Integrating Visual Context into Language Models for Situated Social Conversation Starters

Ruben Janssens, *Student Member, IEEE,* Pieter Wolfert, *Student Member, IEEE,*
Thomas Demeester, Tony Belpaeme, *Member, IEEE*

Abstract—Embodied conversational agents that interact socially with people in the physical world require multi-modal capabilities, such as appropriately responding to visual features of users. While existing vision-and-language models can generate language based on visual input, this language is not situated in a social interaction in the physical world. We present a novel task called Visual Conversation Starters, where an agent generates a conversation-starting question referring to features visible in an image of the user. We collect a dataset of 4,000 images of people with 12,000 crowdsourced conversation starters, and compare various model architectures: fine-tuning smaller seq2seq or image-to-text models versus zero-shot prompting of GPT-3.5, using image captions versus end-to-end image input, and training on human data versus synthetic questions generated by GPT-3.5. Models were used to generate friendly conversation starters which were evaluated on criteria including language fluency, visual grounding, interestingness and politeness. Results show that GPT-3.5 generates more interesting and polite questions than smaller models that are fine-tuned on crowdsourced data, but vision-to-language models are better at referencing visual features, and they can mimick GPT-3.5's performance. This demonstrates the feasibility of deep visiolinguistic models for situated social agents, forming an important first stage in creating situated multimodal social interaction.

Index Terms—Natural language generation, vision-and-language, conversation models, embodied conversational agents.

1 Introduction

At the start of each successful conversation is a proper conversation starter. Conversation starters are often polite questions or remarks, which are not necessarily communicatively informative, but are instead a way of inviting the other to an interaction [1]. Most human relationships are formed through face-to-face interactions, and conversation is central to the construction of a relationship between people. Often, this relationship-forming conversation lacks a task-oriented aspect and instead is a polite exchange of social talk, also known as phatic communication [2] or small talk. One way of doing this is, for example, to comment on something perceptible, like asking why a conversation partner is wearing their sunglasses on a rainy day. Evidence shows that when we approach a stranger to talk, we frequently rely on publicly displayed cues as a source of material for initial conversations [3].

Laver [4] described that phatic communication aims to establish a connection, or "solidarity", between conversation partners. He noted that people will often refer to factors specific to the time and place of the interaction, being phrased as questions that refer to the conversation partner, such as "How do you like the sunshine, then?", employing what he calls "other-oriented tokens".

Corresponding author is Ruben Janssens (ruben.janssens@ugent.be). Ruben Janssens and Tony Belpaeme are with IDLab-AIRO at Ghent University - imec. Thomas Demeester is with IDLab-T2K at the same institution. Pieter Wolfert is at Radboud University but primarily contributed to this work while he was with IDLab-AIRO.
This research received funding from the Flemish Government (AI Research Program), the Horizon Europe VALAWAI project (grant nr. 101070930), and the European ROBotics and AI Network (euROBIN, grant nr. 101070596).
Manuscript received November 25, 2023; revised July 5, 2024.

Phatic communication has also been shown to be employed in many situations, such as nurses aiming to close the emotional distance with their patients through observations like "That's a really lovely blouse you're wearing. That colour suits you" [5], advice being taken more after phatic communication is used [6], and even second language learning being aided by also practicing phatic communication in the new language [7] – all of which are domains where artificial agents such as social robots are often employed [8], [9].

Small talk has been suggested as being equally important to establish a relationship or bond with artificial agents [10], [11]. When the agent personalises its interaction, by referring to an individual's features or preferences, the perception of the robot improves and secondary outcomes, such as learning with a robot, increase [12], [13]. This requires the artificial agent to have visual perception, but also the skills to place it in a temporal and linguistic context.

Yet, when we try to have a conversation with an embodied artificial agent, it is unlikely that it can meet the criteria we have for human interlocutors. Getting artificial systems, such as social robots, to both understand their environment and reference it in a conversation, is a challenging objective in the fields of Human-Robot Interaction (HRI) [14] and Natural Language Processing (NLP). Nevertheless, having artificial systems understand their surroundings, with the ability to weave that understanding into an open-domain conversation, is key to a successful human-AI interaction.

Conversational AI has recently made large progress towards fluent open-ended social conversation, but still lacks the ability to process social multi-modal inputs. We showcase how current AI models can be adapted to gain

this ability: we created a system that starts a friendly chit-chat conversation with a user, by asking a question that is based on a visual feature of the user or their environment. We call this task "*Visual Conversation Starters*": generating a visually grounded conversation-starting question based on an image of the user, which represents the visual input an embodied AI system such as a social robot would have when interacting with them.

This research fills a hole in previous work on multi-modal conversational AI, which is either not situated (e.g. using an image as a shared conversation topic instead of as information about the conversation partner and the environment), or not social (e.g. task-oriented dialogue such as spatial referring expression comprehension).

We explore different architectures to tackle this problem, shown in Figure 1, comparing systems that first transform the visual input into a textual description with those that use the visual input directly, comparing systems that are fine-tuned on a data set for this task with larger models that are not fine-tuned but are prompted (i.e. zero-shot inference), and comparing systems that are trained on human-generated data with those that are trained on AI-generated data. As part of these efforts, we have collected a data set of 4000 images of people, augmented with three conversation-starting questions written by humans, as well as three questions generated by AI.

Finally, we evaluated the different model architectures by studying how people perceive the questions they generated on five criteria: language correctness, correct visual reference, interestingness, specificness, and politeness. We evaluate the models on images from the test data and images of users interacting with a social robot during a science festival.

The data set with human-written questions was originally presented in [15], along with a pilot evaluation of a locally trained model that was trained on the data set and uses image descriptions as representations of the visual input. This paper strongly builds on the insights from this small-scale pilot study [15], with several additional (as well as more recent) models (including ChatGPT), a more extensive emprirical validation, redesigned human evaluation including on real-world input, and a range of new insights.

This research answers the following research questions, in the unique setting of situated social conversations:

1) Can language models generate accurate and specific references to visual context in a social setting, by using image captions as intermediate representation, or by using end-to-end image-to-text models?
2) How well do large language models generate polite, interesting questions that correctly refer to the visual environment without being fine-tuned, compared to smaller models that can be run locally but require fine-tuning?
3) How well do language models perform on this task when fine-tuned on training data that is created by humans compared to when fine-tuned on training data created by large language models?

The data set, fine-tuned models, and example model output on test data are available through `github.com/rubenjanss/visual-conversation-starters2`.

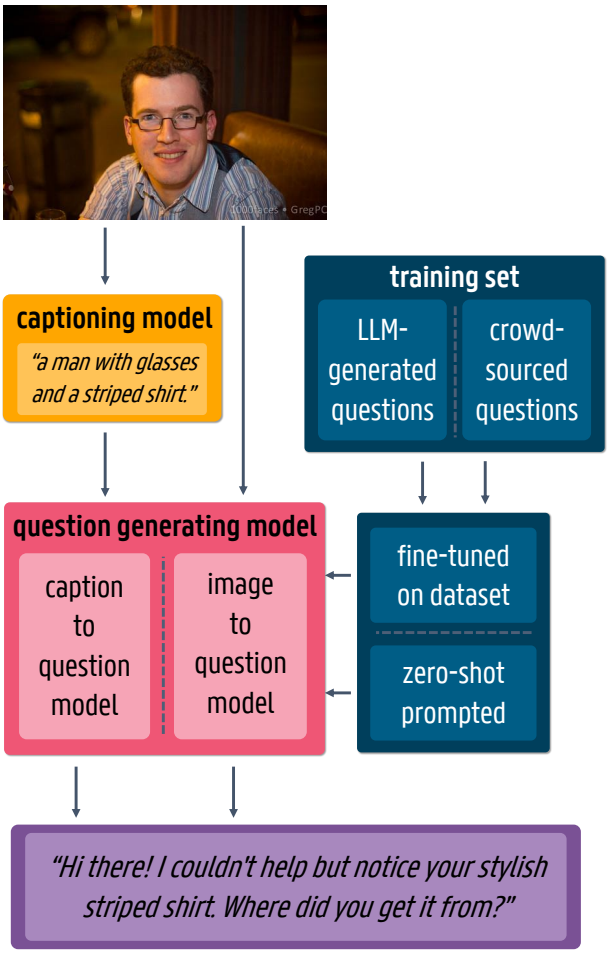

Fig. 1. Overview of the work presented in this paper, showing the different options: using captions or using an end-to-end image-to-question model, fine-tuning the system or using zero-shot prompting, and when fine-tuning, using the crowdsourced or LLM-generated questions.

