Building Common Ground in Dialogue: A Survey

Tatiana Anikina[a,*], **Alina Leippert**[a] **and Simon Ostermann**[a]

[a]German Research Center for Artificial Intelligence, Saarland Informatics Campus, Germany

Abstract. *Common ground* plays a crucial role in human communication since it helps to establish shared knowledge. However, common ground is also a heavily loaded term that may be interpreted in different ways depending on the context. The scope of common ground ranges from domain-specific and personal shared experiences to common sense knowledge. Representationally, common ground can be uni- or multi-modal, and static or dynamic.

In this survey, we attempt to systematize different facets of common ground in dialogue and position it within the current landscape of NLP research that often relies on the usage of language models (LMs) and task-specific short-term interactions. We outline different dimensions of common ground and describe modeling approaches for several grounding tasks, discuss issues caused by the lack of common ground in human-LM interactions, and suggest future research directions. This survey serves as a roadmap of what to pay attention to when equipping a dialogue system with grounding capabilities and provides a summary of current research on grounding in dialogue, categorizing 448 papers and compiling a list of the available datasets.

1 Introduction

Common ground [32, 155, 10] has been studied in a variety of settings by linguists, computer scientists, and philosophers alike. Common ground in dialogue can be defined as a set of shared beliefs between the interlocutors. However, as pointed out in Markowska et al. [116], a more complete definition should also include other components such as shared desires, intentions, and goals. From a Natural Language Processing (NLP) perspective, conversational grounding is important for building trustworthy dialogue systems that can reliably use shared knowledge in conversation [122]. Despite the widespread usage of chat-based language models (LMs), common ground is still often overlooked and not evaluated when comparing the performance of different models. The question arises: **Can we trust LMs and their generated outputs without building some common ground first?** In order to answer this question, it is necessary to clearly define and separate different dimensions of common ground and also to reflect on how these can be modeled and evaluated. To that end, the contributions of this survey are as follows:

(1) We describe different **dimensions of common ground** in dialogue that capture modality, type, and scope (Section 3);
(2) We survey **approaches towards modeling** common ground based on several grounding tasks (Section 4);
(3) We identify **potential problems** caused by the lack of common ground in LM-based dialogues (Section 5) and propose **future research directions** (Section 6).

* Corresponding Author. Email: tatiana.anikina@dfki.de

Figure 1. Word cloud based on the term frequency and abstracts of 448 papers mentioning the terms "common ground" and "dialog".

2 Methodology

In order to survey current research and also cover a variety of common ground definitions and modeling approaches we started by collecting a list of all papers published in the ACL Anthology since 2015 that mention "common ground" or "grounding" in their abstracts together with "dialogue", and then examined them in terms of the definition of common ground and which approaches were used to model it. We started with 448 papers and focused on those that address different dimensions of common ground relevant for dialogue processing. Specifically, we focused on the papers that discuss modality (e.g., textual, visual, multimodal grounding), type (static vs. dynamic), and scope of grounding (commonsense, domain or contextual knowledge). According to the classification of literature reviews outlined in Paré and Kitsiou [134], this survey can thus be considered as both *descriptive* and *narrative* since it aims to provide an overview of the available work and identify some trends for grounding in dialogue. Simultaneously, it is more focused on a qualitative interpretation of prior knowledge. The inclusion of papers in the survey was determined based on the relevance of their topics to both grounding and dialogue as reflected in the abstract and assessed by the authors (see Appendix for more detail on the methodology and statistics). Figure 1 shows a word cloud based on the word frequencies of the most common terms from the abstracts of 448 papers. Unsurprisingly, *dialogue*, *model*, and *knowledge* are the most frequent terms; although terms such as *generation*, *context*, *visual*, and *multimodal* are also commonly used, which reflects the challenging and multifaceted nature of common ground research.

3 Common Ground Dimensions

Common ground contributes to successful communication through dialogue. The concept is thereby easy to define in terms of its relevance, but what information is needed to form common ground is fleeting. As noted by Chandu et al. [19], there is often no such thing as an "axiomatic common ground". Successful grounding in

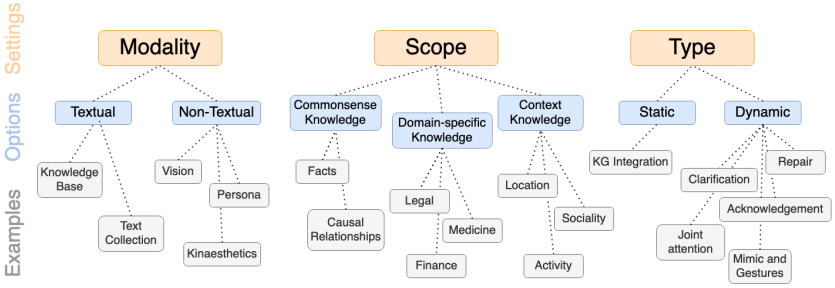

Figure 2. Common ground dimensions.

human-machine-communication is usually only evident in whether the goal of the task is reached; or through assessment of the quality of the conversation. While the common ground does consist of prior knowledge (such as world, commonsense or knowledge about previous events), much of the common ground of a conversation is built and established *during* communication. A system must thereby adapt to the evolving context of the conversation and the newly acquired knowledge.

There is not one way to establish a grounded dialogue between human and machine. As Chandu et al. [20] note, grounding is often performed with the goal of supporting a more defined end purpose task. How grounding is achieved heavily depends on the purpose of the conversation, e.g., whether the goal is to find a an object in a shared environment or to enrich a dull chit-chat with more interesting or personal, user-targeted facts. Researchers have proposed different methods for establishing common ground up to now. We aim to systematize the existing approaches and categorize them in terms of the following dimensions (see Figure 2):

1. The **modality** through which the conversation is grounded (e.g. textual, visual, multimodal)
2. The **scope** of the grounded information (e.g. commonsense, domain-specific or contextual knowledge)
3. The **type** of grounding (static or dynamic)

Building on this classification, we present a roadmap for incorporating common ground into dialogue systems.

3.1 Modality

Dialogue participants often integrate external contexts into the conversation, and these become part of the common ground [157]. Grounding can thereby connect the conversation to the environment: Grounding utterances in the real world allows models to account for what is missing or cannot be learned from conversational data. There exist many forms of external contexts. Strub et al. [157] for example mention the physical environment, a collaborative task the participants work on, a map they use for coordination or a database they want to access. Real world contexts that ground a conversation can thus be derived from different modalities, which Chandu et al. [20] classify into:

- **Textual modality**: e.g. plain text, entities/events, knowledge bases
- **Non-textual modality**: e.g. images, speech, videos

In recent years, many tasks that go beyond a single modality (in NLP: the textual one) have been proposed with the help of neural architectures [133]. Parcalabescu et al. [133] address the need for an appropriate definition of multimodality when the information receiver and processor is a machine learning system. The authors propose a *task-relative* definition: The task determines what information is relevant and how it can be stored, thereby indicating under which circumstances multiple modalities are necessary. Only in cases where different language representations (e.g. speech and image of a text) cannot be converted into one another without losing task-relevant information, they depict multiple modalities.

3.1.1 Grounding in the Textual Modality

Conversations between user and agent can be grounded in additional textual input that goes beyond the conversation history. This could be an external knowledge graph or other textual sources, providing world or domain knowledge.

Textual resources can be used to incorporate knowledge from the human world into the conversation between human and machine. As an example, Ghazvininejad et al. [55] model knowledge-grounded conversations with the goal to produce more contentful utterances grounded in the real world, i.e. taking into account not only the conversation history, but also external facts. To achieve this they retrieve various facts from textual sources such as Wikipedia and Foursquare, selecting the facts relevant to the conversation context. Similarly, language understanding can be improved by injecting commonsense knowledge into a conversation via knowledge graphs, providing background knowledge that machines otherwise lack as this information cannot be learned merely from conversational data [212, 144, 192, 135]. The additional knowledge can also come in the form of domain knowledge, as in Zhu et al. [216]. Focusing on the example of the music domain, their system uses structural background knowledge represented in the knowledge base to discuss and recommend songs to a user. Other works focusing on knowledge base integration and LM-based knowledge generation include, e.g. [23, 102, 105, 98].

