When social robots see our sketches: evaluating human perception of a robot and a VLM model performance in a drawing task

Viktoria Paraskevi Daniilidou[1], Nikolai Ilinykh[1] and Vladislav Maraev[1,2]

[1]Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg,
[2]ISIR - Institut des Systèmes Intelligents et de Robotique, Sorbonne Université

danivictori26@gmail.com, nikolai.ilinykh@gu.se, vladislav.maraev@gu.se

Abstract

We introduce a multimodal framework for interactive drawing in a robot-assisted second language learning scenario. In this scenario, humans are asked to draw objects and spatial relations between them, while a social robot that uses a vision-language model (VLM) to analyse whether the drawings are correct. The correctness decision that is passed to the human is coming from a Wizard-of-Oz (WoZ) setup. Therefore, we use it to indirectly evaluate the quality of VLM's predictions. We show that the task is very challenging for a VLM and approaching evaluation of VLM's performance is important: focusing on the correctness of prediction of certain features (objects, relations) provides a different evaluation picture from when the model is evaluated on prediction of the content of the image as a whole. We also examine how the appearance of the social agent and the type of feedback influence perception of the agent by the participants through a questionnaire. The comparison of verbal feedback, generated by the large language models, against simple pattern-based feedback did not show any significant effects whereas the robot's appearance change indicated significant difference in user ratings concerning naturalness of the agent and its social presence.[1]

1 Introduction

Both children and adults use drawing and sketching as everyday tools to describe or explain things. Previous work has shown that drawing can serve both as a measure of cognition and as a generative learning activity, engaging perceptual, motor, and memory systems in ways that support learning. However, the benefits of drawing depend on guidance and the conditions under which drawings are produced (Fan et al., 2023). Social robots can act as tutors or peers and have been shown to enhance learning outcomes in classrooms (Belpaeme et al., 2018). This motivates exploring drawing in interaction with an embodied system that can guide the learner.

To provide useful guidance, an embodied system must have a component that can recognise objects being drawn and their spatial arrangement. Vision-and-language models (VLMs) are well suited for this task, having shown strong performance in integrating visual perception with linguistic understanding in tasks such as image captioning (Bernardi et al., 2016) and visual question answering (Agrawal et al., 2016). However, these models are typically evaluated on datasets that depict static, often fully rendered images of objects or scenes such as MSCOCO (Lin et al., 2015) or Conceptual Captions (Sharma et al., 2018). What remains unexplored is how such models operate on *continuously evolving visual input* such as a sketch being drawn over time. This type of visual information is particularly challenging to interpret as the same object undergoes multiple transformations during the drawing process. Such dynamic, process-oriented input is especially relevant in contexts like language learning or human-robot interaction, where visual meaning unfolds progressively rather than being presented as a completed image.

In this study we examine drawing as a guidance-based interaction task and evaluate the suitability of VLMs for interpreting continuously evolving visual input. We focus on a foreign-language learning scenario in which participants draw simple object configurations corresponding to spatial prepositions (e.g., "an apple in the box") while interacting with a robot tutor. This setup allows us to explore two complementary aspects: how drawing functions as a medium for guided interaction, and how effectively a VLM can ground the evolving sketch in language as it is being produced. In parallel, we investigate how integrating VLM-generated feedback into the robot's responses influences the learner's

[1]Code and data are available at: https://github.com/Viktoriada26/DrawingWithaSocialRobot

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 239–252
February 26–March 1, 2026.

perception of the interaction.

Specifically, we address the following research questions:

Q1 What is the performance of a VLM as an interpreter of sketches in a human–robot interaction setup?

Q2 How does VLM-generated feedback influence the perception of a social robot?

This paper reports the following contributions:

1. We build a **multi-modal pipeline for interactive drawing** which pairs a social robot with a VLM to judge spatial relation in free hand sketches. We evaluate this pipeline with English-speaking participants engaged in a task of learning spatial prepositions in Greek.

2. We collect a **dataset of real-time sketches**, covering six spatial prepositions and a fixed object set. We store a human-provided ground-truth label per drawing indicating whether the intended relation is present in the image.

3. We use our dataset of real-time sketches **to evaluate the performance of a VLM** in a zero-shot scenario.

4. We report **empirical insights into human–robot drawing practice** in the context of using VLMs to provide feedback on the correctness of objects and relations in sketches.

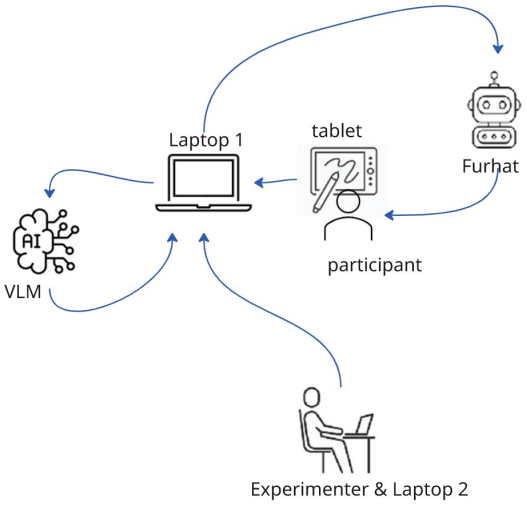


Figure 1: Experimental setup: the participant draws on a tablet while Laptop 1 executes the pipeline. A VLM processes images and sends outputs back to Laptop 1. Furhat conveys VLM-generated and scripted feedback to the participant, while a remote experimenter manually annotates and controls the experiment.