2 Related Work

2.1 Personalising interactive AI through multi-modal information

Multi-modal sensory inputs are an essential part of human-human communication [16], and previous research has argued that embodied conversational agents should also possess multi-modal capabilities, such as the ability to interpret non-verbal cues, to build trust and empathy with users [17]. For conversational agents in general, adding vision to language is expected to create engaging experiences, especially when becoming embodied as a virtual agent or as a social robot [18].

Previous work has suggested that embodied AI systems can be made more engaging through personalisation, even specifically by personalising how the system greets the user. For example, in the field of embodied intelligent tutoring systems, Cooney et al. argued that robotic teaching assistants could be made more engaging by using personalised greetings [19]. Glas et al. also argued that social robots can give users a sense of familiarity and warmth by greeting them with personalised comments, and built a system to generate such greetings based on novel appearances or

consistent behaviours that were observed, although these features could not yet be autonomously detected through visual input [20]. Pantazopoulos et al. presented a system that supports social conversations based on the visual scene and showed that this system is helpful and entertaining [21].

2.2 Natural language generation

The field of Natural Language Processing (NLP) has recently seen important progress with the advent of large pre-trained models built with the Transformer architecture [22]. Their success follows from the idea of transfer learning: such models can be efficiently pre-trained to absorb textual knowledge from large amounts of data with a simple training objective that does not require manual annotation (e.g., predict a masked word in an input sentence). These are then trained further (or 'fine-tuned') on the intended downstream task for which typically limited training data is available. Masked language models such as BERT [23] lead to very strong context-aware representations of words, i.e. tokens, to be used in downstream NLP tasks such as information extraction or sequence classification, whereas auto-regressive models such as GPT [24] are well-suited for generating natural language. For generating a sequence of text in response to a given sequence (e.g. machine translation or text summarization), pre-trained sequence-to-sequence (seq2seq) models such as T5 [25] or BART [26] are often used.

Recently, very large language models such as GPT-3 [27] have shown impressive performance at *in-context learning* or *zero-shot learning*. They perform well on a multitude of tasks while only having been given relatively few (< 100) training examples of these tasks, or even without any examples.

The release of ChatGPT also brought considerable attention to neural language models, showing impressive performance on many different tasks, often without needing any examples. ChatGPT, which is based on the GPT-3.5 model, and now available with GPT-4 as well [28], is particularly successful at maintaining multi-turn conversations and generating text in a natural, conversational tone.

2.3 Vision-and-language tasks

The progress in natural language processing has expanded to models that combine language and vision, working on tasks such as image captioning (describing an image in a short textual description), visual question answering (answering natural language questions about an image) [29], and visually grounded dialogue (expanding visual question answering into a multi-turn conversation about an image) [30]. This has even been expanded into dialogues about videos [31].

Some vision-and-language tasks are more directed at social and affective interactions, such as Personality Captions [32] and Image-Chat [33]. "Personality Captions" aims to generate personality-dependent image captions. Image-Chat takes this one step further and combines image captioning with the ability to have a conversation, as its goal is to generate a conversation-starting utterance as a response to an image, or to reply to an utterance in a conversation about that image. These tasks have been combined with an open-domain conversational agent in Multi-Modal Blender, an extension of the Blender chatbot that can also engage in dialogue that is grounded in an image [34]. More recently, the MMDialog data set was released, which is similar to Image-Chat, but is larger and is composed of real-world dialogues that were collected from social media instead of being created by crowdworkers [35].

More specifically focused towards question generation, Visual Question Generation [36] aims to generate conversation-starting questions based on an image, with the image being the shared discussion topic. Image-Grounded Conversations [37] expands upon this by providing textual context (such as a caption) with the image and a suitable response to the question. Finally, Emotional Dialogue Generation [38] focuses on response generation, also providing a data set of images, captions, questions and answers, and studies the role of sentiment analysis in producing emotionally appropriate responses.

However, the tasks and data sets mentioned above have limitations and are not fully suited for the social situated AI applications. All previously mentioned data sets contain visual data as shared conversation topic, instead of as visual context from the first-person perspective of one of the conversation partners: this means they are not situated in the physical world. First-person (also called egocentric) data sets exist as well, but, just as most multi-modal data sets, they are not social: they are mostly focused on question answering or simple dialogues about explicit features of the image (e.g. "What is this person doing?") instead of social dialogues that require implicit reasoning about visual features [39], or they are goal-oriented such as for navigation [40], spatial reasoning or embodied planning for robotics [41], or retail shopping [42].

Some data sets from a first-person perspective with conversational annotations exist, such as the Visually-Grounded First-Person Dialogue (VFD) data set [43], which contains first-person images of conversation partners, with annotations of probable utterances by conversation partners together with a possible verbal and non-verbal response. These situations are, however, more oriented towards conversations about specific actions or in specific settings (e.g, playing music or sports), and are less suited for general-purpose social talk. Egocentric video data sets have been released as well, such as EgoCom [44], which contains a mixture of non-visually grounded social conversations and conversations where objects in the environment are described, and Ego4D [45].

2.4 Vision-and-language models

Models for these tasks are typically designed using an encoder-decoder architecture, like seq2seq models, but with an image model as encoder and a language model as decoder. Traditionally, these models consisted of a CNN-based image encoder and an LSTM-based text decoder, such as in [32] and [36]. In more recent work, these have been replaced with Transformer-based encoders and decoders.

Recently, vision-and-language models have followed similar trends as language-only models: building very large models that are pre-trained on simple tasks that do not require human annotation. For these models, this is called vision-language pre-training, often using pairs of images

and their alt-text, scraped from the internet [29], [46]. BLIP [46] and GIT [47] are examples of such models that combine Transformer-based vision encoders and text decoders in various ways and use vision-and-language pretraining to achieve state of the art performance on vision-and-language tasks after being fine-tuned for these tasks.

There are also models that claim good performance in a few-shot setting on vision-and-language tasks, such as DeepMind's Flamingo model [48], although this model has not been publicly released. BLIP-2, which is public, also claims good few-shot performance on vision-and-language tasks such as visual question answering and image captioning [49].

3 Data

In order to train and evaluate models that generate questions to start a human-AI conversation, we required a data set of visual inputs and corresponding conversation starter questions. More precisely, this visual input consists of a single image that shows what the system would see when interacting with a user. For each of these images, we also need the textual output the model should produce, i.e. a phrase that would be interesting for the system to ask and that explicitly refers to a visible element. We describe first how the images of this data set were collected, followed by the questions.

3.1 Images

First, we explored which image sets are available in related tasks. There are similar tasks for which conversations are modelled based on an image [33], [36], but here the image takes the role of a shared discussion topic and does not show the conversation partner. Data sets for object detection and image captioning also have a less human-centric focus, which makes them unsuitable for our task.

To overcome this problem, we created our own data set that is appropriate for situated AI applications. We followed the same procedure as taken by others working on image-to-text tasks [33]. We selected a subset of appropriate images from a larger data set and created our own annotations for these images. Given that data sets for related tasks vary in size from 5,000 to 1 million images, the limited resources for annotation, and the characteristics of the models that were to be used for this task, we aimed to collect at least 5,000 images.

We decided on using the YFCC100M data set [50] as the basis for our own data set. It contains 100 million images and videos, with varied content, collected from Flickr. It is used as a source data set by related work like Image-Chat [33] and Visual Genome [51]. Given its size and large variety of subjects, it allowed collecting the required situated AI-appropriate images. Through its use in related research, it also provides comparability. With our task in mind, we defined a relevant image as a photo that contains exactly one person facing the camera.

First, the YFCC100M browser [52] was used to select images with the keyword "person", yielding 73k images. Face recognition was used to retain only images containing a single person, further reduced to only those where the face covered at least 5% of the area of the image, which effectively discards images where the person is only in the background. Face detection was done using the YOLOFace model[1]. Only images with a minimum width of 300 pixels were kept, and finally, unrealistic and unusable images were discarded after manual inspection. This resulted in a total of 7928 images to be included in our data set. We will refer to this data set as the Flickr data set in the rest of this paper.