3.1.2 Multimodal Grounding

Besides the textual modality, an increasing number of grounding-related tasks is multimodal, for example modeling an interplay between language and vision. Another possibility is to ground a dialogue in what is specific to the user, e.g. emotion or persona [49]. Pustejovsky and Krishnaswamy [139] argue that multimodal dialogue includes multiple aspects: (1) *Co-situatedness* and *co-perception* of the agents (i.e. how they perceive the environment and interpret the situation), (2) *Co-attention* of a shared situated reference (e.g. referring to the objects in the environment through language, gestures, visual clues), and (3) *Co-intent* of a shared goal which is especially relevant for problem-solving and collaborative tasks.

Grounding in Vision Possible multimodal tasks are visual games [157] or holding a dialogue about visual scenes [35], where shared representations help to ground meaning. Referring expression comprehension with visual features [65] is another important aspect of building common ground. A range of work where multimodal grounding is relevant focuses on dialogue applications. Such applications of visually-grounded systems can reach beyond interacting with a smart assistant, e.g. by helping visually impaired users to understand their surroundings or online content [22, 58]. They can also help to quickly gain an overview in search and rescue missions where an operator is 'situationally blind' but can interact via language [35] and contribute to grounding in the shared physical context [194].

Grounding in Persona More personalized and engaging conversation can be approached by grounding a dialogue in user-specific attributes such as persona or emotion [115, 180, 24, 197]. These attributes can be represented both through a textual and non-textual modality. A persona can be formed through a textual profile description of the user, including e.g. their personal interests and occupation [202]. Distributed representations for a persona can also be learned through conversational content such as a user's speaking style [95]. A multimodal approach for persona grounding is introduced by Ahn et al. [3], who enrich facts about user's personality with their pictures posted on social media and the corresponding comments.

Grounding in Kinaesthetics Another modality that can ground a dialogue by incorporation of the environment of user and system is kinaesthetics. The perception of one's own body position and movements is an important dimension for embodied systems in human-computer interaction [46]. Navigation agents that ground the conversation in the environment can help a user navigate through space [107], helping for instance persons with visual impairments, by retrieving landmark destinations and providing visual information to the user [104]. Communication grounded in spatial dimensions and actions can also be used to help a tourist reach a target location [38]. For such tasks, landmark recognition, user localization and natural language instructions are needed. Navigation instructions further require grounding in visual objects (e.g. "stop at the door") and geometric structure and directions (e.g. "turn left") [63].

Other modalities to ground the conversation in include e.g. gaze and nodding [5]. As early as in 1996, Dillenbourg et al. made the observation that the process of grounding is not bound inside one interaction mode but instead crosses different modalities. They suggest that agents should instead be capable of *modality-independent* grounding mechanisms, flexibly adjusting to a conversation's interaction-style.

3.2 Scope

Communicating on human terms requires more than just knowing the meanings of words; It demands a deep, integrated understanding of how, when, and why to use them. An important factor for successful communication is an "understanding of the shared world" [9]. Clark [31] propose to define different types of scope of common ground, namely **communal** common ground (speaking the same language, sharing a hobby or profession, leading to a communal lexicon of technical jargon and naming conventions) and **personal** common ground, i.e. joint (linguistic) experiences, leading to a private lexicon.

Based on our survey of recent publications and how they define the scope of common ground, we widen this classification and propose to distinguish between commonsense knowledge, domain-specific, and contextual knowledge.

Commonsense Knowledge. It is fundamental to everyday conversations and therefore plays a substantial role for the grounding process of a conversation. People use commonsense knowledge to understand and enrich conversations with related information or to picture what they do not understand. Moreover, commonsense knowledge enables reasoning about previously unseen events [150, 131]. Building a model grounded in commonsense knowledge is a challenging task, largely because there is no clear definition and there are certain aspects of commonsense knowledge that only come to light in the corresponding situations. As commonsense knowledge is rarely made explicit in natural conversation [36, 56], it is often not represented in conversational datasets and dialogue systems lack grounding in the real world [55].

Currently, no universally agreed upon strategy for encoding commonsense knowledge exists [145]. Commonsense knowledge can either be **implicit** in the training data or **explicit** in external knowledge sources. A common approach for adding explicit commonsense knowledge to a dialogue system is to harness an external KG, such as CONCEPTNET [154] or ATOMIC [150]. To enrich the conversation, the response generation is conditioned on the previous utterance and the related external facts (as in Young et al. [191] or Zhou et al. [212]). While grounding in external commonsense knowledge is helpful especially for concepts which are poorly represented in conversational training data [198], Davison et al. [36] point to the drawback that while commonsense knowledge bases do contain high-quality information, their coverage would be low. Richardson and Heck [145] observe a shift from grounding with KGs to using neural models for learning commonsense knowledge implicit in text data. Importantly, popular LMs produce natural sounding text but the responses may fail to integrate correct knowledge and the facts can be distorted [64]. Moreover, Bian et al. [12] observe that while GPT performs well on many commonsense benchmark tasks, it has drawbacks in domains which require a deeper understanding of human behaviour, such as social or temporal commonsense knowledge. Commonsense can not be learned from descriptions only, it requires both reasoning and inference abilities.

Domain Knowledge. Domain knowledge plays a crucial role in task-oriented dialogues and the required knowledge goes beyond commonsense or what can be inferred based on the conversational history, such knowledge varies a lot across different domains and often requires an access to some ontology or specialized knowledge base. For instance, when booking a hotel, it is important to know about the details like available room types, dates, and services. In the context of emergency response [60, 6], domain knowledge may include information about the responder roles, their responsibilities and equipment. The dialogue system should be able to correctly interpret and utilize the specialized terminology, e.g. it must know that UGV stands for an Unmanned Ground Vehicle in this domain.

There are different ways of integrating domain knowledge in dialogue systems apart from the direct fine-tuning on the domain data. Qian and Yu [140] propose a domain-adaptive dialogue generation approach called *DAML* (*Domain-Adaptive Meta-Learning*) that employs a two-step gradient update during training, allowing the dialogue system to capture general features across various tasks while enhancing its sensitivity to new domains, so that it can efficiently adapt with just a few training samples. Pryor et al. [138] employ a neural-symbolic approach to incorporate symbolic knowledge into the latent space of a neural model, effectively integrating domain knowledge and guiding the induction of dialogue structure. Suresh et al. [162] introduce a dialogue generation framework that can generate high-quality dialogue data for different domains using LMs and

Chain-Of-Thought approach [175].

Contextual Knowledge. Situational grounding requires linking the content of an utterance to its *meaning in the specific context* [139]. The context heavily determines the utterance's interpretation. In situated dialogue, where conversation partners share time and space, grounding can take the form of links to entities in this shared space [74]. However, context has many dimensions which go beyond just a shared space in which an utterance is voiced. Context carries memories of previous utterances, the background or purpose of the conversation, the interrelation and dynamics between the conversation partners, including social and emotional connotations [62].

A reliable and efficient conversational system should adapt to the former context dimensions, along with interactions users feel comfortable with in a specific context. Moreover, different communication styles can be preferred depending on culture, e.g. regarding the expressiveness of emotions, rhetorical style and directness [118].

Katayama et al. [73] define the following contextual dimensions: Location (e.g. home or public), Sociality (e.g. alone or group), Activity (e.g. walking or driving) and Emotion (e.g. neutral or happy). Whether a user feels excited or annoyed, is busy or has time for chit-chat, should receive consideration in finding a suitable conversation strategy, e.g. by eliciting a more discreet as opposed to an entertaining continuation [101]. Kola et al. [84] encourage situation-awareness in agent development such that "agents should provide support that is consistent with the user's goals and preferences", taking into account *situation cues* and *social relationship features*. The four context dimensions proposed by Katayama et al. [73] provide a starting point for assessing a system's context considerations and can be expanded according to task and goal.

3.3 Type

Regarding the type of grounding, we distinguish between **static-symbolic grounding** and **dynamic-collaborative grounding**, following findings in literature including [90, 11, 19] (Table 1).

- **Static-symbolic grounding**: The common ground is the ground truth external data, e.g. a KG or the shared perceptual environment.
- **Dynamic-collaborative grounding**: The common ground is formed interactively, e.g. through clarification and negotiation between user and agent.