2 Related work

Vision–language models VLMs are neural networks that combine visual and linguistic processing to perform different multi-modal tasks such as object recognition (Russakovsky et al., 2015), image captioning (Bernardi et al., 2016), or video question answering (Nguyen et al., 2024). They are typically trained on large image-text datasets and integrate vision and text encoders with cross-attention mechanisms to align the two modalities (Li et al., 2025). Earlier models like CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023) have achieved strong performance on static image tasks, while ever larger models like Qwen2-VL (Wang et al., 2024) demonstrate strong multimodal reasoning and can follow natural-language instructions through prompting. While these models perform well on static image tasks, little is known about their ability to interpret dynamically evolving visual input, such as sketches drawn over time. Fadeeva et al. (2024) has shown that a VLM can process images together with time-ordered pen strokes in online handwriting. However, our setting differs, as we focus on drawings of objects and evaluate whether the spatial relations between them match the instruction. Given a text instruction (e.g., "draw an apple on the box") and a learner's sketch, the model must decide whether the depicted scene satisfies the stated relation.

Robot-assisted language learning Robot-Assisted Language Learning (RALL) investigates how robots can support the acquisition of linguistic and communicative skills (Randall, 2019). Most studies have focused on vocabulary learning (Van den Berghe et al., 2019), while grammar and spatial relations remain less explored. To our knowledge, interactive drawing with a robot haven't previously been explored. Studying prepositions through drawing extends RALL research into multimodal, spatially grounded interaction, consistent with the current focus on embodied, multimodal dialogue. Following prior RALL practice, our setup combines autonomous components with a Wizard-of-Oz (WoZ) evaluation to maintain reliable feedback delivery.

3 Methodology

This project introduces a pipeline for a robot-assisted multimodal language learning task and integrates (i) a social robot (Furhat), (ii) a vision-language model (VLM), (iii) a voice-based interface for interaction, (iv) a canvas-based drawing tool and (v) a real-time WoZ evaluation mechanism performed by the experimenter. The task of the participant is to learn spatial relation expressions in a non-native language.

Our system can be defined as task-oriented, because its goal is to guide the user through drawing tasks and provide feedback. The user provides visual input (a drawing), while the system responds via spoken output (voice), matching the definition of multimodal dialogue systems given by Oviatt (2007). The system captures the sketch, interprets it, and responds accordingly. The present study uses Furhat, a back-projected anthropomorphic robot head (Al Moubayed et al., 2012), to deliver feedback concerning the drawing.

The WoZ-guided participant-robot interaction is organized into structured stages each represented by a distinct state or group of nested states. These stages include: (1) *instruction*, where the user is told what to draw (e.g., "Draw an apple on a box"); (2) *drawing*, where the user makes the sketch; and (3) *feedback*, where the system delivers either scripted or generated by a VLM, depending on the experimental condition. In parallel, the VLM incrementally analyses each drawing and predicts correctness, without influencing the interaction. Transitions between states are triggered by discrete events such as drawing update, model responses, or timed delays.

3.1 Dialogue system and a social robot

The system uses a voice-based interaction layer that serves mainly to provide instructions and feedback, while the participant draws. Our design minimises spoken user input while still allowing the system to provide continuous verbal guidance. By treating the user's drawing actions as responses the system maintains an interactive flow without requiring additional speech. This setup helps the user remain focused on the drawing activity, making the interaction smoother and supporting engagement with the drawing task.

The dialogue manager is based on the *statecharts* formalism which extends finite-state machines with hierarchy and concurrency, explicit events and guards, actions, timers, and history (Harel, 1987). In our project, this lets us handle certain actions and events happening during the experiment such as the image capture in a deterministic and reproducible way, which is important to get consistent experimental results. We use a statecharts-based dialogue manager implemented with XState[2] library for TypeScript. We present the statechart for dialogue manager and WoZ-based dialogue control in Appendix C.

The interaction with a Furhat robot is controlled through the Remote API[3] to activate the robot and to send the behaviours from our experimental software to the robot. We use two Furhat characters (Isabel and Titania[4]). For the drawing task instructions we use a female Greek voice and for the rest of the interaction a female English voice.[5] We also script a head and eye gesture, in which when Furhat "looks down" the head tilts down and the gaze lowers, giving a clear impression that the robot is looking at the participant's drawing. The goal of these multimodal cues is to help the participant understand that the robot's attention is directed toward the sketch, even though the movements themselves are rather mechanical.

3.2 Drawing interface

The drawing interface constitutes the primary modality for user interaction with the system and supports colour selection, erasing, and clearing the canvas. Drawing strokes are rendered in real time using the HTML5 Canvas 2D API, which processes user input such as mouse movements and translates them into visible lines on the canvas. After N seconds of user inactivity, current canvas snapshot is encoded as a base64 image[6] and transmitted to the VLM for labelling. In our experiment we set N to 5 seconds based on our estimate and a series of trials.

3.3 Vision–Language Model

The evaluation of user drawings plays a central role in assessing how well participants understand and

[2] https://stately.ai/docs/xstate

[3] https://docs.furhat.io/

[4] For clarity of name, we refer to name Titania (rather than the original name Titan) to avoid confusion, and both characters were designed as female.

[5] Specifically we used the following Azure voices: "el-GR-AthinaNeural" for Greek and "en-US-AvaNeural" for English.

[6] Base64 encoding is a binary-to-text encoding scheme that converts binary data into an ASCII string, allowing it to carry data stored in a binary format across text-based channels.

visually express spatial relationships. Before conducting the user study, we experimented with keyword matching over VLM-generated image descriptions. For example, if the drawing task was "Draw an apple on a box" the system would search the model's description for the presence of the words "apple", "box", and the spatial preposition "on". If all keywords were found, the drawing was considered correct. This approach was tested only during development and was excluded from both the pilot study and the main experiment.

While this method was simple and computationally efficient, it quickly proved to be insufficiently robust. The core issue was the variability in language: large language models often use synonyms, paraphrasing, or alternative spatial expressions. For instance, a phrase like "a fruit over a square" could describe the intended drawing, but would fail the keyword-match due to lexical mismatch.

To address these issues, we adopted a more structured and interpretable correctness prediction approach. Each drawing task is decomposed into the following elements:

- Object 1 (Obj1): e.g., "apple"
- Object 2 (Obj2): e.g., "box"
- Spatial Relation (R): e.g., "on"
- Full Sentence Match (S): an assessment of whether the entire description semantically matches the target sentence, e.g., "an apple on the box"

For each drawing, the model assigns a boolean value (true/false) to each of the four components. This representation allows us interpret the model's decisions consistently and to analyse failure in terms of missing objects, incorrect spatial relations, or a mismatch with the instruction.