Fig. 2. Sample of images from the training data set. Associated training example questions: *(upper left)* `Have you always had freckles? Were you born with red hair? Is white your favourite colour?` *(upper right)* `Is that cool hat typical of your country? Do your parents also have blue eyes? Where did you get that scarf?` *(lower right)* `Do you go to bars often? What kind of drinks do you like? Where did you get that necklace?` *(lower left)* `Are you eating soup? Do you put the braids in your hair? Is your favourite colour yellow?`

3.2 Questions

For each image appropriate conversation starters were needed. We collected these through crowdsourcing on the Amazon Mechanical Turk (MTurk) platform. Each crowd worker evaluated a unique image, and their task was to come up with three conversation-starting questions. Remuneration was in line with the US minimum wage, and ethical guidelines of the university were followed.

4000 images were randomly selected from the larger Flickr data set. Of those, 3471 images were annotated by crowdworkers. We annotated the other 529 images ourselves, resulting in a total of 4,000 annotated images with 12,000 associated questions. This data set was split into training (3000 images), validation and testing sets (both 500 images). Through a qualitative visual inspection, we found that, even without explicit annotation guidelines, most of the questions focus on appearance (e.g., accessories, clothing, or hair). They mostly tend to follow similar structures as well, such as `Where did you get your...?`, `How long have you had...?`, or `Do you like having...?`.

A sample of images from the training set is shown in Figure 2, along with the crowdsourced questions associated with them.

1. https://github.com/sthanhng/yoloface/

4 MODELS

To answer the three research questions presented in the introduction, we explore six different model architectures. We will investigate whether the different model architectures are able to refer to the visual input correctly, while generating natural, polite and interesting questions and maintaining the embodiment of the conversation (i.e. making sure the questions address the person to whom the visual features belong, without referring to the image itself).

The first approach uses textual descriptions, i.e. captions, as representations of the visual input and fine-tunes a language model using the questions gathered through crowdsourcing (the "caption-to-question model"). This approach transforms the problem into a text-to-text problem, allowing us to use powerful language models. However, the captioning model can also be a bottleneck: only features that the captioning model is trained to recognize, can be used by the question generating model. Errors made by the captioning model will also propagate in the rest of the system.

In the second approach, we do not use the captions anymore but directly fine-tune a vision-and-language model on the human data (the "image-to-question model"). This eliminates the reliance on the captioning model. However, this makes the task more complex, as the model needs to both recognize relevant features in the image and transform this into a good sentence, possibly requiring more data to train the model on.

For the third approach, we prompt a large language model, namely GPT-3.5, without fine-tuning it, and use captions as a representation of the visual context. Large language models contain more parameters than the other models used in this paper (ca. 12-175B parameters for GPT-3.5 compared to 400-700M for BART and BLIP), have shown impressive zero-shot performance and are trained on a much larger data set, hopefully leading to more varied and human-like output. We also explore the zero-shot performance of multi-modal (vision-to-text) large language models, but discover that they are not yet capable of understanding this specific task.

Finally, we use GPT-3.5 to generate synthetic data and use this to fine-tune the caption-to-question model and image-to-question model. As these models are smaller than GPT-3.5, they can be used without needing a large GPU cluster or paying for an API, and allow for faster inference times and lower energy usage. We will investigate whether the smaller models are able to absorb the larger model's skills that are necessary for this task.

For each architecture, we carry out a preliminary manual evaluation of the models on the validation set, because automatic metrics like BLEU and ROUGE have shown weak or no correlation with human judgement on this task [15]. Manual evaluation was performed by a researcher scoring conversation starters generated for a subset of 50 images from the validation set on five criteria:

- **Language:** How good is the writing of this question, with regards to grammar, spelling, and clarity?
- **Visual reference:** The question correctly mentions something that is visible in the image.
- **Interestingness:** How interesting do you find this question as a conversation starter?
- **Specificness:** How specific is this question?
- **Politeness:** Provided the question is asked in a friendly way, how polite is this question?

All criteria were rated on a discrete scale from 1 to 5, except for interestingness and specificness, for which a scale of 1 to 7 was used to capture more nuanced differences. The mean score over all dimensions (after normalizing all criteria to a 0-1 scale) is used to select the best model version.

Note that this initial model evaluation holds only limited representative power for user judgement, as it is carried out by only one (expert) rater. The purpose of this evaluation is to replace the automatic metrics that are typically used with the validation data set for early stopping, selecting hyperparameters, and other model options. For this reason, no statistical tests are carried out on these results.

4.1 Caption-to-question model using crowdsourced data

The caption-to-question model uses an existing captioning model as-is, followed by a sequence-to-sequence model that will be fine-tuned. This is the same approach as used in the pilot study presented in [15]. We first compare available captioning models and then fine-tune the sequence-to-sequence model.

Captioning model

Recently released captioning models include BLIP-2 [49], GIT [47], and GRiT [53]. BLIP-2 and GIT achieve state-of-the-art performance on the COCO data set, which is a standard benchmark for image captioning. While these models output one-sentence captions that summarise the entire image, GRiT generates "*dense captions*": one-sentence descriptions of multiple individual elements of the image. The model excels on the Visual Genome data set, which is the benchmark for dense captioning.

Dense captioning provides an advantage over regular image captioning: the multiple short sentences, each describing a different aspect of the image, provide a richer description than the regular captioning images that tend to focus on a few features in their caption. For example, for a given image, GIT would generate the (regular) caption `a woman in sunglasses and a man in the reflection of the sky`, while GRiT would generate the dense captions `black sunglasses on face. reflection of a person in a mirror. the scarf is purple. the woman has red hair.`

We compare BLIP-2, GIT and GRiT by having them generate a caption for 50 images from the validation set, and manually scoring those captions for whether they correctly reference visual features of the image ("visual reference" score), and how specific the features they describe are ("specificness" score). A researcher manually scored these captions on a 5-point scale for both criteria.

On average, the BLIP-2 model received a visual reference score of 4.32 and a specificness score of 4.7. GIT received 4.7 and 4.16, and GRiT 4.06 and 4.72. It seems that there is trade-off between correctly referencing visual features and including more specific features in the caption. We test both

TABLE 1
Validation Results of the Caption-to-Question Model (BART)

Caption model	Learning rate	Epochs	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
GIT	5e-5	1	5.00	2.62	6.00	5.00	5.00	0.781
	1e-5	8	5.00	4.18	5.82	6.04	5.00	0.888
	5e-6	13	4.98	**4.64**	5.72	5.88	4.96	**0.899**
GRiT	5e-5	3	5.00	1.94	6.00	6.00	5.00	0.780
	1e-5	6	5.00	4.04	5.64	5.70	4.98	0.862
	5e-6	6	5.00	4.00	5.76	5.76	5.00	0.867

TABLE 2
Validation Results of BART With Beam Search, Top-K Sampling, and Top-P Sampling

Decoding strategy	Parameter	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
Greedy search		4.98	4.22	5.66	5.80	4.98	0.874
Beam search	beam width = 3	5.00	4.20	5.72	5.92	4.98	0.880
	beam width = 10	4.98	4.64	5.72	5.88	4.96	**0.899**
Top-k sampling	$k = 9$	4.98	4.36	5.74	6.04	4.98	0.892
	$k = 11$	5.00	3.96	5.54	5.66	4.98	0.854
Top-p sampling	$p = 0.50$	5.00	4.00	5.62	5.94	4.96	0.867
	$p = 0.90$	5.00	4.40	5.76	5.90	4.98	0.891

GIT and GRiT in the remainder of this work, to compare the impact of richer but less correct captions with those that are certainly correct but contain less specific information, and see how the language models will handle the difference between those captions.

Sequence-to-sequence model

Provided with captions generated by the captioning model and example questions crowdsourced for images from the Flickr data set, we can fine-tune a sequence-to-sequence model to generate the visual conversation starters based on captions.

We fine-tuned the sequence-to-sequence model BART [54], which is particularly effective for text generation tasks such as question answering and summarization tasks, due to its architecture containing both a bidirectional encoder and autoregressive decoder. As our task is conceptually the same as a summarization or question answering task (generating a sequence of words of variable length based on another sequence of words, also of unknown length), this seemed to be the most applicable architecture, as opposed to encoder-only or decoder-only models like BERT or GPT that do not work in a sequence-to-sequence fashion.

