Towards Neuro-Symbolic Approaches for Referring Expression Generation

Manar Ali[*1,2], **Marika Sarzotti**[*2,4], **Simeon Junker**[1,3],
Hendrik Buschmeier[1,2], **Sina Zarrieß**[1,3]
[1]CRC 1646 'Linguistic Creativity in Communication', Bielefeld University, Germany
[2]Digital Linguistics Lab, Bielefeld University, Germany
[3]Computational Linguistics Group, Bielefeld University, Germany
[4]Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy
{manar.ali|simeon.junker|hbuschme|sina.zarriess}@uni-bielefeld.de
marika.sarzotti@studenti.unitn.it

Abstract

Referring Expression Generation (REG) has a long-standing tradition in computational linguistics, and often aims to develop cognitively plausible models of language generation and dialogue modeling, in a multimodal context. Traditional approaches to reference have been mostly symbolic, recent ones have been mostly neural. Inspired by the recent interest in neuro-symbolic approaches in both language and vision, we revisit REG from these perspectives. We review relevant neuro-symbolic approaches to language generation on the one hand and vision on the other hand, exploring possible future directions for cognitively plausible models of reference generation/reference game modeling.

1 Introduction

Referring Expression Generation (REG) in visual scenarios is a traditional and widely studied task in cognitively motivated work on Natural Language Generation (NLG). At its core, the task consists of generating an expression that refers to a visual object in a given scene, in a way that an addressee can identify the intended target (Reiter and Dale, 2000). Although this task may seem basic and constrained at first, it is multifaceted and involves overcoming several implicit or explicit challenges at the intersection of language and vision. These challenges include segmenting and understanding the low-level visual input (*visual processing*), determining the properties of the referential target that distinguish it from all distractors (*content determination*), and, finally, formulating the conceptual information into well-formed linguistic expressions (*linguistic realization*), see Schüz et al. (2023).

Existing research in REG has approached this problem using two different methodologies, see Figure 1: The landscape can be roughly divided into *symbolic* and *neural* (or *visual*) approaches,

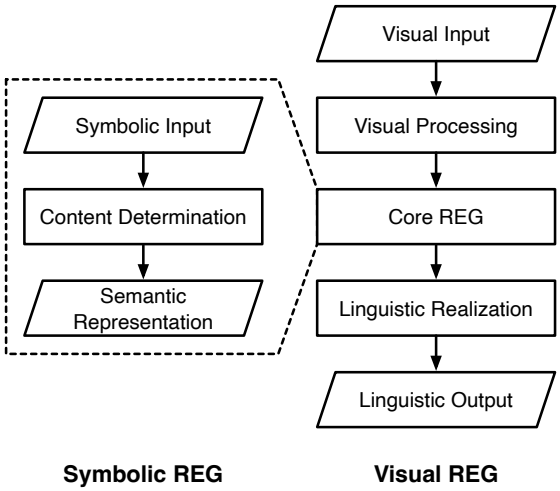


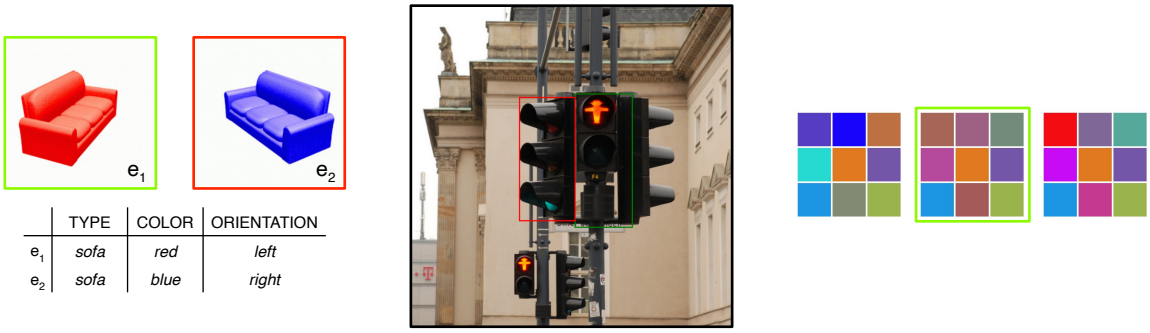
Figure 1: Conceptual illustration of processing steps in common models for symbolic and neural-visual Referring Expression Generation (REG; Schüz et al., 2023). While symbolic REG focuses primarily on selecting discriminative properties of the target, low-level inputs and natural language outputs require further processing steps or more general methods.

each with their own characteristics. Symbolic methods offer controllable, transparent, and cognitively plausible ways of pragmatic reasoning, but most approaches focus on specific challenges (i.e., content determination), and it is difficult to apply the algorithms to natural scenarios due to their dependence on symbolic inputs. In contrast, neural methods can be easily applied to more natural or complex scenarios, as the systems are trained end-to-end, implicitly learning all the necessary steps from visual processing to linguistic realization. However, neural approaches are notoriously difficult to control, their cognitive plausibility is debatable, and the exact processing methods are generally concealed due to the black-box nature of neural systems.