Despite moving to a correctness prediction prompt, initial experiments in the development process showed that the model's predictions were not always reliable. To manage these inconsistencies and ensure experimental control and better user experience, a complementary manual evaluation procedure was introduced, described in section 3.4

We selected a VLM from the text-generation-with-multimodal-input family and deployed LLaVA-34B via Ollama[7] on a local server in order to preserve the privacy of the participant data. LLaVA-34B integrates image understanding with text generation, allowing it to evaluate drawings by analysing object presence and spatial relations. In our setup, the social robot functions as a language-learning assistant, while the VLM serves as a backend evaluator. We assign this role through a few-shot system prompt (Appendix, Figure 11), providing textual examples of correct and incorrect drawings.

3.4 WoZ real-time manual evaluation

The study used a Wizard-of-Oz setup in which, although the system included a VLM, real-time evaluation and feedback were controlled by an experimenter to ensure reliable task progression and interaction quality. As part of this setup, an experimenter in an adjacent lab monitored participants' drawings through a one-way window and a screen sharing, and issued a binary judgement (correct / incorrect), independently of the VLM's output. These judgments controlled task progression and the robot's feedback, while VLM outputs were recorded for offline analysis only and never used to alter the live interaction. The manual label therefore served as the authoritative ground truth during the study, consisting of a single true/false decision recorded at the moment of the experimenter's input and not providing separate annotations for individual objects or spatial relations.

4 Experiment

The experimental procedure is shown in Figure 2.

Before the session begins the participant receives an instruction sheet. The session then starts with the user's mouse click, which immediately triggers the main experiment loop. Once the experiment is complete, the participant fills out the post-session questionnaires. We recruited 16 participants (10 male, 6 female; 23–33 years) and used a within-subject design. Participants interacted alone with the social robot, while the experimenter monitored remotely and two cameras recorded the interaction from different angles. Each participant completed three drawing tasks in each condition in randomized order, comparing scripted binary feedback with a VLM-generated context-sensitive feedback. Face and feedback mapping with the two robot characters Isabel and Titania was manipulated between subjects to counterbalance character effects, with some participants seeing Isabel paired with VLM feedback and Titania paired with

[7] https://ollama.com/library/llava:34b

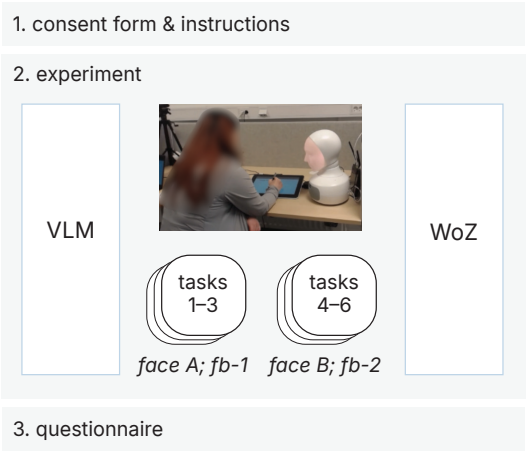


Figure 2: The procedure of the experiment. The participant is assigned with one of the four combinations of robot faces and types of the feedback (*fb*).

scripted feedback and others seeing the reverse pairing. Before the session each participant received a handout containing the two object labels apple and box, images, phonetic transcriptions in the International Phonetic Alphabet and the set of Greek and English spatial prepositions with pronunciation guidance. The handout also included simple 2D schematic illustrations showing different spatial configurations to support comprehension (see Appendix A). In each session, participants drew the verbally prompted spatial relation. After the experimenter evaluated their drawing, they received feedback, either detailed VLM feedback describing object positions and relations or simple scripted feedback indicating correct or incorrect, before proceeding to the next task. Sessions were conducted individually, with task order randomized to ensure balanced exposure across conditions.

Below are examples of the feedback used:

- **VLM (correct, generated)**, e.g., "It looks like the apple is placed inside the box — great job following the instructions!"
- **VLM (incorrect, generated)**, e.g., "The apple is not in front of the box; it's positioned on top or beside it."
- **Scripted feedback:** "Bravo, the drawing is correct." / "Unfortunately, you need to do something more."

Once finished with the drawing, participants completed questionnaires. Social agent user interaction was measured with the Artificial Social Agent Questionnaire (Fitrianie et al., 2022), a seven point scale assessing believability, usability, likeability, sociability and related agent attributes. Six additional task specific items evaluated task comprehension, understanding of the Greek spatial terms, perceived difficulty, clarity of instructions, usefulness of the feedback and satisfaction with the drawing. Each participant completed one questionnaire for each robot face in order to examine whether evaluations differed between Isabel and Titania (Appendix D). Participants were unaware that the type of feedback was different for each character, therefore differences between VLM feedback and scripted feedback were examined indirectly. Completing both questionnaires took around ten minutes.

5 Results and discussion

5.1 Social robot interaction

We first examine the results from the completed questionnaires. A question-by-question analysis considered three factors: **character** (Titania vs. Isabel), **feedback** (VLM vs. scripted) and **congruence** (congruous vs. incongruous). Conditions were treated as congruous when Titania's appearance matched scripted feedback and Isabel's appearance matched VLM feedback; the remaining pairings were treated as incongruous.

Per-question comparisons were conducted using paired-sample t-tests, as each participant rated both conditions across all three factors. According to Figure 3, participants consistently rated Isabel higher than Titania. Specifically, Q1 ("The agent's appearance makes me think of a human") showed a significant difference ($t = 4.189$, $p = 0.0008$, $d = 1.047$), as did Q3 ("The agent seems natural from its outward appearance") ($t = 2.334$, $p = 0.0339$, $d = 0.584$). Q20 ("The agent has a social presence") revealed a positive trend favouring Isabel ($t = 1.939$, $p = 0.0716$, $d = 0.485$). Across most items, Isabel received higher ratings, indicating that participants were particularly sensitive to the robot's visual appearance.