The pre-trained BART model was fine-tuned starting from the `facebook/bart-large` checkpoint accessed through the HuggingFace platform, using the training part of the Flickr data set consisting of 3000 images, with 3 questions for each image.

We compared the model's performance when using the (single-sentence) captions generated by GIT as input with the dense captioning generated by GRiT. Furthermore, we compared three learning rates: 5e-5, 1e-5, and 5e-6. We employed early stopping when fine-tuning the model: the model was always trained for 20 epochs, but evaluated on the validation set after every epoch. The training length with the best validation performance was then selected. Beam search with beam width of 10 was always used as the decoding strategy.

Table 1 shows the results of the manual evaluation after early stopping. The regular captioning model GIT seems to outperform the dense captioning model GRiT, in particular due to its much higher visual reference score. This means the higher visual reference score observed in the earlier evaluation of the captioning models is also reflected when applying them together with the caption-to-question model. However, GRiT's higher specificness score is not reflected here, perhaps because the example questions in the data set do not use the very specific elements provided by GRiT.

With regards to the learning rate, a lower learning rate seems to lead to better performance, especially with regard to visual reference. The learning rate of 5e-5 in particular is not sufficient for this task, as most model versions trained using this learning rate produced the same question for each validation image.

Finally, we investigated the impact of using different decoding strategies, on the BART model using GIT captions trained for 13 epochs with learning rate 5e-6. Decoding strategies are the algorithms with which the language model generates the most likely sequence of tokens. The most intuitive decoding strategy is greedy search: selecting the most probable token one step after another. Beam search is an extended version of this, where n (the "beam width") sequences of tokens are being evaluated in parallel, adding the most probable next token to each of the n sequences at once. Top-k and top-p sampling are other extensions of greedy search, where the next token at every step is randomly selected from either the k most probable tokens or the set of tokens with a cumulative probability mass of p. Top-p sampling, also called nucleus sampling, has been suggested to generate less bland text [55].

Besides greedy search, we evaluated the model's validation performance using beam search, with beam widths 3, 5, 10, and 15, using top-k sampling ($k \in [3, 15]$ with increments of 2) and top-p sampling ($p \in [0.05, 0.95]$ with 0.05 increments). Table 2 shows the best and worst results of each decoding strategy on the validation set. The impact of changing the decoding strategy is limited for BART, although the beam width, k, or p value still needs to be selected carefully, especially with regard to the visual reference score. Beam search with a beam width of 10 is selected

TABLE 3
Validation Results of Image-To-Question Models

Model	Learning rate	Epochs	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
ViT-GPT-2	5e-5	15	4.92	4.34	5.68	5.64	4.90	0.869
	1e-5	12	4.98	4.10	5.62	5.62	5.00	0.862
BLIP	5e-5	10	5.00	4.52	5.52	5.44	4.88	0.869
	1e-5	4	5.00	4.82	5.68	5.88	4.98	**0.909**
GIT	5e-5	5	5.00	3.72	6.00	6.00	4.84	0.861
	1e-5	5	5.00	4.60	5.74	5.76	5.00	0.897

TABLE 4
Validation Results of BLIP With Beam Search, Top-K Sampling, and Top-P Sampling

Decoding strategy	Parameter	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
Greedy search		4.98	4.24	5.67	5.78	4.96	0.874
Beam search	beam width = 3	4.96	4.84	5.64	5.84	4.98	0.905
	beam width = 10	5.00	4.82	5.68	5.88	4.98	**0.909**
Top-k sampling	$k = 3$	4.87	4.50	5.41	5.76	4.98	0.873
	$k = 15$	4.50	4.07	5.35	5.12	4.57	0.789
Top-p sampling	$p = 0.25$	4.96	4.74	5.74	5.89	5.00	0.906
	$p = 0.80$	4.17	4.00	5.20	5.39	4.54	0.771

as the best decoding strategy and is used for the caption-to-question model in the following evaluations.

4.2 Image-to-question model using human data

We study whether the newly released Transformer-based vision-and-language models can learn to generate conversational questions that accurately refer to visual features by being fine-tuned on our data set.

Three models were selected, that are recently released and achieve state-of-the art performance on image captioning tasks: GIT, BLIP, and ViT-GPT-2. They were all accessed through the HuggingFace platform, using checkpoints `microsoft/git-large-coco`, `Salesforce/blip-image-captioning-large`, and `nlpconnect/vit-gpt2-image-captioning`. As is evident from the checkpoint names, the models were already fine-tuned for image captioning after being pre-trained. We hypothesise that this approach will work better than fine-tuning the model with our data set directly after it was pre-trained, as the models should be able to leverage the training data for image captioning (with training sets containing a number of images that is in the order of hunderds of thousands), through transfer learning.

We train each of the three models using learning rates 5e-5 and 1e-5, and use early stopping as described in the previous section. Beam search with a beam width of 10 was always used as the decoding strategy. Table 3 shows the results of the manual evaluation of these models after early stopping. All models were trained for 20 epochs, except for the ViT-GPT-2 model at learning rate 5e-5 which was trained for 15 epochs, and BLIP at learning rate 5e-5, which was trained for 50 epochs. Except for ViT-GPT-2, it seems that the lower learning rate is better, and that training for longer than 20 epochs is not needed. In particular, GIT, when trained using learning rate 5e-5, always produces the same one or two sentences, for all images in the validation set. BLIP and GIT achieve markedly better performance than ViT-GPT-2 with regards to visual reference, and similar performance on the other criteria. The BLIP model, trained for 4 epochs at learning rate 1e-5, is used in the remainder of this paper.

We also investigated the influence of different decoding strategies on the image-to-question models. We evaluated the model's validation performance using greedy search, beam search, with beam width 3, 5, 10, and 15, top-k sampling ($k \in [3, 15]$ with increments of 2) and top-p sampling ($p \in [0.05, 0.80]$ with 0.05 increments). Table 4 shows the results of each decoding strategy with its best and worst parameter.

Interestingly, the model generates nonsensical questions at high values of p or k. These questions can be semantically nonsensical (e.g. *"Does your hat look much like riding a bike?"*) or even use combinations of tokens that do not represent words in the English language (e.g. *"Have you had your ear ringsrals served intland of your pearls?"*). For some images, the model generated an empty string. Top-p sampling with $p > 0.80$ was also attempted, but resulted in mostly nonsensical questions, so these results were not included in the evaluation.

In general, higher values of p or k lead to more diverse formulations of questions and more specific elements being recognised, also when compared to beam search, but also to higher chances of language mistakes or nonsensical sentences. Some examples of specific elements that are recognized by top-p or top-k sampling but not by beam search are: a guinea pig, face painting, a convention, a cultural performance. As is logical with more specific features being referenced, there are also more visual reference mistakes. Using beam search with a beam width of 10 performs best, so this decoding strategy is used in following evaluations.

4.3 Zero-shot large language model using captions

Large language models such as GPT-3.5 (ChatGPT) have shown great zero- or few-shot performance in both natural conversations and many natural language tasks. We design and evaluate different prompts to investigate the performance of this model on our task and we compare the use of dense captions generated by GRiT with regular captions

TABLE 5
Validation Results of GPT-3.5 With Different Prompts and Captioning Models

Captioning model	Prompt	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
GRiT	Simple	4.92	4.24	6.22	6.60	4.96	0.917
GIT	Simple	5.00	4.64	6.24	6.24	4.94	0.928
	Advanced	4.96	4.58	6.08	6.24	4.9	0.916

generated by GIT. We use the model checkpoint which has been retroactively named `gpt-3.5-turbo-0613`.

First, we compared the two captioning systems, using a simple prompt. GPT-3.5 prompts are split up into a system and user prompt, the former setting the tone for the entire conversation, with the latter being a first conversational turn. Our simple prompt consisted of the system prompt "*You are meeting a person in real life. Write a question to start a conversation with this person that references something you see.*" and user prompt "*You see the following: "*, followed by the caption.

The GPT-3.5 results were evaluated in the same manner as the previous models: questions were generated for the same subset of 50 images from the validation set, and manually rated on five criteria. Results of this evaluation are shown in Table 5.