Against this background, neuro-symbolic approaches in computational linguistics and NLP are currently attracting considerable research inter-

[*]These authors share first authorship.

Proceedings of the 2025 CLASP Conference on Language models And RePresentations (LARP), pages 38–50



(a) Symbolic input (b) Natural visual input (c) Abstract visual input

Figure 2: Examples for different REG settings. Settings like (a) (van Deemter et al., 2006) have been traditionally addressed with symbolic approaches, whereas the settings in (b) (Kazemzadeh et al., 2014) and (c) (McDowell and Goodman, 2019) call for (partially) neural approaches, due to the lack of symbolic input representations.

est: By combining neural and symbolic processing methods, it becomes possible to build systems that retain the flexibility and performance of neural systems, but become more robust, controllable, and transparent (Hamilton et al., 2024).

In this paper, we review existing symbolic and neural approaches to REG and discuss how these lines of work can be integrated using neuro-symbolic approaches. We argue that symbolic processing methods can be applied to different stages in a neural REG processing pipeline, potentially leading to more transparent, cognitively plausible, and robust REG systems. What exactly is considered a neuro-symbolic system is not always consistently defined. Here, we treat approaches as neuro-symbolic that include neural modeling components, but also methods for reasoning about symbolic information.

2 Background

2.1 Referring Expression Generation

Symbolic REG Generating references to objects has been a long-standing field of interest in computational linguistics see Krahmer and van Deemter 2012 for a survey). Influential work (e.g., Dale 1989; Dale and Haddock 1991) started to focus on algorithms for *content selection*, comparing properties of a target object with potential distractors to determine a set of properties that can be formulated into discriminative descriptions. Building on a Gricean notion of pragmatics (Grice, 1975), algorithms are considered successful if they provide sufficient information to identify the intended referent without being overly informative. A prime example is the Incremental Algorithm (Dale and Reiter, 1995), which iterates through attributes in a defined order of preference, selecting those that rule out any distractors until only the target remains.

Subsequent work extended the scope by including, e.g., relational descriptions (Krahmer and Theune, 2002; Krahmer et al., 2003), references to sets of multiple targets (Horacek, 2004; Gatt and van Deemter, 2007), or notions of prominence or salience to pre-select contextually relevant distractors (Kelleher and Kruijff, 2006; Belz et al., 2010).

Much of the work in symbolic REG consists of deterministic, rule-based search algorithms for content determination that operate on symbolic knowledge bases (see Figure 2a). However, there are alternative approaches such as the probabilistic PRO (van Deemter et al., 2012) and RSA (Frank and Goodman, 2012) models or the graph-based algorithm in Krahmer et al. (2003). Further approaches combine content determination with linguistic realization (Horacek, 1997; Stone and Webber, 1998; Siddharthan and Copestake, 2004), see van Deemter (2016). However, the reliance on symbolic input information remains as a characteristic feature.

Neural REG Visual environments are commonly used as prime examples or application domains for symbolic REG algorithms, but the reliance on symbolic inputs largely prohibits direct application in natural visual scenarios where this information is not available (Schüz et al., 2023). In recent years, work on *visual REG* has reformulated the task as an *image-to-text* generation problem, enabled by corpora such as RefCOCO (Kazemzadeh et al. 2014; see Figure 2b) and more general advances in neural vision-and-language modeling. Here, the goal is to generate descriptions from raw visual representations of objects in natural images.

Similar to image captioning (Vinyals et al., 2015),

neural REG models are commonly trained end-to-end and follow the encoder–decoder scheme, where raw visual inputs are transformed into intermediate representations by an image encoder and then passed to a language decoder. Hence, neural approaches to visual REG differ fundamentally from their symbolic counterparts: Low-level perceptual inputs replace the high-level symbolic information, and while symbolic approaches often focus on content determination, neural systems cover all steps from visual processing to linguistic realization, although the exact processes are largely concealed in the connectionist structures of the neural systems.

Much of the existing work revolves around methods to optimize the discriminative power of generated expressions (see Schüz et al., 2023), for example by including different simulations of addressee behaviour (Mao et al., 2016; Luo and Shakhnarovich, 2017; Yu et al., 2017; Schüz and Zarrieß, 2021), enriching visual input representations with discriminative information (Yu et al., 2016; Liu et al., 2017) or directing systems to pragmatically relevant features (Li and Jiang, 2018; Tanaka et al., 2019; Liu et al., 2020; Kim et al., 2020; Sun et al., 2022). Other works focus on aspects beyond discriminativeness, e.g., iteratively refined expressions (Zarrieß and Schlangen, 2016; Ye et al., 2023), effects of decoding methods (Zarrieß and Schlangen, 2018), generating diverse expressions (Panagiaris et al., 2020, 2021), REG in visual dialogue (Willemsen and Skantze, 2024) or the role of visual scene context (Junker and Zarrieß, 2024). More recently, work on neural REG has started to incorporate vision-language models (Bracha et al., 2023; Guo et al., 2024; Liang et al., 2024) and referring expressions have been included in multitask frameworks (Wang et al., 2022b; Lu et al., 2023; You et al., 2023; Xiao et al., 2024), although with focus on the inverse referring expression comprehension task.