A static-symbolic approach is used in Ji et al. [68], who use knowledge graph grounding to reduce hallucinated responses. What is missing from static-symbolic approaches, as emphasized by Benotti and Blackburn [11], is the aspect of *error recovery* through negotiation of meaning, which becomes relevant in dynamic-collaborative grounding. In the dataset *GrounDialog*, Zhang et al. [205] focus on dialogues where participants are provided with dissenting information. The naturally arising need to negotiate and clarify therein automatically leads to dynamic-collaborative grounding. Grounding success in a dynamic setting depends on effectively communicating the mutually shared information until a common ground between user and agent is established, while in static grounding it relates to the ability of the agent to successfully link the query to the data [19].

4 Modeling Approaches

In this section, we provide an overview of recent modeling approaches defined by several common grounding tasks: knowledge (static and dynamic), vision, and persona grounding. We acknowledge that this categorization is not exhaustive, and there are more grounding-related tasks that can be considered (e.g., kinaesthetics and multimodal grounding). However, for the purpose of this survey, we focus on three distinct categories of tasks that are prominent in the current scientific literature on grounding.

Table 1. Static vs. dynamic types of grounding.

	Static-symbolic	Dynamic-collaborative	Grounding Motivation
Larsson [90]	*Symbol Grounding:* Connect symbols (e.g. words) to world via perception.	*Communicative Grounding:* Interactively update CG in dialogue	Speakers need to converge on shared meaning.
Chandu et al. [19]	*Static Grounding:* CG is the external data, assuming its universality.	*Dynamic Grounding:* CG is built via interaction and clarification.	Axiomatic CG does not exist and needs to be established in real world.
Benotti and Blackburn [11]	*Symbol Grounding:* Link symbols with perception, e.g. language grounded in vision.	*Collaborative Grounding:* Reach mutual understanding incrementally through dialog.	Human perception is unstable and depends on memory, capabilities, perspective.

4.1 Knowledge Grounding Tasks

Knowledge-based grounding can be based on static or dynamic knowledge with a single or multiple sources of knowledge that need to be integrated for successful communication.

4.1.1 Static Knowledge Grounding

Static knowledge integration typically involves external knowledge bases, graphs, or document collections. For instance, Zhao et al. [209] propose a model (*KnowledGPT*) that uses a knowledge selection module and **jointly optimizes selection and response generation**. KnowledGPT consists of a context-aware encoder and a knowledge selector, trained with a policy-gradient method and a curriculum step that distinguishes between the "hard" and "easy" materials for grounding.

Feng et al. [45] propose the MultiDoc2Dial task and a dataset for modeling **goal-oriented dialogues that are grounded in multiple documents**. The task is to identify which parts of which documents are relevant at each dialogue turn. MultiDoc2Dial task focuses on (1) extracting the grounding span from the document collection and (2) generating the dialogue response given the history and the extracted spans.

Wu et al. [182] address knowledge-grounded dialogue generation with their *Section-Aware Commonsense Knowledge-Grounded Dialogue Generation with Pre-trained Language Model* (*SAKDP*). SAKDP utilizes a PriorRanking network with contrastive learning to estimate the **relevance of the retrieved knowledge facts**. All candidates are clustered into three groups according to their priority. SAKDP then uses section-aware strategies to encode knowledge in a linearized way and applies LMs to encode only the high-priority facts, thus making the encoding process more efficient. Another system called *PLUG* [97] **unifies different knowledge sources** for knowledge-grounded dialogue generation. The approach retrieves relevant information from various sources (e.g. wiki, dictionary, knowledge graph), converts the extracted knowledge into textual format and combines it with the dialogue history.

Chen et al. [26] focus on the task of **knowledge grounded dialogue generation with in-context learning**. Their goal is to produce faithful and informative responses that rely on the dialogue history

as well as the knowledge base. To this end, they propose a retrieval-based framework, *IKA* (In-context Knowledge grounded dialogue Augmenter), combining in-context learning with retrieval techniques and adding the most relevant and diverse demonstrations to the LLM prompt for response generation.

As a part of static grounding, Xie et al. [186] consider **structured knowledge grounding** (*SKG*) and propose the *UnifiedSKG* framework that can standardize different task representations (e.g. semantic parsing, question answering, fact verification). The main idea behind UnifiedSKG is to unify different forms of structured knowledge through linearization. Xie et al. [186] also show that task-specific knowledge can be effectively shared via multi-task prefix tuning, improving the overall performance on the target task.

Another direction for static grounding is to ground the conversation in **social media interactions**. Choudhary and Kawahara [30] emphasize that most of the current work on knowledge-grounded dialogue focuses either on persona or fact-based structured knowledge. Thus, they propose a different approach and present a system that can mimic human responses through modeling social media interactions by training a joint retriever-generator on a mixture of open-domain dialogue data and a collection of Reddit comments.

Other recent work that incorporates static knowledge into dialogue processing uses knowledge graphs and performs entity-agnostic representation learning [214], generates dialogue acts to guide generation through tree-structured reasoning [110], focuses on document-grounded conversations, and uses graphs to capture the inter- and intra-document relations [188].

4.1.2 Dynamic Knowledge Grounding

Dynamic knowledge grounding happens when common ground is formed interactively. This is often achieved through negotiation and clarification [179, 130, 113]. Dynamic changes in common ground can be also modeled as knowledge updates. For instance, Tuan et al. [168] introduce the task of **dynamic knowledge-grounded conversation generation**. They pair every dialogue turn with a knowledge graph that includes a collection of triplets representing entities and relations between them (e.g. *"x IsEnemyOf y"*). The grounding task in this setting involves (1) text generation conditioned on the textual input plus the corresponding knowledge graph and (2) generation of relevant entities after each update of the graph.

Tuan et al. [168] propose a model (*Qadpt*) that predicts the knowledge graph entities and retrieves the relational paths in the graph by applying multi-hop reasoning. Qadpt proves to be beneficial even for zero-shot adaptation with dynamic knowledge graphs. Similarly, topic-grounded dialogues also require keeping track of **topic transitions throughout a conversation**. Wen et al. [178] present a model called Sequential Global Topic Attention (*SGTA*). It uses a latent space to integrate the global-level and sequence-level information and predicts the topic based on the distribution sampling. SGTA exploits topic co-occurrences and models post-to-response topic transitions as well as predicts the next likely topic in dialogue.

Udagawa and Aizawa [170] focus on creating and maintaining common ground in **dynamic environments**. Specifically, they collect a dataset of 5,617 dialogues (*OneCommon* Corpus) that represents entity attributes and their temporal dynamics based on continuous values that correspond to entity movements. Udagawa and Aizawa [170] consider a collaborative reference task as a multi-agent cooperative game. Each agent can observe several entities and exchange information about them with other agents. The task is accomplished successfully if all the agents select the same entity at the end of the game. The proposed model encodes dialogue utterances and utilizes spatial and temporal encoders to integrate the dynamic features.

4.2 Vision Grounding Tasks

Vision grounding is crucial for conversations that revolve around the content of images or videos and there are tasks such as visual dialogue generation and image grounded question answering. Many multi-modal extensions of Transformer models (e.g., VL-BERT [158], VideoBERT [159], LXMERT [165], MTN [93], GTR [18], TransVG++ [40]) allow modeling both texts and images simultaneously and can be applied to such tasks. Below we exemplify several visual grounding tasks and showcase some models for image grounded conversations, visual and video-centered dialogue.

Mostafazadeh et al. [127] introduce the task of multi-modal **image grounded conversations** where natural-sounding conversations are generated about some shared image. This task has both elements of chit-chat and goal-oriented dialogue since the image constrains the topic of conversation.

Kang et al. [72] investigate the task of **reference resolution in visual dialogue**. The goal of this task is to answer a series of questions grounded in some image given the visual input together with the dialogue history. The authors propose Dual Attention Networks (*DANs*) to perform visual reference resolution. Their model consists of two attention modules: *REFER* and *FIND*. First, REFER applies multi-head attention mechanism to learn the relations between the question and the dialogue history. Next, FIND receives as input both image features and the outputs of the REFER module and combines them to perform visual grounding.