Table 5 shows the clear difference between the GRiT and GIT captions, which reflect what was already noted earlier: GIT receives a much higher visual reference score, while GRiT scores higher on the specificness scale. As visual reference mistakes can have a high impact during real-world interactions, we chose to continue with the GIT captions.

This system generates much more specific and interesting questions than the pre-trained systems. First of all, the questions are written more conversationally: compare the GPT-3.5-generated "*Hey there! Nice jacket! I was just wondering where you got it from?*" with the BLIP-generated "*Where did you get your jacket?*". Furthermore, even when referring to the same visual features, GPT-3.5 can ask more interesting questions. For example: BLIP generated "*Where did you get that red flower in your hair?*", while GPT-3.5 generated "*Hi there! I couldn't help but notice your beautiful red flower. Is there a special meaning behind it or is it just for aesthetic purposes?*". Finally, when the caption mentions a specific feature that was not present in the training set, GPT-3.5 can nonetheless ask a question about it, such as "*That's a beautiful flower crown, did you make it yourself?*", while there were no questions about flower crowns in the training set. These aspects together improve the interestingness and specificness scores of GPT-3.5 when compared to BART and BLIP.

However, some questions generated by GPT-3.5 are not very suitable, leading to a suboptimal start of the interaction. For example, some questions are rather generic, e.g. asking someone where they got their stylish outfit, which is a question that could be asked to anyone. Some questions also refer to an image (e.g. "*Hey, I noticed you in the background of that photo laughing. What was going on in that moment? "*) or talk about the user in the third person (e.g. "*Who is the girl in the photo? She looks interesting.*"). Furthermore, some questions are extremely verbose and specific or reference multiple aspects of the image but only ask about one of them (e.g. "*Hey there! I couldn't help but notice your charming smile in that blue jacket. Is it a new one or is it your favorite go-to jacket?"*), which feel unnatural. Also, the model sometimes hallucinates: it will refer to some visual feature that is not present in the caption, (e.g. referring to a "*vibrant red dress*" while the caption is "*a woman walking in the street*", or being more specific than the provided information allows, e.g. referring to a "*Nikon D3500*" while the caption only mentions a camera). Hallucination is an often-observed phenomenon in deep learning based natural language generation [56]. Finally, the formatting of the output is also inconsistent. Sometimes, the model will first write an explanation of why it would ask this question, and sometimes it will write quotation marks around the question, but not always, making it difficult to parse the output to feed it to a situated interactive system.

We tried several prompting techniques to resolve these problems and improve GPT-3.5's output, which led to the "Advanced prompt" shown in Table 6.

The first technique was making the prompt more detailed, describing the persona the model should take on, the situation of the interaction, and the requirements of the questions. Secondly, we prompted the model using step-by-step instructions. This approach, called Chain-of-Thought prompting, has been shown to improve reasoning capabilities in large language models [57]. The often-occurring mistakes were explicitly addressed in these steps, guiding the model's "thought process" to minimize chances of these mistakes. To facilitate parsing of the output, the prompt asks the model to delimit the questions with XML tags, as it often included the answers to the intermediate steps in the output.

TABLE 6
Final GPT-3.5 Prompt

System prompt
You are a friendly and engaging social robot. You are meeting a person in real life and want to start a conversation with them. Your task is to write an interesting question to start a conversation with this person. This question should address this person and should reference something that you can see about this person. The question should be as specific as possible to that person and to what you can see, it should not be a question that you can ask just about anyone. The question should also not be inappropriate, biased or politically incorrect in any way. Do not mention a photo. Do not mention any other people you might see.

User prompt
Let's think step by step. Step 1: Choose one aspect that is explicitly mentioned in this description, and forget about any picture or camera: `<caption>` Step 2: Make sure this aspect was explicitly mentioned in the description. Step 3: Write a conversation-starting question about this aspect, referring to something happening in this moment. Delimit the question with XML tags <question></question>.

Two approaches were tested but not included in the final prompt. We tried adding negative examples to the more detailed prompt: adding three questions which talk about the person in the third person, and explicitly saying these questions are not good. However, this approach was not effective at reducing the presence of this mistake in the model's output.

We also implemented self-criticism: feeding the model's output back into the model and asking it whether this followed all the instructions, or even asking it to rewrite a question that did not follow the instructions [58]. However, this approach performed inconsistently and often incorrectly for this task, even when reducing the temperature parameter of GPT-3.5.

The final prompt, that was used for the evaluation labelled "Advanced prompt" in Table 5, is shown in Table 6.

On the validation set, this prompt receives a slightly lower score than the simple prompt – however, qualitatively, the results of this advanced prompt showed to be more consistent, concise, and seemed to show fewer hallucinations and occurrences of the mistakes described above. Therefore, in order to minimise mistakes that could be disruptive to the interaction, the advanced prompt is chosen for following evaluations.

4.4 Zero-shot image-to-question model

Some vision-and-language models have recently been released that claim zero-shot capabilities. Notably, the large language model GPT-4 was announced to be multi-modal, but this functionality is not available at the time of writing this paper. BLIP-2 [49], which is open-sourced, also boasts zero-shot image-to-text generation: an example prompt they use is "*Write a romantic message that goes along this photo.*"

We tested BLIP-2 through the HuggingFace platform, using the `Salesforce/blip2-opt-6.7b` checkpoint. However, the model did not seem to fully understand the prompts we tested. A simple prompt, "*Q: How would you start a conversation with the person in this image about something you see. A:*", resulted in "*I would start with a smile and a hello.*", which is conversation-starting, but neither a question nor referencing the image. Other simple prompts resulted in descriptions of the image, e.g. "*I see a woman with a smile on her face.*"

More advanced prompts, which include example questions, also did not result in satisfactory outcomes. One such prompt is "*Q: How would you start a conversation with the person in this image mentioning something you see. A: Where did you get that white sweater? Q: How would you start a conversation with the person in this image about something you see. A: Do you like having long hair? Q: How would you start a conversation with the person in this image about something you see. A:*". Advanced prompts such as this one either led to repeating the last example question, or to questions that did not reference the image, such as "*Do you like to travel?*" or "*What is your favourite colour?*"

BLIP-2 can also be configured with the Flan-T5-XXL language model instead of the OPT-6.7B model. Flan-T5-XXL contains more parameters and is instruction-tuned, compared to the smaller OPT-6.7B model that is trained using unsupervised learning, so it should perform better [49]. However, the results are still not satisfactory.

For some images, reasonably good questions were generated, such as "*What is the reason you are wearing a leather jacket?*" and "*What are you doing on the beach?*". However, for many images that contain less salient features, the model would ask "*What is the name of the person in the picture?*". These results were obtained with a step-by-step approach, using a slightly modified version of the prompt used with GPT-3.5.

Focusing the step-by-step reasoning prompt on correct visual reference, led the model to generate questions like "*What is the color of your jacket*" or "*What is the woman wearing?*", that mention a visual feature but whose answer is already visible in the image, or are talking about the user in the third person, so that are not interesting to start a social conversation.

The step-by-step approach could also be modified to feed back the model output for the first step into the model, to generate the second step. When asking the model to describe everything you can see about this person and their surroundings, it shows a detailed awareness of the image ("*the person is wearing a brown leather jacket and has long brown hair*"). However, asking it to generate a question based on that output results in questions like "*What color is the wall behind the woman?*", disregarding the previously produced output. Furthermore, this approach is conceptually very similar to the caption-to-question approach, except for the addition of visual context during the caption-to-question step.

Shorter prompts, that focused more on the interestingness of the question, often resulted in the question "*What is your favourite colour?*". When prompting the model to start a conversation with the person in the image, sentences like "*Hey, I'm a photographer and I'm looking for a model*" or "*Hi, you look really pretty today, what are you wearing?*" were generated, which are not appropriate.

Therefore, we conclude that zero-shot text generation capabilities of image-and-language models available at the time of writing this paper are not yet sufficiently mature for this situated social task and do not offer a unique added value when compared to the other approaches.