2.2 Neuro-symbolic approaches in general

Neuro-Symbolic AI is a growing research field concerned with the development of AI systems which should be able to simulate and integrate the two cognitive processes commonly considered as the core of intelligent behaviour, namely the ability to learn from experience and to reason on what has been learned (Valiant, 2003). Researchers have been trying to pursue this goal by combining neural networks and deep learning methods, excellent at handling parallel computation, unstructured data, and pattern recognition, with purely symbolic approaches, typically leveraging formal logic or structured representations, which are verifiable and data-efficient, and allow structured and logical reasoning about data and its patterns (Garcez and Lamb, 2023; Hamilton et al., 2024).

The problem of how neural networks can handle and represent symbolic knowledge has been present in the literature since early attempts to computationally model brain processes (Bader and Hitzler, 2005). Over the past decade, however, it was Deep Learning that got the most attention in research and application. Lately, it was argued that in order to achieve rich, semantically sound and explainable AI systems, research efforts should focus on the integration between methods affording reasoning abilities and Deep Learning (Garcez and Lamb, 2023), resulting in a new interest in neuro-symbolic integration. To clarify and systematize the work on neuro-symbolic integration, highlighting similarities and differences among the various contributions, taxonomies have been devised. The most well-known was proposed by Kautz (2022) and further streamlined in Hamilton et al. (2024):

Sequential Sequential architectures are current the dominant approach in Deep Learning when the input and output of neural networks are symbolic in nature, such as in the case of Natural Language Processing, where symbolic inputs, namely words and word sequences, are converted into vectors and processed by a neural network.

Nested Nested architectures are those that loosely couple a symbolic reasoning system, such as a problem solver or a planner, with a neural component that will guide certain decision processes. One instance is DeepMind's AlphaGo (Silver et al., 2016), where a Monte Carlo tree search algorithm is paired with a neural network tasked to evaluate game states and suggest moves.

Cooperative Cooperative architectures include a neural component which receives raw inputs, such as images' pixels, and converts them into symbolic data structures, for instance graphs or logic-based representations, which will be used by a symbolic reasoner. One example system is DeepProbLog (Manhaeve et al., 2018), which involves a neural network which parametrizes the truth distribution of predicates with respect to an input, and a probabilistic logic program for reasoning with them.

Compiled Compiled architectures are tightly coupled approaches, as there is no modular sub-division to handle learning and reasoning. In fact, these systems involve standard neural networks undergoing training regimes based on symbolic rules, by having knowledge compiled into the training set or the network's weights, or enforced via specific optimization functions. They are instantiated by Logic Neural Networks (Riegel et al., 2020), where symbolic rules are embedded directly into the architecture, as neurons in the network's layers represent specific logical operations, and Logic Tensor Networks (Badreddine et al., 2022), which are optimized to maximize the satisfiability of grounded (represented as real-valued tensors) formulas.

Ultimately, a neuro-symbolic system could have a fully integrated architecture where the symbolic reasoning component is embedded in the neural one. Hamilton et al. (2024) include this potential architecture in the *Nested* class, though, to this day, there are no implemented solutions that truly embody this definition.

3 Neuro-symbolic approaches to REG

Encoder-decoder models in vision-language generation tasks like REG always combine neural and symbolic aspects, as they map raw inputs (images) to symbolic outputs (text). However, in most approaches for visual REG (Section 2.1) the transformation from perceptual to symbolic information takes place at the very end of the processing pipeline and merely consists of a final mapping over the model's vocabulary during inference, without any reasoning processes involving those symbolic units. In this section we describe existing approaches for reference generation that go beyond this level of neuro-symbolic integration, and include further sources of symbolic information or symbolic reasoning processes.

Chamorro-Martínez et al. (2021) propose a system for referring expression generation (REG) that combines deep learning with symbolic processing. They use a Mask R-CNN model to segment images and detect objects with associated confidence scores. Fuzzy modeling is then applied to derive color attributes and spatial relationships between objects, which are represented in a graph structure—nodes represent objects with category and color labels, and edges represent spatial relations, all annotated with fuzzy confidence values. This symbolic graph is used by a content selection algorithm to identify the most discriminative properties for referring to each object.

Tsvilodub et al. (2024) present a neuro-symbolic Iterative Model (IM) for referring expression generation, inspired by the Incremental Algorithm (Dale and Reiter, 1995). The model combines large language models (LLMs) with symbolic reasoning. An LLM-based utterance proposer generates simple candidate descriptions, which a second LLM module evaluates for semantic adequacy. A symbolic contrastivity selector then assesses how well each description distinguishes the target from distractors. If no maximally contrastive expression is found, the process iterates by adding more details. Designed for visual tasks, the model avoids processing raw visual input by working with verbal scene descriptions.

In Junker and Zarrieß (2024), the low-level target representations used as input in their encoder-decoder models are supplemented by symbolic *scene summaries* that represent the relative area in the visual context covered by different types of objects, in order to support the robustness of referring expressions under visually challenging conditions. The results show that by including scene-level symbolic information, the models can correctly infer the type of the target object, even when visual representations of the target are severely distorted.

