# Hybrid Procedural Semantics for Visual Dialogue: An Interactive Web Demonstration

Lara Verheyen[1], Jérôme Botoko Ekila[1], Jens Nevens[1], Paul Van Eecke[1,2] and Katrien Beuls[3]

[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[2]Itec, imec research group at KU Leuven, E. Sabbelaan 51, 8500 Kortrijk, Belgium
[3]Faculty of Informatics, University of Namur, Rue Grandgagnage 21, 5000 Namur, Belgium

**Abstract**

Visual dialogue refers to the task in which a conversational agent needs to hold a meaningful and coherent conversation with a human interlocutor about a scene they observe. To tackle this task, we introduce a novel methodology that makes use of (i) a novel data structure, called the conversation memory, which holds information that is incrementally conveyed in the conversation and (ii) a hybrid procedural semantic representation that is grounded in both the visual input and the conversation memory. In this paper, we present a demonstration that showcases this novel methodology. In this demonstration, a user can interact with a visual dialogue agent and discuss an image of their choice. While the agent is answering questions, the user can follow the agent's reasoning process. Due to its explainable and interpretable nature, the novel methodology can be used in a wide range of application domains, especially when it is important that the system is human-interpretable. We believe that this novel methodology of hybrid procedural semantics combined with a conversation memory paves the way for building truly intelligent and explainable systems that are able to hold human-like conversations.

## 1. Introduction and background

The task of visual dialogue as introduced by [1] requires an agent to correctly answer a series of questions about visual input. However, these questions are not independent from each other. In many cases, answering these questions involves resolving coreferences with respect to earlier dialogue turns. Compared to the task of visual question answering, the interdependent questions are an extra complexity inherent to the task of visual dialogue. Here, the answers do not only have to be grounded in the visual context, but also in the conversational context.

The CLEVR-Dialog dataset [2] was especially designed to be a diagnostic benchmark for the visual dialogue task. The dataset consists in dialogues discussing images from the CLEVR dataset [3]. In the course of a dialogue, an agent initially receives an image and a caption describing some parts of the contents of the image. Then, ten questions are respectively asked and answered. The goal for the agent is to answer each question correctly. An example dialogue about the image in the upper right corner of Figure 1 would be:

C: There is a green object in the middle.
Q: What is its shape? A: Sphere
Q: And its size? A: Small
Q: Is there an object to its left? A: Yes
Q: How many other objects are in the image?
A: 4

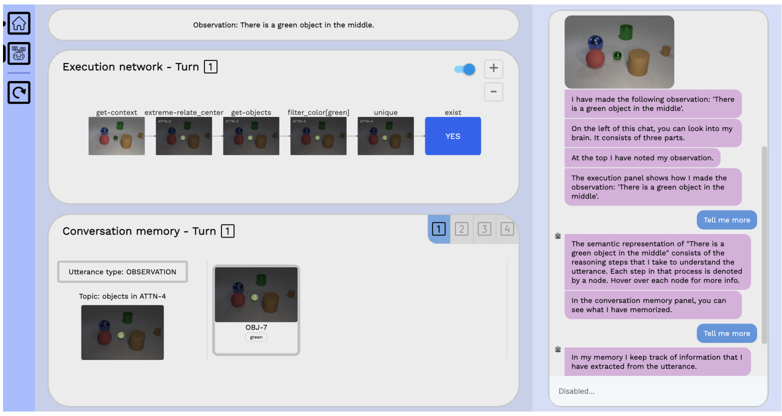
A visual dialogue agent must thus be capable on the one hand of solving coreferences in the conversation, and on the other hand of grounding references in the image. Traditionally, the task of visual dialogue is tackled with neural network approaches. For example, [1] introduced an encoder-decoder architecture with encoders that have late-fusion, hierarchical encoding and memory networks. Other approaches take a more explicit approach and use mechanisms that explicitly represent the dialogue history. For example, [4] and [5] make use of an associative memory to represent the previous questions with their corresponding attentions. In combination with this associative memory, [5] use a neural module networks architecture [6].