Kim et al. [77] address the **visual dialogue grounding task** in the context of question answering. They find that some questions can be answered by only looking at the image while others require both image and dialogue history. Therefore, they decide to maintain both models (image-only and image-history) and combine them in different ways. Specifically, they experiment with ensembling and consensus dropout fusion with shared parameters. The combined model demonstrates complementary gains for image-only and image-history models.

Video-grounded dialogue is also explored in [136, 92, 174]. The video grounding task involves modeling **video features across both spatial and temporal dimensions** as well as dialogue features that include dialogue history and interactions between the turns. Le and Hoi [92] extend GPT-2 models and formulate a video-grounded dialogue task as a sequence-to-sequence processing that combines both visual and textual representations. Their proposed model *VGD-GPT2* captures dependencies between different modalities at the spatio-temporal level (for videos) and token-sentence level (for dialogues).

Qin et al. [141] approach the task of **answering video-grounded questions in dialogue** using a Dual Temporal Grounding-enhanced Video Dialog model (*DTGVD*) that predicts turn-specific temporal regions while filtering out irrelevant video content and grounding free-text response in both video frame and dialog history. The proposed approach is based on the UniVL [108] visual-language model. DTGVD finds temporal segments in the dialog as well as contextually relevant video segments and grounds the response generation in both. This approach employs contrastive learning by utilizing grounded turn-clip pairs as positive samples and other turn-clip pairs as negative ones. The model is trained with answer generation and contrastive losses and achieves state-of-the-art results on several benchmark datasets for video-grounded dialogues.

Wang et al. [174] introduce a new Video-grounded Scene&Topic AwaRe dialogue (*VSTAR*) dataset and propose benchmarks for dialogue understanding based on **scene and topic segmentation as well as video-grounded dialogue generation**. Their experiments demonstrate that visual information is very important for the topic boundary detection and including such information can improve the performance by 7.1% F1. They also show that segment information is helpful for dialogue generation and the current encoder-decoder models still struggle to make full use of the visual input for the video-grounded dialogue generation.

4.3 Persona Grounding Tasks

Lim et al. [100] emphasize that it is important to ground **external knowledge and persona** simultaneously and propose a model called *INFO* that grounds persona information together with the external knowledge. They implement the knowledge and persona selector for the grounding task using poly-encoder and adopt retrieval-augmented generation to reduce the hallucinations and generate more coherent and engaging responses.

Majumder et al. [115] explore persona-grounded dialogue and focus on inferring simple implications of persona descriptions. For instance, if someone likes hiking, they probably also like nature. Majumder et al. [115] utilize commonsense knowledge bases to expand the set of persona descriptions. They also experiment with the **fine-grained persona grounding**, so that the model has to choose between different persona sentences when generating a dialogue response. To this end, the model uses variational learning to sample from various persona descriptions achieving good scores on the Persona-Chat dataset [203] with consistent and diverse responses.

Wang et al. [173] introduce a framework for **decoupling knowledge grounding** into different sources to aid response generation (e.g. persona, documents, memory). Their framework (*SAFARI*) can make a decision of whether to include a specific knowledge source and when to do so while generating a response. *SAFARI* has three modules that are responsible for planning, retrieval and assembling of information. Wang et al. [173] also construct a personalized knowledge-grounded dialogue dataset (*Knowledge Behind Persona*) where responses are conditioned on multiple knowledge sources for more informative and persona-consistent dialogues.

Gao et al. [51] focus on conversational agents that can infer listener's personas to generate appropriate responses and maintain consistent speaker profiles. They introduce *PeaCoK*, **a large-scale persona commonsense knowledge graph** with 100K human-validated facts, structured around five persona dimensions: characteristics, routines and habits, goals and plans, experiences, and relationships. *PeaCoK* is built by extracting and generating persona knowledge from existing knowledge graphs (ATOMIC [149], COMET [15]) and LMs.

Other related work includes modeling partner personas in dialogue [106], disentangling and recombining persona-related and persona-agnostic parts of the dialogue response [180] as well as personalized conversation generation based on journal entries [132].

5 Common Ground in LMs

Effective conversation requires building common ground that is ideally multimodal, dynamic, integrates commonsense and contextual knowledge. However, in the modern age of large LMs (LLMs) common ground is typically **defined by the user** or it is based on the **static training data**. For instance, prompts can be used to include persona, domain and task information, but this information is tailored in a way that represents the user's point of view. Moreover, inputs to LLMs are usually based on text even when they reference other modalities (e.g. describing location, time or emotions).

LLM outputs may also contain hallucinations or incorrect assumptions. Jiang et al. [69] distinguish between two potential sources of factual hallucinations: **insufficient knowledge** within the model's parameters and knowledge memorization coupled with the **lack of generalization** capability. However, some hallucinations can also be a product of miscommunication caused by the lack of common ground: Shaikh et al. [152] show that LLMs tend to generate text with less conversational grounding (on average, 77.5% less likely compared to humans) and often presume common ground instead of building it incrementally over time.

Language models often exhibit reduced rates of grounding acts and show **poor grounding agreement with humans**. Existing supervised fine-tuning and preference optimization datasets are potential sources of this problem [152], because such datasets are meant for training models that simply "follow" instructions based on a limited number of interactions. Models trained on such data never learn how to build common ground throughout the conversation and adjust it depending on new inputs.

Jokinen [70] also emphasize the need for grounding in LM-based dialogue systems and argue that such systems need to build a shared understanding of dialogue context and intents, grounding generated utterances in real-world events and potentially bridging the gap between neural and symbolic computing. Furthermore, Schneider et al. [151] benchmark both open- and closed-source LMs on the grounding tasks and find that both are equally good at classifying grounding acts but **identifying grounded knowledge** proved to be very challenging and it is better handled by close-source LMs.

Another important issue is that LLM-based conversational agents may fail to generate safe and appropriate responses [39] and often go along with a problematic user input, generating offensive and toxic language. Kim et al. [82] aim to address this issue by proposing *GrounDial* that **grounds dialog responses in commonsense social rules** and does not require any additional fine-tuning. GrounDial combines in-context learning with guided decoding that follows human norms to generate more safe and appropriate responses.

Yu et al. [192] emphasize that commonly used techniques for dialogue response generation are based on **Chain-of-Thought (CoT)** [175] or **Retrieval Augmented Generation (RAG)** [208]. However, both methods have important drawbacks. On the one hand, CoT may overestimate the capabilities of LLMs by treating them as isolated knowledge sources, while the knowledge stored in LLMs can be outdated, and LLMs are prone to hallucinations [67]. On the other hand, approaches like RAG underplay the internalized knowledge of LLMs and mostly rely on external sources. Yu et al. [192] propose a different approach that considers **LLM as a collaborator** and includes several Thought-then-Generate stages to identify knowledge demands and then find relevant information via Demands-Guided Knowledge Retrieval.

Chiu et al. [28] draw attention to another limitation of LLMs in the context of grounded task-oriented dialogue: LLMs are difficult to steer towards **task objectives** and they have difficulties with handling **novel grounding**. To address these limitations Chiu et al. [28] propose an interpretable grounded dialogue system that combines **LLMs with a symbolic planner** to perform grounded code execution and response generation. The proposed system has a reader that uses LLM to convert utterances into executable code functions which represent the core meaning and map language to symbolic ac-

tions, and a symbolic planner that can plan over the symbolic actions and determine the next response. The task progress is tracked via Bayesian reasoning and information gain objective. This approach achieved promising results on the OneCommon task from [169] that involves collaborative reference resolution.

Factual inconsistency of LLMs [148] is another important concern for building a conversational agent. Previous work has shown that LLMs can generate factually incorrect responses even when provided with valid knowledge sources [67]. Xue et al. [189] tackle the inconsistency issue in knowledge-grounded conversations by enhancing the factual knowledge expression via extended Feed Forward Networks (FFN) in Transformers and apply reinforcement learning that implicitly aligns dialogue responses with gold knowledge using factual consistency preference. The parameters in extended FFNs are updated based on the knowledge-related tokens that appear in both grounding knowledge and response, e.g., if the term "Argentina" is part of the response it can be enhanced with the factual knowledge "Argentina won the 2022 World Cup champion."