Apart from those works, the Rational Speech Acts framework (RSA; Frank and Goodman, 2012; Frank et al., 2016) emerges as the most prominent approach for integrating neural processing and symbolic reasoning. Here, generally, Bayesian inference is used to model pragmatic behaviour, in terms of rational speakers (S_1) that reason about how literal listeners (L_0) would understand utterances produced by literal speakers (S_0).

Andreas and Klein (2016) propose an approach for generating contrastive scene descriptions in a reference game involving visual scenes as targets and distractors. Ignoring distractor context, a neural language model acting as the literal speaker S_0 takes encoded images and produces descriptions of them. A neural literal listener L_0 takes an image description and a set of possible referents and produces a distribution over candidate scenes, for each indicating the probability that this scene is the referent described. Finally, a RSA reasoning speaker S_1 ties those models together by drawing a set of samples from S_0 and using Bayesian inference to select a description scored high by both S_0 and L_0. Similar to Tsvilodub et al. (2024), this

system relies on symbolic feature representations for objects depicted in the scenes.

In their work on pragmatically informative image captioning, Cohn-Gordon et al. (2018) follow the same intuition, but apply the pragmatic reasoning at each step of the iterative inference process. Here, S_0 is a character-level image captioning model, consisting of a CNN encoder and an LSTM decoder. At each decoding step, S_0 outputs a probability distribution over possible continuations of a partial caption consisting of the start token in the initial run. For each possible continuation, L_0 returns a distribution over potential target images. Finally, S_1 takes the L_0 distribution over images and re-weights the S_0 predictions for possible continuations by L_0's ability to infer the correct target image with this continuation.

The decoding algorithm in Vedantam et al. (2017) pursues the same idea, but with word-level captioning models and without the recursive back-and-forth between the speaker and listener agents as defined in the RSA model.

Several papers in REG have adopted the idea of performing pragmatic reasoning during the inference of otherwise context-agnostic generation models: Schüz and Zarrieß (2021) directly apply this approach to REG using the discriminative decoding methods from Cohn-Gordon et al. (2018) and Vedantam et al. (2017), but define targets and distractors as objects within a single image rather than as separate images. Here, at first, the bounding box content for a visual target object is encoded and passed to the model decoder. During decoding, output probabilities are compared at each step with the predictions of the same model when processing distractor objects instead of the target. On this basis, the token probabilities for the target are adjusted in favor of words that have a higher probability for the target than for distractors. In line with findings from image captioning (Schüz et al., 2021), the authors show that this method increases both the pragmatic informativeness and the linguistic diversity of generated expressions.

Zarrieß and Schlangen (2019) use a similar method to reason about possible categorizations of target objects, assuming that very specific terms should be avoided when models are uncertain about object categories. Again, they incorporate RSA-style reasoning into the iterative decoding process. However, their model does not reason about which words are informative for identifying the target, but about which terms should be used for the target to avoid erroneous descriptions, given the uncertainty about object categories. They show that their model generates more expressions without any nouns or category labels, consistent with the hypothesized strategies for describing unknown objects. With respect to an external listener model, the proposed strategy increases the resolution accuracy for most categories of objects.

Finally, White et al. (2020) consider further possibilities for how the Rational Speech Acts framework can be incorporated into neural generation models. In addition to a *full RSA* model, which includes an exhaustive reasoning process, where all possible utterances are tested for how effectively they allow the trained listener model to identify the target, they also consider a *sample re-rank* model which resembles Andreas and Klein (2016)'s approach in that a smaller number of candidate utterances are sampled from the speaker model and then re-ranked by the listener. In addition, they present a model that *amortizes* the computational costs of exhaustive RSA reasoning by directly optimizing a speaker model with respect to the utterances that an RSA model would prefer. To this end, during training, at each optimization step an utterance is sampled from the speaker model to be trained, which is then evaluated by the listener model and translated into training signals depending on its communicative success. This transfers the symbolic reasoning process from the inference to the training stage; the subsequent decoding process can thus be carried out using computationally more efficient methods. The results show that the amortized model almost achieves the pragmatic effectiveness of the full RSA model, but is significantly more efficient.

Overall, neuro-symbolic processing remains an exception in REG and related tasks. Apart from Junker and Zarrieß (2024), symbolic components generally target the linguistic level rather than the visual processing of inputs. Most commonly, RSA or related approaches are used to reason about the pragmatic informativeness of linguistic symbols (characters, words, or sentences), sometimes as part of the training procedure (White et al., 2020). Similar to content selection in symbolic REG, Chamorro-Martínez et al. (2021) and Tsvilodub et al. (2024) employ similar procedures at a more conceptual level, i.e., with regard to the question of which attributes best describe the referent, regardless of the concrete realization.

Regarding Hamilton et al. (2024)'s taxonomy, the approaches can be placed at different levels:

The addition of symbolic inputs renders Junker and Zarrieß (2024) a *sequential* system, while Tsvilodub et al. (2024) can be seen as *nested* with symbolic components controlling the entire process. Chamorro-Martínez et al. (2021) and inference-level RSA variants are *cooperative* because deep learning methods form the basis for symbolic reasoning. Finally, White et al. (2020)'s amortized model is a *compiled* system where symbolic reasoning is integrated into the training regime.