Taking inspiration from this last approach, we introduce a novel methodology, based on two concepts: a conversation memory and a hybrid procedural semantics. These novel techniques allow a visual dialogue agent to ground the questions in

*IJCAI 2022: Workshop on semantic techniques for narrative-based understanding, July 24, 2022, Vienna, Austria*

lara.verheyen@ai.vub.ac.be (L. Verheyen); jerome@ai.vub.ac.be (J. B. Ekila); jens@ai.vub.ac.be (J. Nevens); paul@ai.vub.ac.be (P. V. Eecke); katrien.beuls@unamur.be (K. Beuls)

CEUR Workshop Proceedings (CEUR-WS.org)



Observation: There is a green object in the middle.

Execution network - Turn 1

Conversation memory - Turn 1

Utterance type: OBSERVATION

Topic: objects in ATTN-4

OBJ-7

green

I have made the following observation: 'There is a green object in the middle'.

On the left of this chat, you can look into my brain. It consists of three parts.

At the top I have noted my observation.

The execution panel shows how I made the observation: 'There is a green object in the middle'.

Tell me more

The semantic representation of "There is a green object in the middle" consists of the reasoning steps that I take to understand the utterance. Each step in that process is denoted by a node. Hover over each node for more info.

In the conversation memory panel, you can see what I have memorized.

Tell me more

In my memory I keep track of information that I have extracted from the utterance.

Disabled...

**Figure 1:** Main interface of the demonstration, with a chat window on the right and the agent's reasoning process on the left.

both the conversational context and the visual context. To represent the history of a dialogue, the conversation memory keeps track of relevant information conveyed in the dialogue. To understand a question correctly, the agent maps the question onto a meaning representation in terms of procedural semantics [7, 8, 9]. This meaning representation consists of the conceptual operations that an agent needs to execute in order to answer a question. Procedural semantics has been used successfully for the task of visual question answering [10, 11, 12]. We have extended the procedural semantics that was designed for the task of visual question answering [12] to include operations that cover the incremental nature of dialogues. Moreover, the meaning representation is executed in a hybrid way, where some steps are executed symbolically (in particular operations related to reasoning processes) and others are executed subsymbolically (in particular operations responsible for perception). The hybrid method contrasts with the approach of [5], who use a neural module networks approach with queries that are executed subsymbolically.

In the demonstration that is introduced in this paper, our novel methodology is explained didactically in a step-wise fashion. The interactive web demonstration can be found at https://ehai.ai.vub.ac.be/demos/visual-dialog/[1]. The main interface of the demo is shown in Figure 1. Through this interface, the user can discuss an image with a visual dialogue agent. The user chooses an image and selects a question. While the agent computes an answer to the question, the reasoning steps that the agent is performing are shown. The goal of the demonstration is for the user on the one hand to reflect on what it takes to hold a meaningful and coherent conversation, and on the other hand to gain a deep understanding of our innovative methodology based on hybrid procedural semantics.

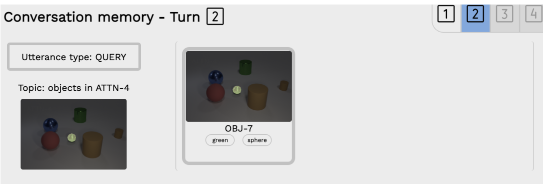
## 2. Methodology

The novel methodology that we designed for solving visual dialogue tasks builds on two foundational ideas: the use of a *conversation memory* and a *hybrid procedural semantics* that is grounded in both visual input and the conversation memory.