4.5 Fine-tuned models using synthetic data

As GPT-3.5 is able to generate much more interesting and specific questions, with a similar visual reference score to the fine-tuned caption-to-question model, the question arises of whether the lower performance of the fine-tuned models is due to the data that they were trained with, or a fundamental limitation of the smaller models, perhaps also indicating that a locally run image-to-question model with comparable performance is unfeasible. To answer these questions, we fine-tuned both the caption-to-question model BART and the image-to-question model BLIP on a data set generated by GPT-3.5.

First, we created a data set of the same size as the original data set. The same images were used, and three questions per image were generated by GPT-3.5. These images were generated using the advanced prompt described in Section 4.3, with the additional step "*Step 4: Write two more conversation-starting questions about an aspect from the description, referring to something happening in this moment.*

TABLE 7
Validation Results of BART and BLIP Trained on Synthetic Data

Model	Learning rate	Epoch	Language	Visual reference	Interestingness	Specificness	Politeness	Mean
BART-GPT	5e-6	9	4.98	4.66	6.06	6.22	4.96	0.923
BLIP-GPT	5e-6	5	5.00	4.68	5.98	6.16	5.00	0.922

Delimit each question with XML tags <question></question>." appended to the user prompt.

Then, both BART and BLIP were fine-tuned on this data set, starting from the same original checkpoints as used when fine-tuning them in sections 4.1 and 4.2. For BART, the data was again coupled with the GIT captions, as they showed to provide the best performance in previous steps. For both models, learning rate 5e-6 was used, and the models were trained for 20 epochs but were again evaluated after every epoch for early stopping. Beam search with beam width 10 is used as decoding strategy, as this consistently showed to provide the best results.

Table 7 shows the results of BART and BLIP when fine-tuned on the GPT-3.5 generated data. The labels BART-GPT and BLIP-GPT are used to disambiguate these model versions from the ones that were trained on the crowdsourced data. These results are encouraging, as they show results that are very similar to the original GPT-3.5 evaluation and are still better than the results of the originally fine-tuned BART and BLIP models. BLIP only slightly excels BLIP-GPT with regards to visual reference.

5 Human evaluation

In order to assess the human perception of the performance of the models, we compare the performance of different model architectures with each other and with human performance through a crowdsourced human evaluation.

5.1 Evaluation setup

The models will be assessed on the same five criteria as used in Section 4, using the descriptions shown there. For each of the five criteria, a human rater evaluates a question for a given image on a Likert item with seven points.

While in Section 4, five-point scales were used for language, visual reference, and politeness, we opted to use seven-point scales for all criteria in the crowdsourced evaluation to reduce cognitive load for crowdworkers. Note that the scores reported in this human evaluation should not be compared with those in Section 4, as those were not scored by multiple crowdworkers, and only served to compare model versions within the same architecture type.

This evaluation method is adapted from the pilot study presented in [15], but was changed in multiple ways. First, all items use a seven-point scale, instead of binary or ternary scales, to capture more nuanced differences in quality. Second, the item "Specificness" was added, to disambiguate the concepts of interestingness and specificness, which were linked together in the previous work. Finally, the item "Politeness" was previously called "Appropriateness". This was changed to remove any potential confusion with "Visual reference", which could be interpreted as appropriateness of the question for the situation.

Each of the five models was evaluated on a test set of 50 images - the same images for each model. 25 of these images were taken from the testing part of the data set gathered from Flickr, as described in section 3.1. These images were collected using the same methodology as the training set, but were not seen by the models during training.

The other 25 images were gathered in human-robot interaction experiments at the Nerdland Science Festival in Belgium. Participants stood in front of a Furhat social robot, which greeted them with a conversation-starting question generated by the generative model presented in [15], and then had a short scripted conversation with the robot. These conversations were recorded using an Intel RealSense webcam, and one frame was extracted from 25 of these video recordings to be used as test input for these models. These pictures are much more representative of real-world input, as the Flickr images contain more salient features than would be realistic for real-world images. Furthermore, this also tests the generalizability of the models to data that is distributed differently from the training data.

Each of the five models generated a question for each of the 50 test images. Additionally, for each of the test images, a human-written question was also evaluated. For the Flickr images, this was one of the three questions that were gathered in the original crowdsourcing campaign described in Section 3.2. For the Nerdland images, we conducted a new crowdsourcing campaign to gather questions. This followed the same methodology as for the Flickr images, but was conducted on Prolific instead of MTurk, as we saw that the quality of data gathered on MTurk has worsened when compared to the original campaign.

The resulting 300 image-question pairs were grouped into series of 18 items, with each series containing at least two image-question pairs from each model and from the crowd-workers, with no image occurring more than once in a series. Each series additionally also contained two attention checks (Instructional Manipulation Checks [59]), where the participants were not shown an image but rather explicitly instructed which scores they should select. Participants who failed both attention checks were asked to stop the study and were not included in the final results. The images and attention checks were presented in a randomly shuffled order. Each series was rated by 4 participants, and no participant could rate more than one series. Participants were first shown 6 example image-question pairs with ratings shown, in order to calibrate them to the scale.

Raters were recruited on the crowdworking platform Prolific, which is noted for providing high-quality data [60]. Only participants who are fluent in English were eligible for the study. They also had to have indicated in their Prolific account that they are comfortable taking part in a study where they would be deceived, as we only revealed at the end of the study that some questions were generated by AI, and where they could be exposed to harmful content, as

TABLE 8
Crowdsourced Evaluation Results (Scores Range From 1 to 7)

Model	Language		Visual reference		Interestingness		Specificness		Politeness		Mean	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BART (caption-based)	6.55	0.63	4.61	2.57	3.35	0.96	5.17	1.00	4.55	0.94	4.84	0.75
BLIP (image-based)	6.53	0.49	**5.89**	1.72	3.62	0.94	5.31	0.84	4.76	0.84	5.22	0.62
GPT-3.5 (caption-based)	6.72	0.39	5.23	2.15	4.44	1.19	**5.90**	0.83	5.54	0.77	5.56	0.71
BART-GPT (caption-based)	6.66	0.49	5.35	2.16	**4.59**	1.13	5.60	0.94	5.68	0.95	5.57	0.78
BLIP-GPT (image-based)	**6.79**	0.34	5.63	1.89	4.57	1.23	5.79	0.73	**5.99**	0.99	**5.75**	0.76
Human (image-based)	6.17	0.85	5.84	1.14	3.85	1.13	5.17	1.03	4.82	1.08	5.17	0.68

(a) BART: `Where did you get your earrings?` BLIP-GPT: `Hi there! I couldn't help but notice your lovely smile. What's making you so happy today?` Picture from `flickr.com/harryprayiv`.

(b) BART: `Where did you get your glasses?` GPT-3.5: `Hey, I noticed you're wearing glasses and a colorful shirt. I'm curious, do you think your fashion choices reflect your personality in any way?`

Fig. 3. Selected output of the tested models on (a) an image from the Flickr test set and (b) the Nerdland data set. GPT-3.5 shows more elaborate and specific questions than BART, while BLIP-GPT shows that it is able to mimick GPT's elaborateness and additionally is better at visual reference than the caption-based model.

some of the AI-generated questions could contain harmful biases. Raters were paid 9 GBP per hour, the advised payment rate by Prolific (the minimum rate is 6 GBP per hour), for an estimated completion time of 15 minutes. The actual median completion time was 12 minutes and 10 seconds, making the actual average reward 11.10 GBP per hour.

The research was conducted according to the ethical regulations (General Ethical Protocol) of the Faculty of Psychology and Educational Sciences of Ghent University.

5.2 Results

Figure 3 shows the output of two models for two images from the set of 50 images the models were tested on. The supplementary material contains an overview of all models' output on the 25 Flickr test images.

Table 8 shows the average of all ratings (between 1 and 7) for the different criteria and models. The "mean" column shows the mean of all criteria. The significance of the differences between models is tested using non-parametric tests as the ratings are not normally distributed. The Kruskal-Wallis test shows that there is a significant difference in mean score between the six models (including the human-annotated questions), with $p < 0.001$. Pairwise comparisons between models were done using the Mann-Whitney U test, using the Bonferroni correction for multiple testing.

On the language correctness criterion, all models score highly, and no significant differences are found except for between BLIP-GPT and BLIP ($p < 0.05$). Except for BLIP, all models do score significantly better than the human-written questions, showing a first mark of the low quality of crowdsourced text.