4 Neuro-symbolic NLG and Vision

Only a few approaches in REG surpass a level of neuro-symbolic integration that is trivial for vision-language generation tasks. This section will therefore discuss some neuro-symbolic approaches in two REG-relevant fields: NLG more generally and visual processing in vision-language tasks.

4.1 Natural Language Generation

Graph-based methods One approach is to integrate structured data representations, such as knowledge graphs, into the language generation process. This is often referred to as knowledge injection, where knowledge from external sources is incorporated into models to improve their output quality (Cadeddu et al., 2024). Knowledge graphs represent general-purpose, or domain-specific (Ji et al., 2022) data as nodes (entities) and edges (relations), a flexible and powerful way of encoding knowledge.

Knowledge graphs have been used in various NLG tasks (see Panchendrarajan and Zubiaga 2024 for a survey). In language modeling, knowledge graphs can be used by converting them into vector representations using graph embedding methods and feeding them as input to a language model. Other models adapt existing text-generation models to generate text directly from knowledge graphs (e.g., Koncel-Kedziorski et al., 2019). Knowledge graphs have also been used in dialogue systems. For instance, Zhang et al. (2020) proposed a method that constructs concept graphs from dialogue inputs and expands them to include related one-hop and two-hop concepts from a commonsense knowledge base. These graphs are then encoded into vector representations using a graph neural network. The resulting vectors are integrated with the original input to incorporate external knowledge and guide the model in generating coherent responses. Likewise, knowledge graphs have been used in text summarization tasks, where faithfulness to the original text is essential. Some models (e.g., Wang et al., 2022a) introduce a knowledge graph pipeline that extracts relational triplets from the source text and encodes them using graph embeddings. A filtering step uses a trained classifier to identify key facts from the source by predicting their importance. This allows the model to focus on salient and relevant information. The filtered knowledge graph embeddings are then combined with the hidden states from a BERT-based encoder and passed to the decoder.

Planning and constraint-guided generation Typical neural data-to-text models, which generate text from inputs like databases, often suffer from redundancy and lack factual faithfulness. Puduppully et al. (2019) proposed an alternative data-to-text approach where the input is a record table and the output is a natural language text. Their model explicitly separates content determination and content planning before passing the result to a neural generator for surface realization. The input is first encoded using a neural encoder. A content selection gate then determines relevant content using an attention mechanism over the table entries, followed by a sigmoid activation to determine which content is selected for further processing. Next, the content planning module decides what to say and in what order by generating a sequence of selected records using a pointer network. These plans are learned by aligning the summary text with table records. The resulting plan is then fed into a neural generator, which uses a standard encoder-decoder architecture to produce the output text.

Other data-to-text approaches include LogicNLG (Chen et al., 2020) and Symbolic Reasoning with Entity Scheduling (SORTIE; Zhao et al., 2023) which frame the task as logical data-to-text generation and aim to produce text that is logically consistent with the input data.

Lu et al. (2021) propose logic-guided, constraint-based generation that controls the decoding stage of neural text generation. It uses negative and positive constraints expressed as predicate logic formulas, which are converted into a penalty term and added to the decoding objective. This allows the model to generate fluent output while satisfying symbolic constraints, effectively guiding generation through inference-time decoding.

4.2 Vision

Graph-based methods In order to obtain agents which are able to proficiently understand the tem-

poral, relational and causal dynamics that go into performing everyday house-hold tasks, (Hazra et al., 2023) proposes a benchmark called Egocentric Task Verification (EgoTV), comprising a set of egocentric videos of daily life tasks, accompanied by a natural language description, as well as a novel Neuro-Symbolic Grounding (NSG) approach to counter the low performance exhibited by existing vision-language models on the EgoTV benchmark. The NSG architecture can convert a task description into a graph through its different components. This graph is then grounded in the video frames and the information represented by its nodes is aligned with the video. The NSG approach proposed by the authors indeed proved able to outperform state-of-the-art VLMs in capturing tasks' steps on both the EgoTV benchmark and on a dataset derived from the CrossTask dataset (Zhukov et al., 2019).

Huang et al. (2025) show interest in the understanding of spatio-temporal dynamics in videos as well. They introduce LASER, a neuro-symbolic framework that converts videos into graphs representing the properties and relations of entities at various time points. It then computes the alignment between these graphs and video captions that have been translated into formulas using an extended Linear Temporal Logic. The model is trained using weak supervision and displays enhanced performance compared to previous solutions in capturing relationships and dynamics in a range of video datasets with rich spatio-temporal specifications.

He et al. (2023) aim at applying scene graph generation to human-object interaction detection. They propose the unified model SG2HOI+ based on the Transformer architecture. The model is able to extract semantic-spatial features from images using a bounding box segmentation network, then generating scene graphs using information from said bounding boxes, and finally to convert the scene graphs into human-object interactions. SG2HOI+ was tested on a variety on benchmarks (Visual Genome, V-COCO, HICO-Det), and achieved better results than pre-existing methods.