Daheim et al. [33] also investigate **factuality in the context of document-grounded dialogue generation** and consider two components: a simple ungrounded response generation model that encourages fluency, and a component that encourages responses which can help reconstruct the response grounding document. The proposed method uses Bayes' Theorem to decompose the posterior distribution of a response given context and grounding into these components, and employs scaling factors to promote either greater correctness or fluency of the response. Although this approach results in improved factuality, the online decoding is **computationally demanding**.

Mohapatra et al. [123] provide a benchmark and design various tests to assess how well LLMs handle grounding as both speakers and listeners. As listeners, LLMs should integrate repaired or canceled information and identify ambiguous cases that need clarification. As speakers, they must generate accurate and unambiguous responses. Mohapatra et al. [123] analyze perplexity on the responses that are appropriate and grounded and responses that are fitting but contextually incorrect. They found **a strong correlation between conversational abilities and the size of the models and the pre-training data** with larger models and datasets leading to improved grounding capabilities and lower perplexity for correct responses.

6 Current and Future Research Directions

In order to visualize the distribution of topics that are most prominent in the current research on dialogue grounding, we applied k-means clustering to the abstracts of 448 papers selected for the survey based on keyword match. We then categorize them according to common topics. Figure 4 shows that almost 16% of all papers focus on grounded generation and selection, while 15.2% of the papers are concerned with cognitive and human-centered aspects of grounding. Knowledge grounding and LLM fine-tuning are two other relatively big topics (13% each). Interestingly, multimodal papers represent only 7.4% with a similar number of papers dedicated to learning-based approaches. Multimodal papers with emphasis on spatial grounding are even less prominent (only 5.37%), and, quite worryingly, only 2% of all publications focus on evaluation and benchmarking. These trends show that there are some considerable gaps in the current research on conversational grounding.

As demonstrated in this survey, well-grounded dialogue requires some commonsense knowledge as well as contextual knowledge that heavily depends on the dialogue history and previous interactions, and it may also include some domain-specific knowledge. We

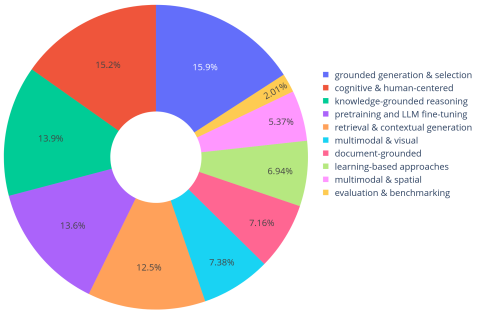

Figure 3. Distribution of topics relevant to conversational grounding based on the papers published in the ACL Anthology since 2015.

believe that the **interdisciplinary collaboration** between different fields such as robotics, cognitive science and linguistics will greatly benefit this research area. There is also a need for more **diverse and realistic datasets** that are not purely text-based or combine text with just one modality e.g. images, the community needs a collection of various types of data that capture different dimensions of grounding with respect to modality, type and scope. More often than not research on grounding addresses the static setting which assumes that we have an access to some graph or a knowledge base, but grounding is a collaborative process and more research needs to be done on incorporating **dynamic features**.

Another important aspect that has only recently started gaining more attention is **evaluation**. For instance, Alghisi et al. [4] investigate the impact of incorporating external knowledge to ground dialogues with retrieval-augmented generation or gold knowledge and emphasize the importance of human evaluation. Chaudhary et al. [21] find that LLM-based evaluation does not align well with human judgments and show that evaluation results are not robust against perturbations. Ghaddar et al. [54] argue that knowledge-grounded dialogue needs to be more thoroughly evaluated with respect to hallucinations.

An important future avenue is research comparing common ground in **human-human and human-robot interactions**. For both, there is a need to collect more diverse data and combine LLM-based generation with neuro-symbolic approaches. E.g., Bonial et al. [13] use an Abstract Meaning Representation formalism to ground language concepts in the robot's world model, and Torres-Foncesca et al. [167] investigate an important dimension of knowledge grounding related to object permanence, i.e., the ability to maintain mental representations of objects even when they are not in view.

More research should be done towards **integrating common ground during pre-training** and fine-tuning stages of LLMs, e.g., by introducing additional loss functions and contrastive learning to distinguish between compatible and incompatible beliefs being formed in a conversation (or by measuring perplexity between correct and adversarial responses [123] and combining generation with grounding reconstruction [33]). Some recent works explore how one can build a knowledge-grounded dialog system that utilizes both dialog history and local knowledge base for response generation with a semi-supervised pre-training [196] and perform large-scale multi-party aware pre-training on conversational data [8] that shows promising results for knowledge grounded conversations.

7 Conclusion

In this survey we provided an overview of different dimensions of common ground and categorized them according to the modality, type and scope. We also discussed existing modeling approaches for

knowledge-based, visual, and persona-based grounding, exemplifying promising research directions and attempts to integrate various aspects of common ground. We talked about leveraging common ground in LLMs, and summarized the issues related to the lack of conversational grounding in such models. We also described current and promising future research directions. We hope that this survey and our annotations[1] will serve as a guide for exploring the broad and dynamic landscape of conversational grounding.

Limitations

The current survey represents just a snapshot of the research on the topics of conversational grounding. This is an interdisciplinary field that ideally involves collaboration between the researchers who work on language, vision, robotics, and cognitive modeling. This work may not include all the relevant and very recent publications due to its scope and focus on the ACL Anthology.

Acknowledgements

This project was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

[1] S. Agarwal, T. Bui, J.-Y. Lee, I. Konstas, and V. Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.728. URL https://aclanthology.org/2020.acl-main.728.

[2] J. Ahn, Y. Song, S. Yun, and G. Kim. MPCHAT: Towards multimodal persona-grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.189. URL https://aclanthology.org/2023.acl-long.189/.

[3] J. Ahn, Y. Song, S. Yun, and G. Kim. MPCHAT: Towards Multimodal Persona-Grounded Conversation. *arXiv preprint arXiv:2305.17388*, 2023.

[4] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, and G. Riccardi. Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue. In S. Mahamood, N. L. Minh, and D. Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 180–197, Tokyo, Japan, Sept. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.inlg-main.15/.

[5] M. Alikhani and M. Stone. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15, 2020.

[6] T. Anikina. Towards efficient dialogue processing in the emergency response domain. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.31. URL https://aclanthology.org/2023.acl-srw.31/.

[7] R. Artstein, J. Boberg, A. Gainer, J. Gratch, E. Johnson, A. Leuski, G. Lucas, and D. Traum. The Niki and Julie Corpus: Collaborative multimodal dialogues between humans, robots, and virtual agents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[1] Please refer to the Appendix for additional statistics, annotated papers and datasets. All materials will be made publicly available upon acceptance.

[8] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Wu, Z. Guo, H. Lu, X. Huang, X. Tian, X. Xu, Y. Lin, and Z.-Y. Niu. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 107–118, Online only, Nov. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-aacl.10. URL https://aclanthology.org/2022.findings-aacl.10/.

[9] C.-P. Bara, S. CH-Wang, and J. Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. *arXiv preprint arXiv:2109.06275*, 2021.

[10] L. Benotti and P. Blackburn. A recipe for annotating grounded clarifications. *arXiv preprint arXiv:2104.08964*, 2021.

[11] L. Benotti and P. Blackburn. Grounding as a Collaborative Process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.41.

[12] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*, 2023.

[13] C. Bonial, J. Foresta, N. C. Fung, C. J. Hayes, P. Osteen, J. Arkin, B. Hedegaard, and T. Howard. Abstract Meaning Representation for grounded human-robot communication. In J. Bonn and N. Xue, editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 34–44, Nancy, France, June 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.dmr-1.4/.

[14] J. Bonn, M. Palmer, Z. Cai, and K. Wright-Bettner. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.601.

[15] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://aclanthology.org/P19-1470.

[16] B. Byrne, K. Krishnamoorthi, S. Ganesh, and M. Kale. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 671–680, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.55. URL https://aclanthology.org/2021.acl-long.55.

[17] J. Cao, A. Suresh, J. Jacobs, C. Clevenger, A. Howard, C. Brown, B. Milne, T. Fischaber, T. Sumner, and J. H. Martin. Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7671–7684, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.513/.