With regards to visual reference, BLIP and BLIP-GPT score the highest, which shows the superior capacity of end-to-end vision-and-language models on this aspect. The spread on the visual reference score is higher than that of the other criteria, both BLIP and BLIP-GPT's visual reference scores are only significantly different from that of BART ($p < 0.05$).

The interestingness, specificness, and politeness criteria show a clear difference between GPT-3.5 and the models that were fine-tuned on the synthetic data on one side, and the models that were fine-tuned on the crowd-sourced data and the crowd-sourced questions themselves on the other side. GPT-3.5 and BLIP-GPT have a significantly higher interestingness, specificness and politeness score than BART, BLIP, and the crowd-sourced questions ($p < 0.05$ when comparing GPT-3.5 and human performance on interestingness, and BLIP and BLIP-GPT on specificness, $p < 0.01$ when comparing BLIP-GPT with BART and human performance on specificness, and $p < 0.001$ in all other cases).

Comparing all architectures on the overall mean score, it is clear that the BLIP-GPT model scores the highest, the difference being significant with BART, BLIP, and the human performance (all $p < 0.001$). BLIP-GPT's high overall score is explained by its high visual reference score, which it shares with BLIP, and its high interestingness, specificness, and politeness scores, which it shares with GPT-3.5 and BART-GPT.

When comparing the Flickr and Nerdland data sets, shown in Table 9, we see a slight but significant downward shift for all models when transferring to the Nerdland data set: the average mean rating for all models and human performance is 5.42 for Flickr and 5.29 for Nerdland ($p < 0.01$).

6 Discussion and Conclusion

6.1 Discussion and Limitations

These results show that the end-to-end image-to-text models provide a significant advantage over the models that use a caption as intermediate representation. Furthermore, it is clear that GPT-3.5 provides more interesting, specific and polite questions than crowdworkers. The BART-GPT and BLIP-GPT results show that GPT-3.5's ability to generate

TABLE 9
Crowdsourced Evaluation Results Split Between Flickr and Nerdland Data Sets (Scores Range From 1 to 7)

Model	Data set	Language		Visual reference		Interestingness		Specificness		Politeness		Mean	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BART (caption-based)	Flickr	6.57	0.72	4.86	2.49	3.38	1.10	5.15	0.80	4.40	0.93	4.87	0.77
	Nerdland	6.52	0.53	4.36	2.67	3.31	0.82	5.18	1.18	4.70	0.94	4.81	0.75
BLIP (image-based)	Flickr	6.57	0.45	6.10	1.24	3.65	0.89	5.29	0.89	4.70	0.74	5.26	0.51
	Nerdland	6.48	0.52	5.68	2.09	3.58	1.01	5.33	0.81	4.81	0.94	5.18	0.72
GPT-3.5 (caption-based)	Flickr	6.77	0.37	5.56	1.83	4.67	1.04	5.95	0.81	5.56	0.81	5.70	0.64
	Nerdland	6.66	0.42	4.90	2.43	4.20	1.31	5.84	0.86	5.51	0.75	5.42	0.77
BART-GPT (caption-based)	Flickr	6.68	0.53	5.64	1.98	4.62	1.17	5.75	0.84	5.63	1.03	5.66	0.75
	Nerdland	6.64	0.45	5.05	2.34	4.55	1.10	5.44	1.03	5.72	0.87	5.48	0.82
BLIP-GPT (image-based)	Flickr	6.81	0.32	6.08	1.19	4.91	1.12	6.00	0.66	6.12	1.12	5.98	0.62
	Nerdland	6.76	0.36	5.18	2.34	4.22	1.25	5.58	0.74	5.85	0.84	5.52	0.82
Human (image-based)	Flickr	6.04	0.89	5.90	1.19	3.60	1.14	5.02	1.24	4.61	1.02	5.03	0.76
	Nerdland	6.30	0.80	5.78	1.11	4.09	1.10	5.31	0.78	5.03	1.11	5.30	0.57

rich questions can be transferred to local models by generating a data set to fine-tune these local models on. Finally, the combination of the image-to-text model with the GPT-3.5 generated training set, is able to achieve good performance on both visual reference and interestingness, specificness and politeness – combining the best of both worlds.

The lower performance of all models on the Nerdland data set can be explained by the different distribution of visual features. On one hand, the Nerdland data set contains less salient features, as they are daily-life snapshots that were not intended as a photo to be published. The Flickr data set contains more pictures of people wearing interesting outfits or doing interesting actions. On the other hand, there might also be features in the Nerdland data set that are not present in the Flickr images. The background of the Nerdland images, for example, often contains some clutter objects that are not related to the person and could also be misinterpreted by the captioning model or image-to-text models because of the lower image quality.

Nevertheless, the lower performance ratings are primarily due to the lower Visual reference score. The models are still able to ask interesting, specific, and polite questions, and still outperform human crowdworkers. The overall shift in performance is relatively small, meaning that the presented approaches are also feasible in real-world settings. Interestingly, the GPT-3.5 model also suffers from the transfer, while it was not pre-trained on the Flickr images.

The difference between the human performance on the Flickr and Nerdland data sets is remarkable as it is opposite to that of the models: crowdworkers performed better on the Nerdland data set. However, this is mostly a reflection of the difference in quality between MTurk and Prolific.

A further limitation of the models that were fine-tuned on the crowdsourced data, is the way that data set of questions was constructed. As both the language and vision-to-language models' capabilities to generate longer, more nuanced questions was not yet certain at the time of constructing the data set, we instructed the crowdworkers to write short questions. GPT-3.5 was not instructed to limit its output in such a way, and the raters seem to evaluate GPT-3.5's longer questions as more polite than the ones written by the crowdworkers. Furthermore, part of the crowdsourced data set was manually annotated by researchers, potentially lowering diversity in that data set.

6.2 Conclusion and Future Work

In this work, we presented a system for a situated AI system to start a social conversation with a user by asking a question that is grounded in visual input, such as the user's apparel. While research in vision-and-language models is growing, related work has not yet focused on social conversations for situated systems.

We compared multiple approaches to this problem in a crowdsourced evaluation study and discovered that current vision-and-language models are sufficiently mature for this problem when being fine-tuned. While image captions can be used as a representation of the visual information for text-to-text models, end-to-end image-to-text models outperform them in terms of correctly referring to visual features, even when being fine-tuned using the same number of training samples, answering our first research question.

As for our second question: we showed that large language models such as GPT-3.5 generate more interesting, specific and polite questions than models that are fine-tuned on questions written by crowdworkers, and even than crowdworkers themselves. Answering our final question, we saw that these qualities can be transferred to smaller, locally trained models, by fine-tuning them on questions that are generated by the large language model, removing the need for expensive human annotations, and showing that "synthetic" data can also be used to train social AI systems. These smaller models are also more controllable than models that require a (paid) API or GPU farms, and can guarantee stability and privacy.

End-to-end image-to-text models that are prompted in zero-shot fashion are not yet sufficiently mature to understand this task at the time of writing. Newly released large image-to-text-models such as GPT-4v [61] should be further tested for situated social dialogue tasks. However, combining all three research questions, we saw that the superior visual reference capacity of small end-to-end image-to-text models can be leveraged for this task and combined with the large language models' better expressive quality by fine-tuning the smaller image-to-text model on a data set generated by a large language model: this approach performed the best of all that were tested, even with a model that can be locally trained and run.

The effects of this conversation-starting system should be tested when embodied in a physical agent like a social robot, and it should be extended into a full conversation that

remains grounded in visual information. Future work will also have to look at how to further optimise language models for subjective criteria such as generating interesting and polite questions while maintaining correct visual grounding and not making the questions too specific, perhaps using more qualitative user feedback. The politeness and appropriateness of these questions is also highly culture-dependent, providing another avenue for future research to fine-tune such a model to different cultures.

Overall, our work has shown that the current vision-and-language technology offers a lot of promise for autonomous multi-modal social interactions with situated and embodied AI systems, which we believe will be a crucial stepping stone for stronger human-robot relationships, and we look forward to more research in this area.

REFERENCES

[1] J. Laver, "Linguistic routines and politeness in greeting and parting," in *Conversational routine*. De Gruyter Mouton, 2011, pp. 289–304.