The conversation memory is a data structure that keeps track of all relevant knowledge that is built up during the dialogue. It is composed of a number of turns, which represent the turns in the dialogue (i.e., the observation and the question-answer pairs). After each turn of the dialogue, information is added to the conversation memory. Each turn in the conversation memory consists of its timestep, the utterance type of the turn, the current topic of the conversation, and a symbolic representation of the mentioned attributes of the objects that were discussed.

The procedural semantics consists of operations that can interact with the conversation memory, when coreferences between turns (e.g., '*it*') need to be resolved. For example, the primitive operation GET-LAST-TOPIC returns the topic of the previous turn, which was stored in the conversation memory. Figure 2 shows the conversation memory after the second turn. The question corresponding to this turn was '*What is its shape?*', the answer that the agent computed '*sphere*'. The utterance type of

[1] A video accompanying the demonstration can be found at: https://youtu.be/D3Ny6kta5d8



Utterance type: QUERY

Topic: objects in ATTN-4

Figure 2: The conversation memory part of the demonstration which shows the conversation memory after two turns.

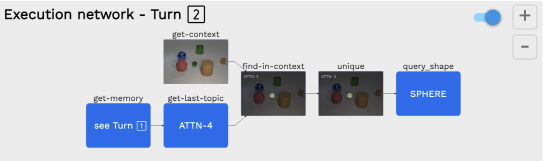


Figure 3: The execution network part of the demonstration which shows the execution of the hybrid procedural semantic representation of the question '*what is its shape?*'.

the question is '*query*'. '*Attention-4*' is the visualisation of the topic of the turn. The attributes that are conveyed in the dialogue are symbolically added to the memory. In this case, the attribute '*green*' was known from the previous turn and the attribute '*sphere*' was added in the second turn. The demonstration gives the user the option to view the previous turns in the dialogue, so that it becomes clear how the information is built up over the dialogue turns.

The hybrid procedural semantics represents the meaning of utterances. The meaning representation is expressed in terms of the conceptual operations that need to be performed to obtain an answer to a question. This makes the meaning representation directly executable. Each operation in the meaning representation is performed either symbolically or subsymbolically. The symbolic operations are operations that typically represent reasoning processes, such as comparisons or operations on the conversation memory (e.g., GET-LAST-TOPIC). Subsymbolic operations are linked to perception and are executed directly on the image (e.g., FIND-CUBES). In most cases, these subsymbolic primitives are implemented by neural networks, in other cases matrix operations on attentions are used (e.g., binary and, or, ...).

An example of the meaning representation of the question '*what is its shape?*' is shown in Figure 3. It consists of the steps GET-MEMORY, GET-LAST-TOPIC, GET-CONTEXT, FIND-IN-CONTEXT, UNIQUE and QUERY-SHAPE. The output of an operation becomes the input to another operation. For example, the output of the operation GET-MEMORY becomes the input to the operation GET-LAST-TOPIC. The operation FIND-IN-CONTEXT gets as input the output of both GET-LAST-TOPIC and GET-CONTEXT. The first operation GET-MEMORY is a symbolic operation that returns the conversation memory. This conversation memory is the input to the next symbolic operation GET-LAST-TOPIC, which returns the symbolic representation of the topic of the previous turn. The operation FIND-IN-CONTEXT retrieves this topic in the image, which is itself the output of GET-CONTEXT. Then, the symbolic operation UNIQUE checks whether the input attention contains just one object. Lastly, the subsymbolic operation QUERY-SHAPE will use a classifier to classify the input attention into a shape category. The output of this last step is the answer to the question. Thus, the answer to the question '*what is its shape?*' is in this case '*sphere*'.

## 3. Visual dialogue demo

Figure 1 shows the main interface of the demonstration. The interface consists of two main parts. On the right, there is a chat window in which the user can interact with the visual dialogue agent, by choosing images and asking questions. On the left, the user can follow the reasoning process of the agent. This part is itself divided in an *utterance*, *execution network* and *conversation memory*. *Utterance* shows the utterance under consideration, which can be an observation (i.e., a statement about the image) or a question. The utterance is then mapped onto a procedural semantic representation, which is shown under *execution network*. The meaning representation can be viewed before execution and after execution. Figure 3 shows the execution process of the meaning representation. The last window shows the conversation memory, which is updated after the agent computes the answer.