As shown by Figure 4, VLM-generated feedback produced medium but non-significant effects (Q21 and Q26). On Q21 ("The agent's and my behaviours are in direct response to each other's behaviour") ($t = 2.029$, $p = 0.0605$, $d = 0.507$), and on Q26 ("The agent's feedback helped me improve or adjust my drawing") ($t = 2.053$, $p = 0.058$, $d = 0.513$), both favouring VLM feedback. These results suggest that participants perceived

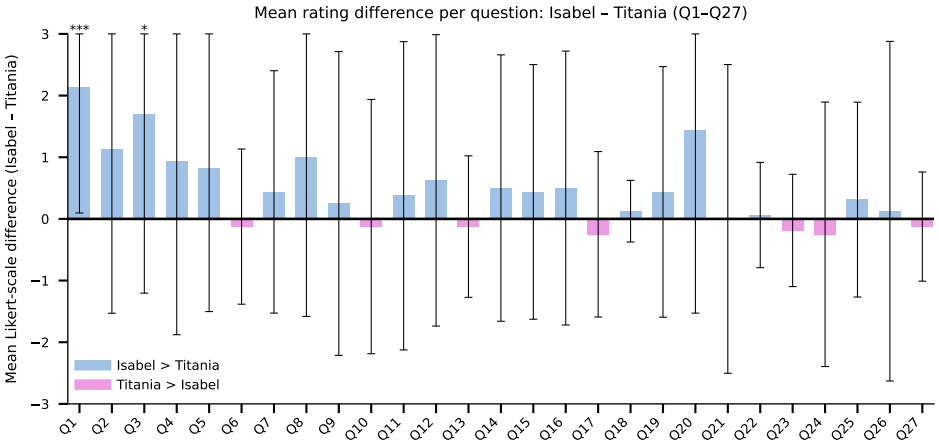


Figure 3: Isabel vs. Titania: mean rating difference per-question (Q1–Q27). Statistical significance: *** $p < .001$, * $p < .05$. Error bars indicate standard deviation.

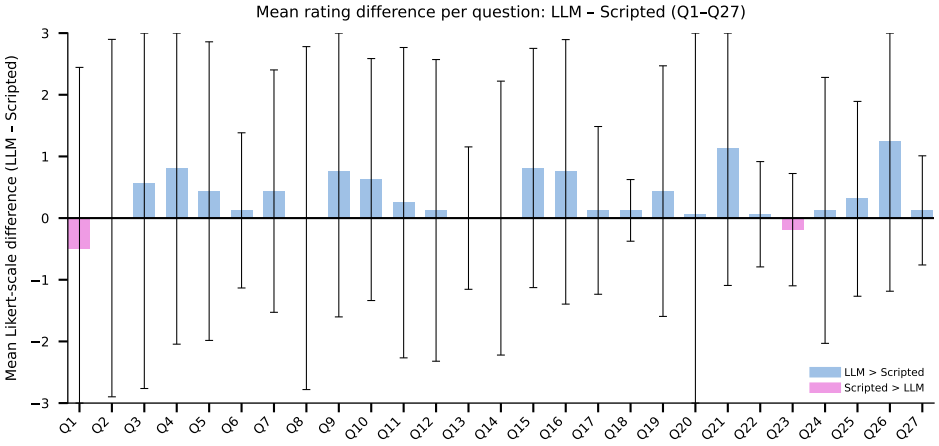


Figure 4: VLM vs. Scripted: mean rating difference per-question (Q1–Q27). Error bars indicate standard deviation.

VLM feedback as somewhat more engaging and useful, even if the differences were not statistically reliable. No significance effects were observed for congruence between character and feedback (Figure 5). The largest trend appeared on Q14 (“ I can rely on the agent”), where congruous pairings received slightly higher ratings ($t = 1.78$, $p = 0.096$, $d = 0.444$), but the effect remained small.

Overall, the robot’s appearance had the strongest impact on user perceptions, while feedback style and character-feedback congruence had smaller, non-significant effects. Effect sizes for appearance-related questions confirm that participants were most sensitive to human-likeness and naturalness, whereas medium effects for feedback suggest that VLM feedback may provide subtle engagement benefits. Free-form participant comments support this interpretation, as participants preferred the human-like Isabel even when paired with simpler feedback, while Titania was rated lower even despite providing more detailed responses for participants assigned with Titania-VLM pairing.

5.2 Evaluation of a VLM on sketches

We evaluate the VLM using two approaches (see Section 3.3): a *sentence-level* correctness of the drawing and a *feature-based* correctness requiring Object 1 (“apple”), Object 2 (“box”) and the spatial relation (“on”) to be correct. This allowed us to identify where differences in model predictions arise. The dataset comprised 427 images, of which 162 were labelled by the first author as correct (38%) and 265 incorrect (62%). As shown in Table 1, for sentence-level correctness, the model achieved moderate accuracy, around 0.52, with a tendency to over-predict correctness. False positives were substantially more frequent than false negatives, indicating the model often marked a

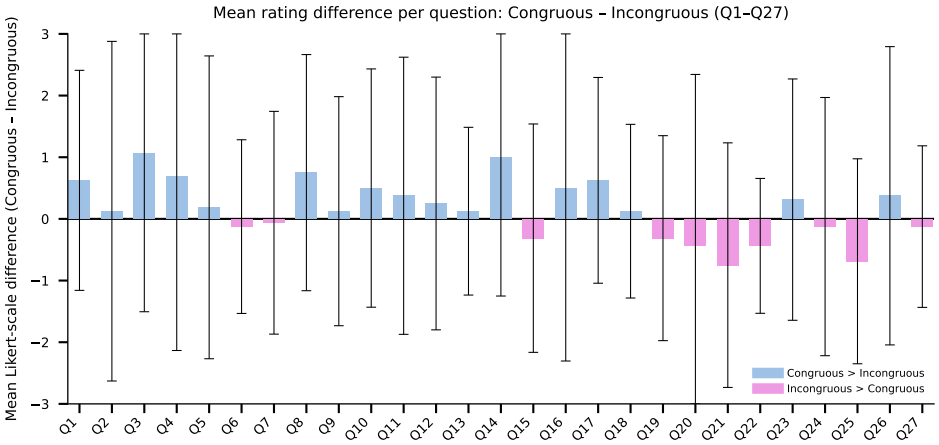


Figure 5: Per-question comparison: Congruous Vs Incongruous: mean rating difference per-question (Q1–Q27). Error bars indicate standard deviation.

