explore prompt engineering for large language-and-vision models in the complex domain of visual dialogue tasks. Our analysis shows that LLaVA can be used to generate utterances in collaborative visual dialogue tasks such as MeetUp. Future work will examine how much game-relevant information prompts should include or if simply describing the input to the model and asking it to "generate a next response, given the input" is sufficient. We will also focus on the evaluation component of generated utterances by examining characteristics relevant to different games in a dialogue.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An opensource chatbot impressing gpt-4 with 90%* chatgpt quality.

Nikolai Ilinykh and Simon Dobnik. 2023. The VDG challenge: Response generation and evaluation in collaborative visual dialogue. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 23–30, Prague, Czechia. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! A corpus of joint activity dialogues in a visual environment. *CoRR*, abs/1907.05084.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix: LLaVA prompts The prompts were designed with two goals. First, we aimed for prompts similar in structure to those used by LLaVA. Second, we created one prompt that simply describes the model's input with basic game context and another that provides more game-relevant information. The differences between the prompts below are highlighted in bold.

Prompt **A**:

> You are a helpful language and vision assistant. You see a chat between two people, A and B. **They are playing a game in which they are trying to find each other in a house.** What you see are the pictures of each **room** they have visited. The **rooms** visited by person A are shown in the top row, and the **rooms** visited by person B are shown in the bottom row. Pictures in each row are arranged in sequence from left to right, representing the order in which they were taken. Person A is currently **in the room** shown in the rightmost picture from the top row, and person B is currently **in the room** shown in the rightmost picture from the bottom row. A and B are having a chat and are trying to ensure that they are in the same room, i.e., **they have to see the same picture**. Each player does not see what the other player sees. Sometimes the chat is empty, which means that the players have not written any messages yet.
> What do you think is the next message based on the information you have about the game, the players, the rooms they have visited, and their chat?
> CHAT:

Prompt **B**:

> You are a helpful language and vision assistant. You see a chat between two people, A and B. You also see **pictures**. **Pictures** seen by person A are shown in the top row of the image, and **pictures** seen by person B are shown in the bottom row. Pictures in each row are arranged in sequence from left to right, representing the order in which they were seen. Person A is currently **seeing the rightmost picture** from the top row, and person B is currently **seeing the rightmost picture** from the bottom row. A and B are having a chat and are trying to ensure that they **see the same picture**. Each person does not see what the other person sees. Sometimes the chat is empty, which means that A and B have not written any messages yet.
> What do you think is the next message based on the information you have about the situation, A, B, and pictures they have seen, and their chat?
> CHAT:

Excerpt from a MeetUp dialogue

(1) t-20 B: What's up
t-21 A: hi
t-28 A: i have found a bathroom
t-29 B: Oh k let me look for it
t-36 A: it has white bathtub in the back of the room, white shower curtain with blue patterns
t-41 A: a stand alone sink on the left
t-50 A: there is tile on the wall with small squares ranging in color between white and brown
t-51 B: I think I found it
t-52 B: toilet
t-53 B: towel rack
t-54 A: no i dont think i see a towel rack
t-69 B: oh
… …
t-106 A: lets meet at the bathroom with pink towels
t-107 A: it is more easily identifiable
t-120 B: ok im there