[2] V. Žegarac, "What is "phatic communication"?" in *Current Issues in Relevance Theory*, V. Rouchota and A. Jucker, Eds. John Benjamins, 1998, pp. 327–362.

[3] H. J. Wiener, "Conversation pieces: The role of products in facilitating conversation," Ph.D. dissertation, Duke University, 2017.

[4] J. Laver, "Communicative functions of phatic communion," *Organization of behavior in face-to-face interaction*, vol. 215, p. 238, 1975.

[5] M. McAllister, B. Matarasso, B. Dixon, and C. Shepperd, "Conversation starters: re-examining and reconstructing first encounters within the therapeutic relationship," *Journal of psychiatric and mental health nursing*, vol. 11, no. 5, pp. 575–582, 2004.

[6] J. Chen, Y. Guo, and J. Duan, "How and when phatic communion enhances advice taking," *Asian Journal of Social Psychology*, vol. 25, no. 4, pp. 611–622, 2022.

[7] F. González Manzo, "Talking big about small talk: A contemporary theoretical model for phatic communication," Ph.D. dissertation, Mount Saint Vincent University, 2014.

[8] C. A. Cifuentes, M. J. Pinto, N. Céspedes, and M. Múnera, "Social robots in therapy and care," *Current Robotics Reports*, vol. 1, pp. 59–74, 2020.

[9] E. Verhelst, R. Janssens, T. Demeester, and T. Belpaeme, "Adaptive second language tutoring using generative ai and a social robot," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1080–1084.

[10] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 2, pp. 293–327, 2005.

[11] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu *et al.*, "Multimodal child-robot interaction: Building social bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, 2012.

[12] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: Personalisation promotes child learning," *PloS one*, vol. 12, no. 5, p. e0178126, 2017.

[13] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 687–694.

[14] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Šabanović, *Human-robot interaction: An introduction*. Cambridge University Press, 2020.

[15] R. Janssens, P. Wolfert, T. Demeester, and T. Belpaeme, "'cool glasses, where did you get them?" generating visually grounded conversation starters for human-robot dialogue," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 821–825.

[16] A. W. Siegman and S. Feldstein, *Nonverbal behavior and communication*. Psychology Press, 2014.

[17] K. Shubham, L. N. N. Venkatesan, D. B. Jayagopi, and R. Tumuluri, "Multimodal embodied conversational agents: A discussion of architectures, frameworks and modules for commercial applications," in *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2022, pp. 36–45.

[18] S. Roller, Y.-L. Boureau, J. Weston, A. Bordes, E. Dinan, A. Fan, D. Gunning, D. Ju, M. Li, S. Poff *et al.*, "Open-domain conversational agents: Current progress, open problems, and future directions," *arXiv preprint arXiv:2006.12442*, 2020.

[19] M. Cooney and W. Leister, "Using the engagement profile to design an engaging robotic teaching assistant for students," *Robotics*, vol. 8, no. 1, p. 21, 2019.

[20] D. F. Glas, K. Wada, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Personal greetings: Personalizing robot utterances based on novelty of observed behavior," *International Journal of Social Robotics*, vol. 9, pp. 181–198, 2017.

[21] G. Pantazopoulos, J. Bruyere, M. Nikandrou, T. Boissier, S. Hemanthage, B. K. Sachish, V. Shah, C. Dondrup, and O. Lemon, "Vica: Combining visual, social, and task-oriented conversational ai in a healthcare setting," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, ser. ICMI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 71–79. [Online]. Available: https://doi.org/10.1145/3462244.3479909

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training."

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[28] OpenAI, "Gpt-4 technical report," 2023.

[29] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao *et al.*, "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends® in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022.

[30] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 326–335.

[31] H. Alamri, V. Cartillier, A. Das, J. Wang, A. Cherian, I. Essa, D. Batra, T. K. Marks, C. Hori, P. Anderson *et al.*, "Audio visual scene-aware dialog," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7558–7567.

[32] K. Shuster, S. Humeau, A. Bordes, and J. Weston, "Engaging image chat: Modeling personality in grounded dialogue," *arXiv preprint arXiv:1811.00945*, 2018.

[33] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] K. Shuster, E. M. Smith, D. Ju, and J. Weston, "Multi-modal open-domain dialogue," *arXiv preprint arXiv:2010.01082*, 2020.

[35] J. Feng, Q. Sun, C. Xu, P. Zhao, Y. Yang, C. Tao, D. Zhao, and Q. Lin, "Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation," *arXiv preprint arXiv:2211.05719*, 2022.

[36] N. Mostafazadeh, I. Misra, J. Devlin, L. Zitnick, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," *CoRR*, vol. abs/1603.06059, 2016. [Online]. Available: http://arxiv.org/abs/1603.06059

[37] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende, "Image-grounded

conversations: Multimodal context for natural question and response generation," *CoRR*, vol. abs/1701.08251, 2017. [Online]. Available: http://arxiv.org/abs/1701.08251

[38] B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan, "Emotional dialogue generation using image-grounded language models," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[39] A. Sundar and L. Heck, "Multimodal conversational ai: A survey of datasets and approaches," in *Proceedings of the 4th Workshop on NLP for Conversational AI*, 2022, pp. 131–147.

[40] H. De Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela, "Talk the walk: Navigating new york city through grounded dialogue," *arXiv preprint arXiv:1807.03367*, 2018.

[41] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *arXiv preprint arXiv:2305.15021*, 2023.

[42] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations," *arXiv preprint arXiv:2104.08667*, 2021.

[43] H. Kamezawa, N. Nishida, N. Shimizu, T. Miyazaki, and H. Nakayama, "A visually-grounded first-person dialogue dataset with verbal and non-verbal responses," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3299–3310.

[44] C. Northcutt, S. Zha, S. Lovegrove, and R. Newcombe, "Egocom: A multi-person multi-modal egocentric communications dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[45] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[46] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[47] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[48] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[49] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[50] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[51] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[52] S. Kalkowski, C. Schulze, A. Dengel, and D. Borth, "Real-time analysis and visualization of the yfcc100m dataset," in *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, 2015, pp. 25–30.

[53] J. Wu, J. Wang, Z. Yang, Z. Gan, Z. Liu, J. Yuan, and L. Wang, "Grit: A generative region-to-text transformer for object understanding," *arXiv preprint arXiv:2212.00280*, 2022.

[54] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[55] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[56] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[57] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[58] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[59] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of experimental social psychology*, vol. 45, no. 4, pp. 867–872, 2009.

[60] B. D. Douglas, P. J. Ewell, and M. Brauer, "Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona," *Plos one*, vol. 18, no. 3, p. e0279720, 2023.

[61] G. A. Abbo and T. Belpaeme, "I was blind but now i see: Implementing vision-enabled dialogue in social robots," *arXiv preprint arXiv:2311.08957*, 2023.

Ruben Janssens is currently a PhD researcher in the IDLab-AIRO research group at Ghent University - imec, researching conversational artificial intelligence for human-robot interaction, focusing on multi-modal language models. He was main organiser of the 2022 Human-Robot Interaction Winter School on Embodied AI. He received his MSc degree in Computer Science Engineering with distinction from Ghent University, Belgium in 2021.

Pieter Wolfert is a lecturer at Radboud University and researcher at ConnectedCare. Previously, at Ghent University, he delved into co-speech gestures in embodied conversational agents. Affiliated with the Donders Institute, his current work explores the nuances of effective communication in interactive settings, emphasizing gestures in human-machine interactions.

Thomas Demeester is assistant professor at IDLab, at the Department of Information Technology, Ghent University-imec in Belgium. After his master's degree in electrical engineering (2005), he obtained his Ph.D. in computational electromagnetics. His research interests then shifted to information retrieval, natural language processing (NLP) and machine learning, and more recently to Neuro-Symbolic AI.

Tony Belpaeme is Professor at Ghent University and Visiting Professor in Robotics and Cognitive Systems at the University of Plymouth, UK. At Ghent he is a member of IDLab - imec. He received his PhD in Computer Science from the Vrije Universiteit Brussel (VUB) and currently leads a team studying cognitive robotics and human-robot interaction. Starting from the premise that intelligence is rooted in social interaction, Belpaeme and his research team try to further the science and technology behind artificial intelligence and social human-robot interaction.