	Sentences			**Features**		
Class	*Prec.*	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*
Correct	.419	.667	.514	.521	.154	.238
Incorrect	.680	.434	.530	.639	.913	.752
F1-macro		.522			.495	
F1-weighted		.524			.557	
Accuracy		.522			.625	

Table 1: Per-class metrics comparing the sentence decision and the triplet decision.

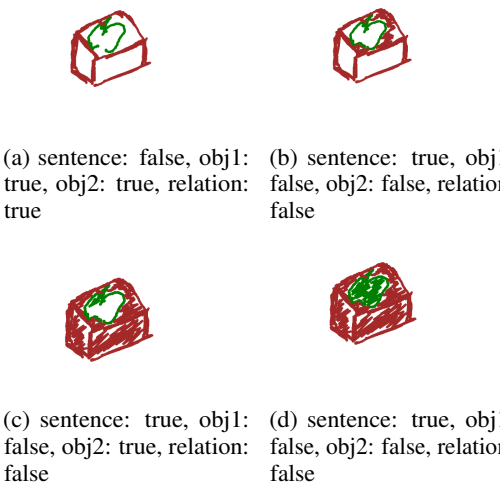
drawing as correct when it was not. Accuracy varied considerably across participants, from 0.42 to 0.74, reflecting the influence of drawing clarity, colour use, and object depiction. The stricter feature-based correctness, which required all three components (object 1, object 2, relation) to be *true*, reversed the error pattern. False negatives were now much more common than false positives, leading to higher overall accuracy, around 0.63. Recall for truly correct drawings was low, indicating that many correct drawings were missed, but incorrect drawings were detected more reliably. This pattern reflects the model's sensitivity to individual drawing features and the stricter evaluation criteria. Figure 16 shows that the sentence-based evaluation labels a much higher number of drawings as correct compared to the feature-based evaluation method. The stricter feature-based criterion comes at the cost of many false negatives, as a number of truly correct drawings are missed. However, Figure 17 shows accuracy across the sequence of drawing snapshots. Overall, feature-based accuracy remains higher for most snapshots, while the sentence-based decision is less stable across the sequence. This stability across the drawing sequence is mainly due to the feature-based method's more reliable identification of incorrect drawings.

Analysis of examples Drawing examples illustrate these patterns. Sentence-level evaluation marked drawings as correct despite incorrect spatial relations, while feature-based evaluation correctly flagged such errors. Conversely, missed detections of an object or relation caused the feature-based evaluation to classify correct drawings and incorrect. Predictions were also unstable across drawing stages. For instance, a participant sketching an apple inside a box (Figure 6) might have all elements correctly detected at first, but after adding colour or small edits, the model could fail to recognise one, leading to misclassification. Such instability suggests that using the model for *real-time* feedback may produce inconsistent or misleading guidance.

Overall, the VLM reliably detects structural and relational errors but shows participant-dependent variability and sensitivity to small edits. Adding context memory so that the model retains recent task descriptions and images, following retrieval-augmented generation ideas (Lewis et al., 2020), could improve stability across incremental changes and reduce label flips, although this still needs to be validated.

6 Conclusions

In this study we propose a language learning pipeline that combines a social robot, a Vision-Language Model (VLM), and real-time human eval-



(a) sentence: false, obj1: true, obj2: true, relation: true

(b) sentence: true, obj1: false, obj2: false, relation: false

(c) sentence: true, obj1: false, obj2: true, relation: false

(d) sentence: true, obj1: false, obj2: false, relation: false

Figure 6: Changes in the model's predictions across drawing stages for the task "an apple inside the box".

uation for interactive drawing tasks involving spatial prepositions. The system integrates multimodal input and supports both offline VLM analysis and Wizard-of-Oz human-controlled feedback. This pipeline is the first step towards more interactive and natural AI-assisted language learning. Our results demonstrate that the VLM could identify most objects and relations, achieving moderate accuracy. However, its instability across drawing stages and sensitivity to edits limit its reliability for autonomous feedback. Human evaluation remains essential for accurate guidance. We also observe that humans are more sensitive to the human-like robot (Isabel) than the robotic one (Titania) in our experiment, while feedback type showed only non-significant trends favouring VLM feedback.

Future work will expand participant numbers, conduct a qualitative analysis of VLM recognition performance, and test alternative VLMs for comparison with different few-shot prompts. A long-term goal is to fine-tune models for sketch-based tasks, once a larger dataset is available.

Limitations

The main limitation of this study is the small sample size (N=16). While sufficient to reveal clear trends, it limits statistical power and the ability to generalize. A larger sample might reveal small but meaningful effects, particularly in user perception measures.

A second limitation concerns the real-time human evaluation of drawings. The human judgment was binary (correct/incorrect) and applied to the drawing as a whole, without distinguishing individual objects or spatial relations. Consequently, partial drawings (e.g., only the apple or only the box) were labelled as incorrect instead of incomplete. This also prevents direct feature-level comparison between human and system evaluations.

Finally, due to the logging setup, some images lacked in-session labels and were later annotated manually by the experimenter. These post-session labels ensured dataset completeness but were produced by a single rater, meaning inter-rater reliability cannot be established.