[18] M. Cao, L. Chen, M. Z. Shou, C. Zhang, and Y. Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.773. URL https://aclanthology.org/2021.emnlp-main.773.

[19] K. R. Chandu, Y. Bisk, and A. W. Black. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.375. URL https://aclanthology.org/2021.findings-acl.375.

[20] K. R. Chandu, Y. Bisk, and A. W. Black. Grounding 'Grounding' in NLP, June 2021.

[21] M. Chaudhary, H. Gupta, S. Bhat, and V. Varma. Towards understanding the robustness of LLM-based evaluations under perturbations. In S. Lalitha Devi and K. Arora, editors, *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 197–205, AU-KBC Research Centre, Chennai, India, Dec. 2024. NLP Association of India (NLPAI). URL https://aclanthology.org/2024.icon-1.22/.

[22] C. Chen, S. Anjum, and D. Gurari. Grounding answers for visual questions asked by visually impaired people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19076–19085. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01851. URL https://doi.org/10.1109/CVPR52688.2022.01851.

[23] C.-Y. Chen, P.-H. Wang, S.-C. Chang, D.-C. Juan, W. Wei, and J.-Y. Pan. AirConcierge: Generating task-oriented dialogue via efficient large-scale knowledge retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 884–897, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.79. URL https://aclanthology.org/2020.findings-emnlp.79.

[24] K. Chen, Q. Huang, D. McDuff, X. Gao, H. Palangi, J. Wang, K. Forbus, and J. Gao. NICE: Neural image commenting with empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4456–4472, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.380. URL https://aclanthology.org/2021.findings-emnlp.380.

[25] N. Chen, Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.570. URL https://aclanthology.org/2023.findings-emnlp.570/.

[26] Q. Chen, W. Wu, and S. Li. Exploring in-context learning for knowledge grounded dialog generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10071–10081, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.675. URL https://aclanthology.org/2023.findings-emnlp.675.

[27] Z. Chen, B. Liu, S. Moon, C. Sankar, P. Crook, and W. Y. Wang. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.197. URL https://aclanthology.org/2022.findings-naacl.197.

[28] J. Chiu, W. Zhao, D. Chen, S. Vaduguru, A. Rush, and D. Fried. Symbolic planning and code generation for grounded dialogue. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7426–7436, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.460. URL https://aclanthology.org/2023.emnlp-main.460/.

[29] H. Cho and J. May. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.218. URL https://aclanthology.org/2020.acl-main.218.

[30] R. Choudhary and D. Kawahara. Grounding in social media: An approach to building a chit-chat dialogue model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 9–15, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-srw.2. URL https://aclanthology.org/2022.naacl-srw.2.

[31] H. H. Clark. Context and Common Ground. In *Encyclopedia of Language & Linguistics*, pages 105–108. Elsevier, 2006. ISBN 978-0-08-044854-1. doi: 10.1016/B0-08-044854-2/01088-9.

[32] H. H. Clark and S. E. Brennan. Grounding in communication. 1991.

[33] N. Daheim, D. Thulke, C. Dugast, and H. Ney. Controllable factuality in document-grounded dialog systems using a noisy channel model. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1365–1381, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.98. URL https://aclanthology.org/2022.findings-emnlp.98/.

[34] N. Daheim, J. Macina, M. Kapur, I. Gurevych, and M. Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.478. URL https://aclanthology.org/2024.emnlp-main.478/.

[35] 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*, pages 326–335, 2017.

[36] J. Davison, J. Feldman, and A. M. Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019.

[37] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. GuessWhat?! Visual object discovery through multimodal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[38] 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.

[39] J. Deng, J. Cheng, H. Sun, Z. Zhang, and M. Huang. Towards safer generative language models: A survey on safety risks, evaluations, and improvements, 2023. URL https://arxiv.org/abs/2302.09270.

[40] J. Deng, Z. Yang, D. Liu, T. Chen, W. Zhou, Y. Zhang, H. Li, and W. Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13636–13652, 2023. doi: 10.1109/TPAMI.2023.3296823. URL https://doi.org/10.1109/TPAMI.2023.3296823.

[41] P. Dillenbourg, D. Traum, and D. Schneider. Grounding in multimodal task-oriented collaboration. In *Proceedings of the European Conference on AI in Education*, pages 401–407, 1996.

[42] T. Dong, A. Testoni, L. Benotti, and R. Bernardi. Visually grounded follow-up questions: A dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31, 2021.

[43] M. Eric, L. Krishnan, F. Charette, and C. D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506.

[44] S. Feng, H. Wan, C. Gunasekara, S. Patel, S. Joshi, and L. Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.652. URL https://aclanthology.org/2020.emnlp-main.652.

[45] S. Feng, S. S. Patel, H. Wan, and S. Joshi. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.498. URL https://aclanthology.org/2021.emnlp-main.498.

[46] J. Françoise, Y. Candau, S. Fdili Alaoui, and T. Schiphorst. Designing for kinesthetic awareness: Revealing user experiences through second-person inquiry. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5171–5183, Denver Colorado USA, May 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025714.

[47] H. Fu, Y. Zhang, H. Yu, J. Sun, F. Huang, L. Si, Y. Li, and C. T. Nguyen. Doc2Bot: Accessing heterogeneous documents via conversational bots. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1820–1836, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.131. URL https://aclanthology.org/2022.findings-emnlp.131/.

[48] F. Galetzka, C. U. Eneh, and D. Schlangen. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 565–573, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.71.

[49] J. Gao, M. Galley, and L. Li. *Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots*. Now Foundations and Trends, 2019. ISBN 1-68083-552-1.

[50] J. Gao, Y. Lian, Z. Zhou, Y. Fu, and B. Wang. LiveChat: A large-

scale personalized dialogue dataset automatically constructed from live streaming. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.858. URL https://aclanthology.org/2023.acl-long.858/.

[51] S. Gao, B. Borges, S. Oh, D. Bayazit, S. Kanno, H. Wakaki, Y. Mitsufuji, and A. Bosselut. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.362. URL https://aclanthology.org/2023.acl-long.362/.

[52] F. Gervits, K. Eberhard, and M. Scheutz. Disfluent but effective? a quantitative study of disfluencies and conversational moves in team discourse. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3359–3369, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1317.

[53] F. Gervits, A. Roque, G. Briggs, M. Scheutz, and M. Marge. How should agents ask questions for situated learning? an annotated dialogue corpus. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 353–359, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.37.

[54] A. Ghaddar, D. Alfonso-Hermelo, P. Langlais, M. Rezagholizadeh, B. Chen, and P. Parthasarathi. CHARP: Conversation history AwaReness probing for knowledge-grounded dialogue systems. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1534–1551, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.90. URL https://aclanthology.org/2024.findings-acl.90/.

[55] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. ISBN 2374-3468.

[56] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*, 2020.

[57] L. Golany, F. Galgani, M. Mamo, N. Parasol, O. Vandsburger, N. Bar, and I. Dagan. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1908–1925, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.106. URL https://aclanthology.org/2024.findings-emnlp.106/.

[58] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00380. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

[59] J. Haber, T. Baumgärtner, E. Takmaz, L. Gelderloos, E. Bruni, and R. Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1184. URL https://aclanthology.org/P19-1184.

[60] T. K. Harris and A. I. Rudnicky. Teamtalk: A platform for multi-human-robot dialog research in coherent real and virtual spaces. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 1864–1865. AAAI Press, 2007. URL http://www.aaai.org/Library/AAAI/2007/aaai07-307.php.

[61] B. Hedayatnia, D. Jin, Y. Liu, and D. Hakkani-Tur. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 298–311, Edinburgh, UK, Sept. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sigdial-1.30.

[62] D. Hofs, M. Theune, and R. op den Akker. Natural interaction with a virtual guide in a virtual environment: A multimodal dialogue system. *Journal on Multimodal User Interfaces*, 3:141–153, 2010. ISSN 1783-7677.

[63] R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. *arXiv preprint arXiv:1906.00347*, 2019.

[64] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

[65] T. Iki and A. Aizawa. Language-Conditioned Feature Pyramids for Visual Selection Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4687–4697, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.420. URL https://aclanthology.org/2020.findings-emnlp.420.