Methods with programmatic descriptions

Gupta and Kembhavi (2023) X presents VisProg, a neuro-symbolic modular model that uses LLMs prompted in a few-shot manner to generate Python-like programs from image captions, questions and instructions. At each step, the programmes invoke one of the 20 modules currently supported (ranging from other LLMs to CLIP-like models and logic and arithmetic reasoning modules). VisProg performs at a high level on a range of of V&L tasks.

Hsu et al. (2023) introduce a modular architecture with neuro-symbolic components to solve 3D grounding tasks. It uses a language-to-code model to convert instructions in natural language asking to identify objects in pictures into symbolic programs. It then extracts object features and relations using an encoder, executes the program using the learned features and retrieves the target object.

Li et al. (2020) focus on jointly modeling camera poses, object locations and scene structures of naturalistic images presenting pronounced pattern regularities, treating the task as an inverse graphics problem, generating a graphic program from an input image, then reconstructing the picture and computing a loss between the reconstruction and the target.

Program-based neuro-symbolic approaches have also been applied to video related tasks. Kulal et al. (2021) introduces a framework designed to enhance human motion understanding in videos. It follows a hierarchical pipeline which first detects key points in videos, then produces both a concrete motion program, by assigning parameters to three motion primitives, and an abstract motion program, which generalizes over the concrete one by capturing higher-level repeated sequences and loops of primitives in the video.

5 Neuro-symbolic REG: Future directions

After reviewing approaches of neuro-symbolic integration methods in the area of REG and its closely related fields of Natural Language Generation and Vision (& Language), we will now discuss future directions to neuro-symbolic REG. In doing so, we will take into account the cognitively plausible properties of neuro-symbolic approaches – in particular, the combination of bottom-up, data-driven learning and structured reasoning about such data based on prior knowledge as two key aspects of human cognitive abilities and intelligent behaviour. In our discussion, we focus on the challenges in two different (but potentially overlapping) REG settings, i.e., interactive reference games set up as visual task-oriented dialogue, and reference generation under naturalistic real-world conditions.

Various neuro-symbolic approaches, previously discussed in Sections 3 and 4, seem fit to be adapted and applied to reference games, specifically *Cooperative* solutions (such as Chamorro-Martínez et al.,

2021; Huang et al., 2025; He et al., 2023), which are centered around the conversion of visual inputs into graphs representing their properties and relations, further used to solve tasks involving reasoning about those. Chamorro-Martínez et al. (2021) already applied such techniques, with an architecture able to generate referring expressions to objects in images employing fuzzy graphs.

Generating referring expressions in dialogical reference games, however, should consider the addressee and treat reference as a collaborative process (Clark and Wilkes-Gibbs, 1986), in which feedback is provided and the interaction history is available. While RSA-based models have a concept of a 'listener', it is an abstract one and not an actual addressee the model is interacting with and to whom the reference could be tailored. Adaptations to the addressee should happen on different levels of processing. Low level adaptations, such as interactive (lexical and syntactic) alignment (Pickering and Garrod, 2004) can be handled neurally during surface realization, whereas more strategic adaptations (Clark and Wilkes-Gibbs, 1986) could be the result of symbolic planning processes, which, however, could result from neural processing of the addressee's multimodal behaviors. If the addressee makes an error in, or is unable to resolve an initial reference, a model such as Chamorro-Martínez et al. (2021) could be adapted by implementing an iterative process, which retrieves the fuzzy graph previously produced and compares the target's node to that of the object wrongly selected by the addressee, in order to identify characteristics that were mistakenly chosen for or left out of the generated referring expression. The symbolic fuzzy graph is therefore a useful intermediate (and mediating) representation, that allows comparison of the speaker's and addressee's conceptualization with features of the target.

Nevertheless, more tightly integrated neuro-symbolic solutions such as those belonging to the *Compiled* class could be useful in modeling reference games, too. Methods including Logic Neural Networks (Riegel et al., 2020) and Logic Tensor Networks (Badreddine et al., 2022) could be used to extract features from images, encode these features as logical formulas, and then impose rules and constraints on them to guide the generation of referring expressions. Adressee's errors and feedback can be accounted for by updating the model weights or logic rules depending on whether the expression generated was precise enough for resolution.

A concern that can easily arise in REG tasks revolves around those cases where the exact category of an object that should be identified is not clear (Zarrieß and Schlangen, 2019). In such an eventuality, the knowledge injection methods discussed in Section 4.1 could potentially prove useful. In fact, through the use of knowledge graphs, it could be possible to provide REG models with knowledge bases granting them world knowledge, and thus informing them of more generic and overarching characteristics of objects that they might encounter in visual inputs, for instance common uses and functions, which could be used in indirect reference.