Ethical Considerations

Participation was voluntary and preceded by informed consent explaining the purpose of the study, recording procedures, and the right to withdraw. Only data necessary for analysis were saved (drawings, task descriptions, timestamps, model outputs, and human labels). Questionnaires were anonymous, and session IDs were used instead of names. The VLM ran on the university server, receiving only sketches and session IDs. No video, audio, or personal data were transmitted or stored. All recordings were kept locally on the experimenter's computer with restricted access, and the final dataset contains only de-identified images.

Participants were not exposed to model errors during interaction. The VLM operated offline, and feedback was based on the experimenter's human label, minimizing the risk of misleading or inconsistent model responses. Overall, the study ensured privacy, transparency, and full de-identification throughout the process.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and the participants who took part in the study for their time and contributions. Nikolai Ilinykh was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Vladislav Maraev was supported by a Swedish Research Council Grant – VR project 2023-00358 – Social laughter for virtual agents (SocLaVA).

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra,

and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, february 21-26, 2011, revised selected papers*, pages 114–130. Springer.

Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics*, 3(21):eaat5954.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures.

Anastasiia Fadeeva, Philippe Schlattner, Andrii Maksai, Mark Collier, Efi Kokiopoulou, Jesse Berent, and Claudiu Musat. 2024. Representing online handwriting for recognition in large vision-language models (2024). *arXiv preprint arXiv:2402.15307*.

Judith E Fan, Wilma A Bainbridge, Rebecca Chamberlain, and Jeffrey D Wammes. 2023. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568.

Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. 2022. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

David Harel. 1987. Statecharts: A visual formalism for complex systems. *Science of computer programming*, 8(3):231–274.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657, Bangkok, Thailand. Association for Computational Linguistics.

Sharon Oviatt. 2007. Multimodal interfaces. *The human-computer interaction handbook*, pages 439–458.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Natasha Randall. 2019. A survey of robot-assisted language learning (rall). *ACM Transactions on Human-Robot Interaction (THRI)*, 9(1):1–36.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Preprint*, arXiv:1409.0575.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Rianne Van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne Van der Ven, and Paul Leseman. 2019. Social robots for language learning: A review. *Review of Educational Research*, 89(2):259–295.

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.

A Instructions sheet

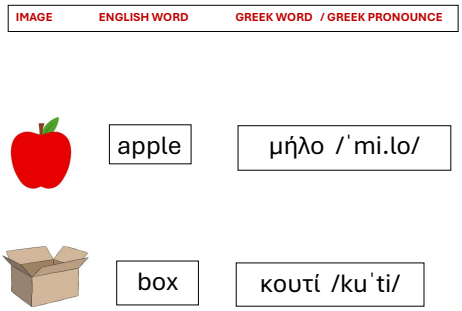


Figure 7: Instructions Sheet (page 2)

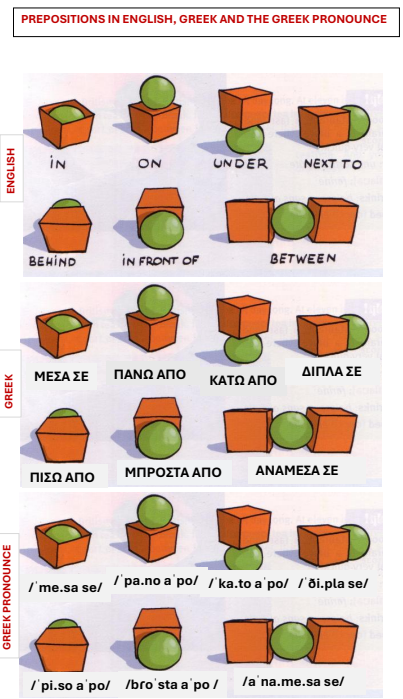


Figure 8: Instructions Sheet (page 2)

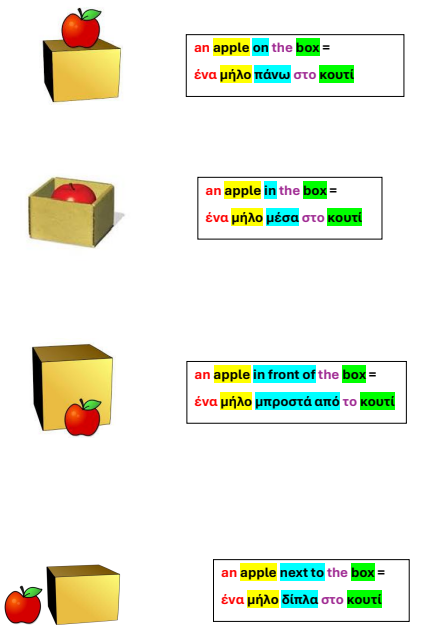


Figure 9: Instructions sheet (page 3)

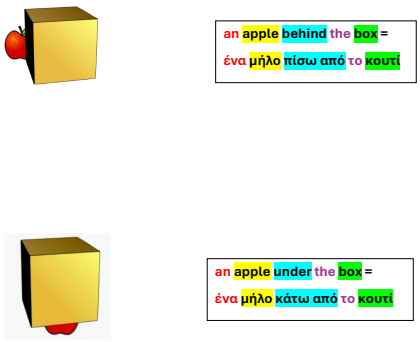


Figure 10: Instructions sheet (page 4)

B VLM Prompts

```
You are a helpful assistant that looks at drawings and decides whether they match a given description.
You will receive a short description and an image.
Your job is to answer four yes/no questions about whether the image contains the right objects and relationships.

Description: "<TASK_DESCRIPTION_ENGLISH>"
Respond ONLY with a raw JSON object and nothing else.
Do NOT include any commentary, explanation, or text after the JSON.
Do NOT use code blocks or wrap the JSON in triple backticks.

DO NOT guess. Only answer "true" if the object or relation is CLEARLY VISIBLE in the image.
If you're unsure or the object is missing, answer "false".

Answer the following questions:
1. Is the following sentence correctly describing the image: "<TASK_DESCRIPTION_ENGLISH>"?
2. Is this correct: the picture contains "<OBJ1>"?
3. Is this correct: the picture contains "<OBJ2>"?
4. Is this correct: the picture contains a relation "<RELATION>" between "<OBJ1>" and "<OBJ2>"?

Please reply with a JSON object:
{
  "sentence": true/false,
  "obj1": true/false,
  "obj2": true/false,
  "relation": true/false
}

Example 1:
Description: "An apple in a box"
Drawing: (a picture showing an apple inside a box)
Response:
{
  "sentence": true,
  "obj1": true,
  "obj2": true,
  "relation": true
}

Example 2:
Description: "An apple next to a box"
Drawing: (a picture showing an apple on top of a box)
Response:
{
  "sentence": false,
  "obj1": true,
  "obj2": true,
  "relation": false
}
```