[66] N. Ilinykh, S. Zarrieß, and D. Schlangen. MeetUp! A corpus of joint activity dialogues in a visual environment. *arXiv preprint arXiv:1907.05084*, 2019.

[67] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

[68] Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng, and P. Fung. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, 2023.

[69] C. Jiang, B. Qi, X. Hong, D. Fu, Y. Cheng, F. Meng, M. Yu, B. Zhou, and J. Zhou. On large language models' hallucination with regard to known facts. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.60. URL https://aclanthology.org/2024.naacl-long.60.

[70] K. Jokinen. The need for grounding in LLM-based dialogue systems. In T. Dong, E. Hinrichs, Z. Han, K. Liu, Y. Song, Y. Cao, C. F. Hempelmann, and R. Sifa, editors, *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 45–52, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.neusymbridge-1.5/.

[71] 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)*, pages 3299–3310, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.267. URL https://aclanthology.org/2020.emnlp-main.267.

[72] G.-C. Kang, J. Lim, and B.-T. Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2024–2033, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1209. URL https://aclanthology.org/D19-1209.

[73] S. Katayama, A. Mathur, M. Van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar. Situation-aware emotion regulation of conversational agents with kinetic earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 725–731. IEEE, 2019. ISBN 1-72813-888-4.

[74] C. Kennington, S. Kousidis, and D. Schlangen. Interpreting situated dialogue utterances: An update model that uses speech, gaze, and gesture information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 173–182, 2013.

[75] I. K. Khebour, K. Lai, M. Bradford, Y. Zhu, R. A. Brutti, C. Tam, J. Tu, B. A. Ibarra, N. Blanchard, N. Krishnaswamy, and J. Pustejovsky. Common ground tracking in multimodal dialogue. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.318/.

[76] S. Khosla. Information extraction and program synthesis from goal-oriented dialogue. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meet-*

ing of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 51–53, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.19/.

[77] H. Kim, H. Tan, and M. Bansal. Modality-balanced models for visual dialogue. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8091–8098. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6320. URL https://doi.org/10.1609/aaai.v34i05.6320.

[78] H. Kim, Y. Yu, L. Jiang, X. Lu, D. Khashabi, G. Kim, Y. Choi, and M. Sap. ProsocialDialog: A prosocial backbone for conversational agents. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.267. URL https://aclanthology.org/2022.emnlp-main.267/.

[79] J.-H. Kim, N. Kitaev, X. Chen, M. Rohrbach, B.-T. Zhang, Y. Tian, D. Batra, and D. Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1651. URL https://aclanthology.org/P19-1651.

[80] M. Kim, C. Kim, Y. H. Song, S.-w. Hwang, and J. Yeo. BotsTalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5149–5170, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.344. URL https://aclanthology.org/2022.emnlp-main.344/.

[81] S. Kim, M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. Hakkani-Tur. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.35.

[82] S. Kim, S. Dai, M. Kachuee, S. Ray, T. Taghavi, and S. Yoon. GrounDial: Human-norm grounded safe dialog response generation. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1582–1588, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.109/.

[83] T. Kodama, H. Kiyomaru, Y. J. Huang, T. Okahisa, and S. Kurohashi. Is a knowledge-based response engaging?: An analysis on knowledge-grounded dialogue with information source annotation. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 237–243, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.34. URL https://aclanthology.org/2023.acl-srw.34/.

[84] I. Kola, P. K. Murukannaiah, C. M. Jonker, and M. B. Van Riemsdijk. Toward social situation awareness in support agents. *IEEE intelligent systems*, 37(5):50–58, 2022. ISSN 1541-1672.

[85] D. Kontogiorgos, E. Sibirtseva, and J. Gustafson. Chinese whispers: A multimodal dataset for embodied language grounding. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 743–749, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.93.

[86] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1058. URL https://aclanthology.org/N19-1058.

[87] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.401. URL https://aclanthology.org/2021.emnlp-main.401.

[88] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi. Navigating connected memories with a task-oriented dialog system. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2495–2507, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.160. URL https://aclanthology.org/2022.emnlp-main.160/.

[89] J. Kruijt and P. Vossen. The role of common ground for referential expressions in social dialogues. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Gyeongju, Republic of Korea, Oct. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.crac-1.10.

[90] S. Larsson. Grounding as a side-effect of grounding. *Topics in cognitive science*, 10(2):389–408, 2018.

[91] D. Le, R. Guo, W. Xu, and A. Ritter. Improved instruction ordering in recipe-grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10086–10104, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.561. URL https://aclanthology.org/2023.acl-long.561/.

[92] H. Le and S. C. Hoi. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5842–5848, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.518. URL https://aclanthology.org/2020.acl-main.518.

[93] H. Le, D. Sahoo, N. Chen, and S. Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1564. URL https://aclanthology.org/P19-1564.

[94] H. Le, C. Sankar, S. Moon, A. Beirami, A. Geramifard, and S. Kottur. DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5651–5665, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.439. URL https://aclanthology.org/2021.acl-long.439.

[95] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[96] S. Li, Y. Yin, C. Yang, W. Jiang, Y. Li, Z. Cheng, L. Shang, X. Jiang, Q. Liu, and Y. Yang. NewsDialogues: Towards proactive news grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3634–3649, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.224. URL https://aclanthology.org/2023.findings-acl.224/.

[97] Y. Li, B. Peng, Y. Shen, Y. Mao, L. Liden, Z. Yu, and J. Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.15. URL https://aclanthology.org/2022.naacl-main.15.

[98] Y. Li, J. Zhao, M. Lyu, and L. Wang. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10551–10564, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.721. URL https://aclanthology.org/2022.emnlp-main.721/.

[99] Y. Li, D. Hazarika, D. Jin, J. Hirschberg, and Y. Liu. From pixels to personas: Investigating and modeling self-anthropomorphism in human-robot dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9695–9713, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.567. URL https://aclanthology.org/2024.findings-emnlp.567/.

[100] J. Lim, M. Kang, Y. Hur, S. W. Jeong, J. Kim, Y. Jang, D. Lee, H. Ji, D. Shin, S. Kim, and H. Lim. You truly understand what I need : Intellectual and friendly dialog agents grounding persona and knowledge. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1053–1066, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.75. URL https://aclanthology.org/2022.findings-emnlp.75/.

[101] B. Liu and S. Mazumder. Lifelong and continual learning dialogue systems: Learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15058–15063, 2021. ISBN 2374-3468.

[102] S. Liu, X. Zhao, B. Li, F. Ren, L. Zhang, and S. Yin. A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.173. URL https://aclanthology.org/2021.emnlp-main.173.

[103] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269.

[104] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. D. Campbell. DRAGON: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics Autom. Lett.*, 9(4):3712–3719, 2024. doi: 10.1109/LRA.2024.3362591. URL https://doi.org/10.1109/lra.2024.3362591.

[105] Z. Liu, M. Patwary, R. Prenger, S. Prabhumoye, W. Ping, M. Shoeybi, and B. Catanzaro. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.104. URL https://aclanthology.org/2022.findings-acl.104.

[106] H. Lu, W. Lam, H. Cheng, and H. Meng. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5200–5212, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.382. URL https://aclanthology.org/2022.naacl-main.382.

[107] S. M. Lukin, F. Gervits, C. J. Hayes, P. Moolchandani, A. Leuski, J. G. Rogers III, C. Sanchez Amaro, M. Marge, C. R. Voss, and D. Traum. ScoutBot: A dialogue system for collaborative navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4016. URL https://aclanthology.org/P18-4016.

[108] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, X. Chen, and M. Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv*, abs/2002.06353, 2020. URL https://api.semanticscholar.org/CorpusID:211132410.

[109] S.-B. Luo, C.-C. Fan, K.-Y. Chen, Y. Tsao, H.-M. Wang, and K.-Y. Su. Chinese movie dialogue question answering dataset. In Y.-C. Chang and Y.-C. Huang, editors, *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 7–14, Taipei, Taiwan, Nov. 2022. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/2022.rocling-1.2/.

[110] W. Ma, R. Takanobu, and M. Huang. CR-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.139. URL https://aclanthology.org/2021.emnlp-main.139.

[111] J. Macina, N. Daheim, S. Chowdhury, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

[112] B. Madureira. Incrementally enriching the common ground: A research path. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 57–58, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.21/.