The demonstration proceeds as follows. First, the user chooses an image. Then, the visual dialogue agent processes the image and utters a statement about the image. The underlying meaning representation of the statement and the hybrid execution network are shown. Meanwhile in the chat, the agent provides more information about the methodology. Next, the agent updates its conversation memory with information from the observation. Afterwards, a few questions appear for the user to choose from. These questions are based on the questions from the CLEVR-Dialog dataset [2]. The visual dialogue agent maps the question to a meaning representation and executes it in a hybrid way. The execution of the meaning representation immediately provides the answer to the question. Then, the agent updates its conversation memory so that it contains all information from this turn. Afterwards, the user can select a follow-up question to continue the dialogue. After a number of turns, the user is given the option to select another image and start a new dialogue. Again, the user can see the same mechanisms at work. This gives them insight into the system and showcases the explainability and interpretability of the methodology.

## 4. Contribution

The goal of this didactic demonstration is twofold. On the one hand, it aims to let users experience the challenges involved in understanding grounded natural language conversations. On the other hand, it aims to showcase our novel hybrid procedural semantics-based methodology for solving visual dialogue tasks. The demonstration especially focusses on showcasing the explainability and interpretability of the reasoning processes performed by the agent, which is one of the major advantages of our approach as compared to other state-of-the-art approaches. This makes our approach particularly interesting to human-centric AI applications, in which explainability and interpretability are a major concern. Possible applications include safety-critical systems such as emergency response platforms and decision support systems that are required to motivate decisions in a human-understandable fashion.

## 5. Conclusion

The demonstration that is proposed in this paper shows the dynamics of two novel techniques for the task of visual dialogue: (i) a conversation memory that is incrementally updated and (ii) a procedural semantics that is executed in a hybrid way. The conversation memory is a representation of all previous turns of the dialogue and is used to solve coreferences with respect to previous turns. The procedural semantics is a meaning representation for utterances in terms of steps that need to be performed. The meaning representation is executed in a hybrid manner with symbolic primitives to execute reasoning operations and subsymbolic primitives that are responsible for operations related to perception. In the interactive web demonstration, the user can ask questions to the agent. While the agent computes the answer, the novel methodology is explained. The explainability and interpretability by design are the main factors that make this novel methodology an ideal set-up for a new generation of intelligent agents that can hold a meaningful and coherent conversation with a human.

## References

[1] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, D. Batra, Visual Dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1080–1089.

[2] S. Kottur, J. M. Moura, D. Parikh, D. Batra, M. Rohrbach, CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 582–595.

[3] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2901–2910.

[4] P. H. Seo, A. Lehrmann, B. Han, L. Sigal, Visual reference resolution using attention memory for visual dialog, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 3722–3732.

[5] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: Proceedings of the 15th European Conference on Computer Vision (ECCV), 2018, pp. 153–169.

[6] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Neural module networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 39–48.

[7] W. A. Woods, Procedural semantics for a question-answering machine, in: Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I, New York, NY, USA, 1968, pp. 457–471. URL: https://doi.org/10.1145/1476589.1476653.

[8] T. Winograd, Understanding natural language, Cognitive Psychology 3 (1972) 1–191.

[9] P. N. Johnson-Laird, Procedural semantics, Cognition 5 (1977) 189–214.

[10] J. Andreas, M. Rohrbach, T. Darrell, D. Klein, Learning to compose neural networks for question answering, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1545–1554.

[11] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Inferring and executing programs for visual reasoning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2989–2998.

[12] J. Nevens, P. Van Eecke, K. Beuls, Computational construction grammar for visual question answering, Linguistics Vanguard 5 (2019) 20180070.