Figure 11: Binary classification prompt used for VLM predictions (placeholders: <TASK_DESCRIPTION_ENGLISH>, <OBJ1>, <OBJ2>, <RELATION>).

```
You are a helpful assistant who provides natural,
concise feedback on visual tasks based on short drawing instructions.
For each drawing task, you have seen the image.
Respond as if speaking naturally to the person who made the drawing
either encouraging them when it is correct or pointing out what went wrong if it is incorrect.

Example 1:
Instruction: Draw an apple in front of the box.
Result: The drawing is correct.
Feedback: The apple is clearly in front of the box, positioned lower and closer to the viewer — well done.

Example 2:
Instruction: Draw an apple under the box.
Result: The drawing is incorrect.
Feedback: The apple is placed next to the box instead of underneath it.

Now evaluate the following:

Instruction: <TASK_DESCRIPTION>
Result: The drawing is <CORRECTNESS>.
Feedback:
```

Figure 12: Few-shot prompt for VLM generated feedback (placeholders: <TASK_DESCRIPTION> and <CORRECTNESS>).

C Dialogue Statechart

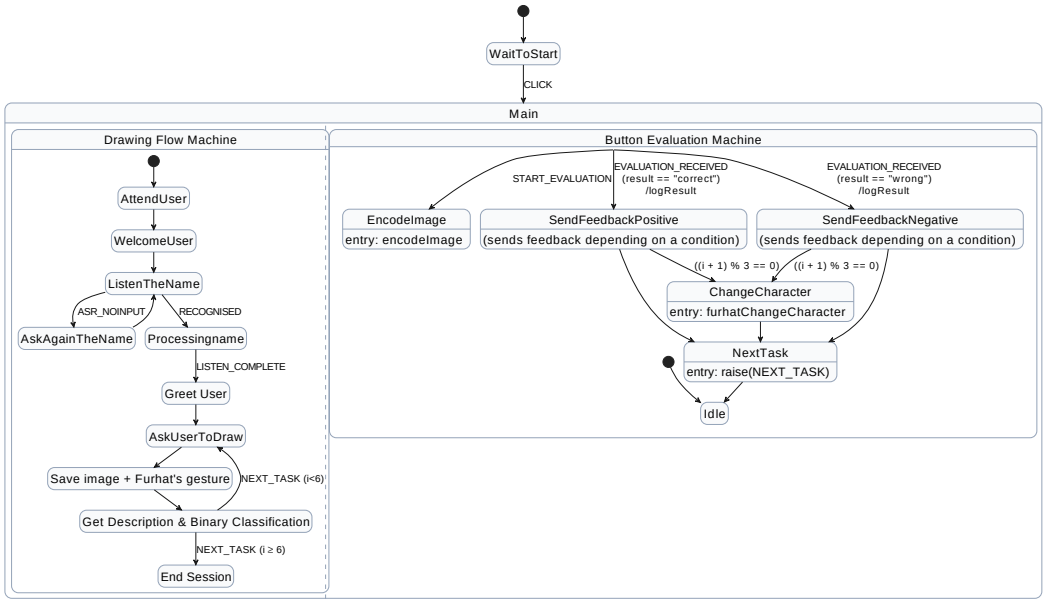


Figure 13: Dialogue manager and WoZ-based dialogue control

D Questionnaire Items

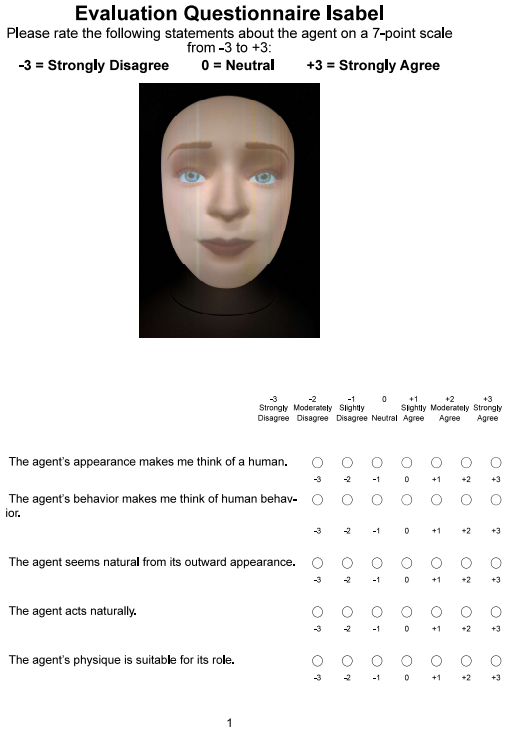


Figure 14: Questionnaire for Isabel (page 1)

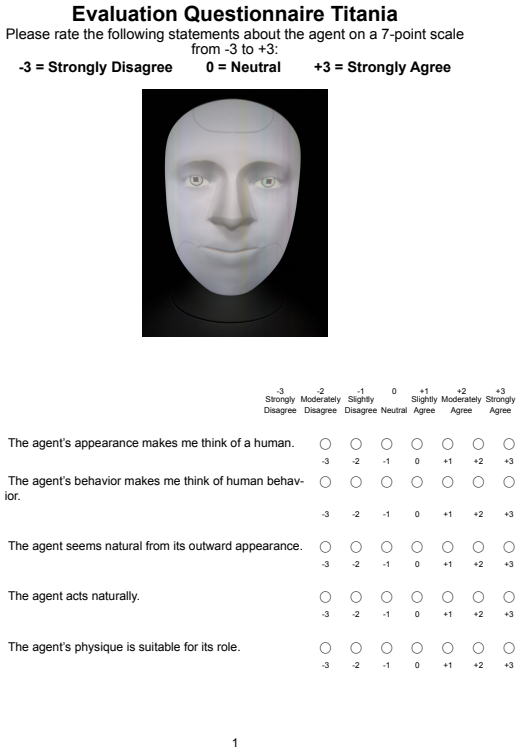


Figure 15: Questionnaire for Titania (page 1)