Cognitively oriented representations of the visual scenes in which referential targets are embedded also appear promising for generating referring expressions that are not only effective but also easy to understand. For example, Võ (2021) provides an in depth analysis of the rules and regularities of real world scenes, referred to as *Scene Grammars*. They include the notion of 'anchor objects', namely objects which are diagnostic of specific environments and serve as points of reference to identify other objects in a scene (e.g., the toilet in a bathroom). Inside visual scenes, objects tend to cluster around certain anchors, making it easier to identify them by restricting the domain of attention. The ability to take into consideration the pivotal role of anchors in object identification could be useful for REG models, as they would be able to abide to the natural way in which people parse visual scenes. *Cooperative* techniques based on graph structures, such as those presented in Section 4.2, could be optimized to recognize anchor nodes and subsequently use them to identify the target object in the scene, focusing on the relevant phrase and using information contained in it, such as the relationship the target has with the anchor, to construct referring expressions which can guide the listener's attention to the target in a way that is meaningful and familiar.

More generally, in vision and language tasks, such as referring expression generation, neuro-symbolic processing has great potential when providing both more autonomy as well as the ability for bidirectional information flow to components on all levels of processing. Neural vision components could implement theories of visual attention that respond to saliency and/or other features (e.g., Gestalts) from the visual scene such that they do not provide an exhaustive representation of the scene but are already selective or operate on a different

level of abstraction and thereby influence object naming or attribute selection when generating references with IA-like algorithms. Conversely, attention and visual processing could also be guided top down through symbolic information that is grounded in an interlocutor's utterance (such as a clarification), the broader interaction history, or the speaker agent's goal. Following theories of 'ecological perception' (Gibson, 1979), this could afford a neural re-conceptualization of objects in the scene, possibly yielding completely different features and object description that fit the speaker's need at a specific moment in a reference game.

6 Conclusion

Neuro-symbolic approaches are gaining considerable interest in computational linguistics and NLP, as they allow to integrate the complementary characteristics of symbolic and neural processing, potentially leading to strong and adaptive, but also transparent and cognitively plausible systems. In this paper, we reviewed existing neuro-symbolic approaches in REG and discussed possible future directions, drawing on related research areas such as NLG and vision. As an inherently multimodal task with defined pragmatic objectives, REG opens up many possibilities for linking these paradigms at different levels, opening up exciting possibilities for further research.

Acknowledgments

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, TX, USA.

Sebastian Bader and Pascal Hitzler. 2005. Dimensions of neural-symbolic integration – A structured survey. In Sergei Artemov, Howard Barringer, Artur d'Avila Garcez, Luis C. Lamb, and John Woods, editors, *We Will Show Them: Essays in Honour of Dov Gabbay*, pages 167–194. King's College Publications.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. Logic tensor networks. *Artificial Intelligence*, 303:103649.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2010. Generating referring expressions in context: The GREC task evaluation challenges. In Emiel Krahmer and Theune Mariët, editors, *Empirical Methods in Natural Language Generation*, pages 294–327. Springer, Berlin, Germany.

Lior Bracha, Eitan Shaar, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. 2023. DisCLIP: Open-vocabulary referring expression generation. In *Proceedings of the 34th British Machine Vision Conference*, Aberdeen, UK.

Andrea Cadeddu, Alessandro Chessa, Vincenzo De Leo, Gianni Fenu, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Salatino, and Luca Secchi. 2024. A comparative analysis of knowledge injection strategies for large language models in the scholarly domain. *Engineering Applications of Artificial Intelligence*, 133:108166.

Jesús Chamorro-Martínez, Nicolás Marín, Míriam Mengíbar-Rodríguez, Gustavo Rivas-Gervilla, and Daniel Sánchez. 2021. Referring expression generation from images via deep learning object extraction and fuzzy graphs. In *Proceedings of the 2021 IEEE International Conference on Fuzzy Systems*, pages 1–6, Luxembourg.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, Vancouver, Canada.

Robert Dale and Nicholas Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Michael C. Frank, Andrés Gómez Emilsson, Benjamin Peloquin, Noah D. Goodman, and Christopher Potts. 2016. Rational speech act models of pragmatic reasoning in reference games. *Preprint*, OSF:f9y6b.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336:998.

Artur d'Avila Garcez and Luís C. Lamb. 2023. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406.

Albert Gatt and Kees van Deemter. 2007. Lexical choice and conceptual perspective in the generation of plural referring expressions. *Journal of Logic, Language and Information*, 16(4):423–443.

James J. Gibson. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, MA.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, NY, USA.

Danfeng Guo, Sanchit Agarwal, Arpit Gupta, Jiun-Yu Kao, Emre Barut, Tagyoung Chung, Jing Huang, and Mohit Bansal. 2024. Prompting vision-language models for aspect-controlled generation of referring expressions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2793–2807, Mexico City, Mexico.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14953–14962, Vancouver, Canada.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2024. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, 15(4):1265–1306.

Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, and Ruta Desai. 2023. EgoTV: Egocentric task verification from natural language task descriptions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15371–15383, Paris, France.

Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. 2023. Toward a unified transformer-based framework for scene graph generation and human-object interaction detection. *IEEE Transactions on Image Processing*, 32:6274–6288.

Helmut Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 206–213, Madrid, Spain.

Helmut Horacek. 2004. On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 70–79, Brockenhurst, UK. Springer.

Joy Hsu, Jiayuan Mao, and Jiajun Wu. 2023. NS3D: Neuro-symbolic grounding of 3d objects and relations. In *In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, Vancouver, Canada.