No.	Questionnaire statement
Q1	The agent's appearance makes me think of a human.
Q2	The agent's behavior makes me think of human behavior.
Q3	The agent seems natural from its outward appearance.
Q4	The agent acts naturally.
Q5	The agent's physique is suitable for its role.
Q6	Learning how to communicate with the agent is quick.
Q7	The agent does its task well.
Q8	I like the agent.
Q9	The agent interacts socially with me.
Q10	The agent has a distinctive character.
Q11	I can see myself using the agent in the future.
Q12	I enjoy interacting with the agent.
Q13	The interaction captured my attention.
Q14	I can rely on the agent.
Q15	The agent can collaborate in a productive way.
Q16	The agent is attentive.
Q17	The agent's behavior does not make sense.
Q18	The agent acts intentionally.
Q19	I see the interaction with the agent as something positive.
Q20	The agent has a social presence.
Q21	The agent's and my behaviors are in direct response to each other's behavior.
Q22	I understood what the agent was asking me to draw.
Q23	I understood the meaning of the Greek spatial terms
Q24	The drawing task was difficult.
Q25	The agent's instructions led to a drawing that made sense.
Q26	The agent's feedback helped me improve or adjust my drawing.
Q27	I am satisfied with my final drawing.

E VLM graphs

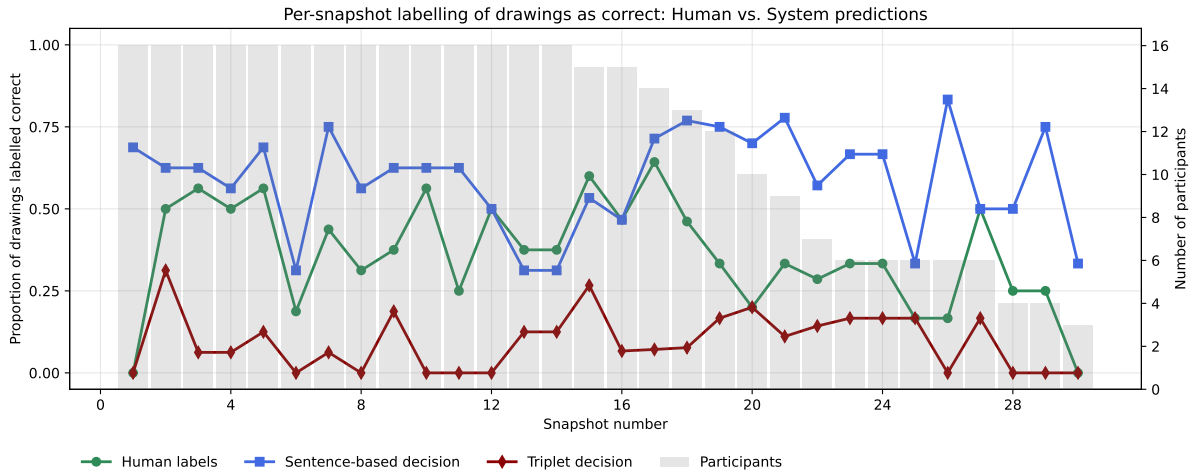


Figure 16: Per-snapshot labelling of drawings as correct: Human vs. System predictions. Only the first 30 snapshots are shown. Two participants exceeded 30 and their later snapshots are omitted for readability.

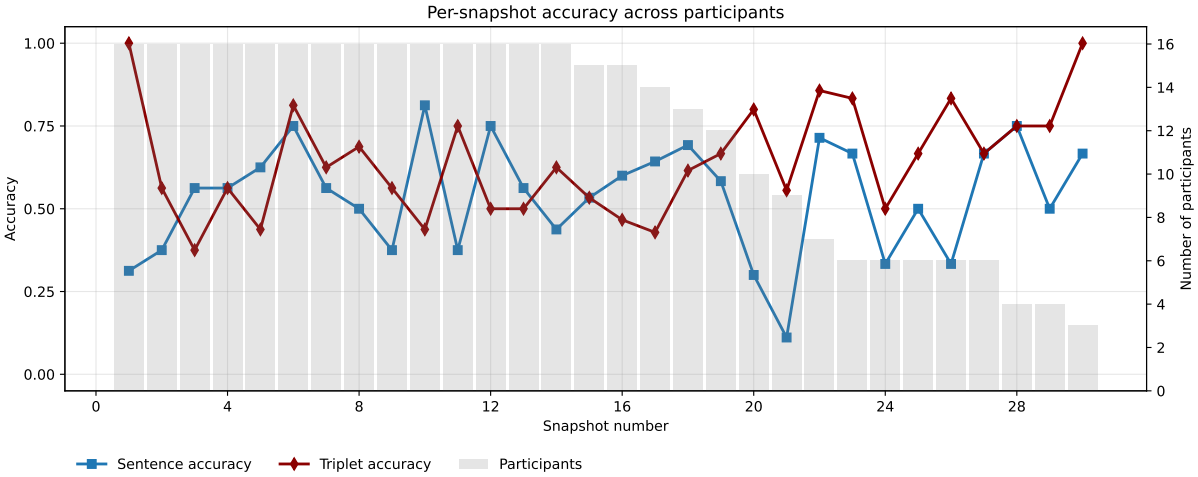


Figure 17: Accuracy per-snapshot across participants. Only the first 30 snapshots are shown. Two participants exceeded 30 and their later snapshots are omitted for readability.