Jiani Huang, Ziyang Li, Mayur Naik, and Ser-Nam Lim. 2025. LASER: A neuro-symbolic framework for learning spatial-temporal scene graphs with weak supervision. In *Procedings of the 13th International Conference on Learning Representations*, Singapore.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Simeon Junker and Sina Zarrieß. 2024. Resilience through scene context in visual referring expression generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 344–357, Tokyo, Japan.

Henry Kautz. 2022. The third AI summer: AAAI Robert S. Engelmore memorial lecture. *AI Magazine*, 43(1):105–125.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1952–1962, Barcelona, Spain (Online).

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. 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)*, pages 2284–2293, Minneapolis, MN, USA.

Emiel Krahmer and Mariet Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI, Stanford, CA, USA.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Sumith Kulal, Jiayuan Mao, Alex Aiken, and Jiajun Wu. 2021. Hierarchical motion understanding via motion programs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6564–6572, Los Alamitos, CA, USA.

Xiangyang Li and Shuqiang Jiang. 2018. Bundled object context for referring expressions. *IEEE Transactions on Multimedia*, 20(10):2749–2760.

Yikai Li, Jiayuan Mao, Xiuming Zhang, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2020. Perspective plane program induction from a single image. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4433–4442, Seattle, WA, USA.

Yaoyuan Liang, Zhuojun Cai, Jian Xu, Guanbo Huang, Yiran Wang, Xiao Liang, Jiahao Liu, Ziran Li, Jingang Wang, and Shao-Lun Huang. 2024. Unleashing region understanding in intermediate layers for MLLM-based referring expression generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 120578–120601.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy.

Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29:5244–5258.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2023. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (Un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134, Honolulu, HI, USA.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31, pages 1–11, Montréal, Canada.

Junhua Mao, J. Huang, A. Toshev, Oana-Maria Camburu, A. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, Las Vegas, NV, USA.

Bill McDowell and Noah D. Goodman. 2019. Learning from omission. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 619–628, Florence, Italy.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2020. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech & Language*, 68:101184.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing. *Expert Systems with Applications*, 251:124097.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conference and 9th AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 6908–6915, Honolulu, HI, USA.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Ryan Riegel, Alexander Gray, Francois Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh Srivastava. 2020. Logical neural networks. *Preprint*, arXiv:2006.13155.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online.

Simeon Schüz and Sina Zarrieß. 2021. Decoupling pragmatics: Discriminative decoding for referring expression generation. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, Gothenburg, Sweden.

Simeon Schüz, Albert Gatt, and Sina Zarrieß. 2023. Rethinking symbolic and visual context in referring expression generation. *Frontiers in Artificial Intelligence*, 6:1067125.

Advaith Siddharthan and Ann Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 407–414, Barcelona, Spain.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Matthew Stone and Bonnie Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 178–187, Niagara-on-the-Lake, Canada.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, 25:2446–2458.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. 2019. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5793–5802, Seoul, Korea.

Polina Tsvilodub, Michael Franke, and Fausto Carcassi. 2024. Cognitive modeling with scaffolded LLMs: A case study of referential expression generation. In *ICML 2024 Workshop on LLMs and Cognition*, Vienna, Austria.

Leslie G. Valiant. 2003. Three problems in computer science. *Journal of the ACM*, 50(1):96–99.

Kees van Deemter. 2016. *Computational Models of Referring: A Study in Cognitive Science*. The MIT Press, Cambridge, MA, USA.

Kees van Deemter, Albert Gatt, Roger P.G. van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2):166–183.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Natural Language Generation Conference*, pages 130–132, Sydney, Australia.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079, Honolulu, HI, USA.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA.

Melissa Le-Hoa Võ. 2021. The meaning and structure of scenes. *Vision Research*, 181:10–20.

Guan Wang, Weihua Li, Edmund Lai, and Jianhua Jiang. 2022a. KATSum: Knowledge-aware abstractive text summarization. In *Proceedings of the 2022 Principle and Practice of Data and Knowledge Acquisition Workshop PKAW*, Shanghai, China.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23318–23340.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, Virtual.

Bram Willemsen and Gabriel Skantze. 2024. Referring expression generation in visually grounded dialogue with discourse-aware comprehension guiding. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 453–469, Tokyo, Japan.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, Seattle, WA, USA.

Fulong Ye, Yuxing Long, Fangxiang Feng, and Xiaojie Wang. 2023. Whether you can locate or not? Interactive referring expression generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4697–4706, Ottawa, ON, Canada.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. FERRET: Refer and ground anything anywhere at any granularity. In *The 12th International Conference on Learning Representations*, Vienna, Austria.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham, Switzerland. Springer.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3521–3529, Honolulu, HI, USA.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 610–620, Berlin, Germany.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg, The Netherlands.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online.

Xueliang Zhao, Tingchen Fu, Lemao Liu, Lingpeng Kong, Shuming Shi, and Rui Yan. 2023. SORTIE: Dependency-aware symbolic reasoning for logical data-to-text generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11247–11266, Toronto, Canada.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3532–3540, Long Beach, CA, USA.