[113] B. Madureira and D. Schlangen. It couldn't help but overhear: On the limits of modelling meta-communicative grounding acts with supervised learning. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 149–158, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.13. URL https://aclanthology.org/2024.sigdial-1.13/.

[114] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

[115] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.739. URL https://aclanthology.org/2020.emnlp-main.739.

[116] M. Markowska, M. Taghizadeh, A. Soubki, S. Mirroshandel, and O. Rambow. Finding common ground: Annotating and predicting common ground in spoken conversations. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.551. URL https://aclanthology.org/2023.findings-emnlp.551/.

[117] M. Mazuecos, F. M. Luque, J. Sánchez, H. Maina, T. Vadora, and L. Benotti. Region under Discussion for visual dialog. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4745–4759, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.390. URL https://aclanthology.org/2021.emnlp-main.390.

[118] J. Miehle, K. Yoshino, L. Pragst, S. Ultes, S. Nakamura, and W. Minker. Cultural communication idiosyncrasies in human-computer interaction. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 74–79, Los Angeles, Sept. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3610. URL https://aclanthology.org/W16-3610.

[119] K. Mitsuda, R. Higashinaka, Y. Oga, and S. Yoshida. Dialogue collection for recording the process of building common ground in a collaborative task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5749–5758, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.618.

[120] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1255. URL https://aclanthology.org/D18-1255.

[121] B. Mohapatra. Conversational grounding in multimodal dialog systems. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 15–17, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.5/.

[122] B. Mohapatra, S. Hassan, L. Romary, and J. Cassell. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.352/.

[123] B. Mohapatra, M. N. Kapadnis, L. Romary, and J. Cassell. Evaluating the effectiveness of large language models in establishing conversational grounding. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.545. URL https://aclanthology.org/2024.emnlp-main.545/.

[124] S. Moon, P. Shah, A. Kumar, and R. Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1081. URL https://aclanthology.org/P19-1081.

[125] S. Moon, P. Shah, R. Subba, and A. Kumar. Memory grounded conversational reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 145–150, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3025. URL https://aclanthology.org/D19-3025.

[126] S. Moon, S. Kottur, P. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, R. Subba, and A. Geramifard. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.96. URL https://aclanthology.org/2020.coling-main.96.

[127] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1047.

[128] M. Moutti, S. Eleftheriou, P. Koromilas, and T. Giannakopoulos. A dataset for speech emotion recognition in Greek theatrical plays. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1040–1046, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.111.

[129] K. Nakamura, S. Levy, Y.-L. Tuan, W. Chen, and W. Y. Wang. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.41. URL https://aclanthology.org/2022.findings-acl.41.

[130] K. Naszadi, P. Manggala, and C. Monz. Aligning predictive uncertainty with clarification questions in grounded dialog. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14988–14998, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.999. URL https://aclanthology.org/2023.findings-emnlp.999/.

[131] S. Ostermann, S. Zhang, M. Roth, and P. Clark. Commonsense inference in natural language processing (COIN) - shared task report. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 66–74, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6007. URL https://aclanthology.org/D19-6007.

[132] S. Pal, S. Das, and R. K. Srihari. Beyond discrete personas: Personality modeling through journal intensive conversations. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7055–7074, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.470/.

[133] L. Parcalabescu, N. Trost, and A. Frank. What is multimodality? *arXiv preprint arXiv:2103.06304*, 2021.

[134] G. Paré and S. Kitsiou. Chapter 9 methods for literature reviews. handbook of ehealth evaluation: An evidence-based approach [internet]. university of victoria, 2017.

[135] J. Park, M. Joo, J.-K. Kim, and H. J. Kim. Generative subgraph retrieval for knowledge graph–grounded dialog generation. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21167–21182, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1179. URL https://aclanthology.org/2024.emnlp-main.1179/.

[136] R. Pasunuru and M. Bansal. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1012. URL https://aclanthology.org/D18-1012.

[137] D. Petrak, T. T. Tran, and I. Gurevych. Learning from implicit user feedback, emotions and demographic information in task-oriented and document-grounded dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4573–4603, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.264. URL https://aclanthology.org/2024.findings-emnlp.264/.

[138] C. Pryor, Q. Yuan, J. Liu, M. Kazemi, D. Ramachandran, T. Bedrax-Weiss, and L. Getoor. Using domain knowledge to guide dialog structure induction via neural probabilistic soft logic. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7631–7652, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.422. URL https://aclanthology.org/2023.acl-long.422/.

[139] J. Pustejovsky and N. Krishnaswamy. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. *Traitement Automatique des Langues*, 61(3):17–41, 2020.

[140] K. Qian and Z. Yu. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1253. URL https://aclanthology.org/P19-1253.

[141] Y. Qin, W. Ji, X. Lan, H. Fei, X. Yang, D. Guo, R. Zimmermann, and L. Liao. Grounding is all you need? dual temporal grounding for video dialog. *ArXiv*, abs/2410.05767, 2024. URL https://api.semanticscholar.org/CorpusID:273228747.

[142] D. Raghu, S. Agarwal, S. Joshi, and Mausam. End-to-end learning of flowchart grounded task-oriented dialogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4348–4366, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.357. URL https://aclanthology.org/2021.emnlp-main.357.

[143] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

[144] A. Razzhigaev, M. Kurkin, E. Goncharova, I. Abdullaeva, A. Lysenko, A. Panchenko, A. Kuznetsov, and D. Dimitrov. OmniDialog: A multimodal benchmark for generalization across text, visual, and audio modalities. In D. Hupkes, V. Dankers, K. Batsuren, A. Kazemnejad, C. Christodoulopoulos, M. Giulianelli, and R. Cotterell, editors, *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 183–195, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.genbench-1.12. URL https://aclanthology.org/2024.genbench-1.12/.

[145] C. Richardson and L. Heck. Commonsense reasoning for conversational AI: A survey of the state of the art. *arXiv preprint arXiv:2302.07926*, 2023.

[146] P. Rodriguez, P. Crook, S. Moon, and Z. Wang. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.655. URL https://aclanthology.org/2020.emnlp-main.655.

[147] F. Ruggeri, M. Mesgar, and I. Gurevych. A dataset of argumentative dialogues on scientific papers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.425. URL https://aclanthology.org/2023.acl-long.425/.

[148] S. Santhanam, B. Hedayatnia, S. Gella, A. Padmakumar, S. Kim, Y. Liu, and D. Hakkani-Tur. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *CoRR*, abs/2110.05456, 2021. URL https://arxiv.org/abs/2110.05456.

[149] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33013027. URL https://doi.org/10.1609/aaai.v33i01.33013027.

[150] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019. ISBN 2374-3468.

[151] P. Schneider, N. Machner, K. Jokinen, and F. Matthes. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 110–120, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.10. URL https://aclanthology.org/2024.sigdial-1.10/.

[152] O. Shaikh, K. Gligoric, A. Khetan, M. Gerstgrasser, D. Yang, and D. Jurafsky. Grounding gaps in language model generations. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.348. URL https://aclanthology.org/2024.naacl-long.348.

[153] K. Shuster, S. Humeau, A. Bordes, and J. Weston. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.219. URL https://aclanthology.org/2020.acl-main.219.

[154] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. ISBN 2374-3468.

[155] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25:701–721, 2002. URL https://api.semanticscholar.org/CorpusID:265871097.

[156] C. Strathearn and D. Gkatzia. Task2Dial: A novel task and dataset for commonsense-enhanced task-based dialogue grounded in documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 187–196, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.21. URL https://aclanthology.org/2022.dialdoc-1.21.

[157] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[158] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

[159] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019. doi: 10.1109/ICCV.2019.00756.

[160] H. Sun, Z. Cao, and D. Yang. SPORTSINTERVIEW: A large-scale sports interview benchmark for entity-centric dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5821–5828, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.626.

[161] M. Sung, S. Feng, J. Gung, R. Shu, Y. Zhang, and S. Mansour. Structured list-grounded question answering. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8347–8359, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.558/.

[162] S. K. Suresh, M. Wu, T. Pranav, and E. Chng. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 673–690. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-naacl.40/.

[163] T. Takenobu, I. Ryu, T. Asuka, and K. Naoko. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of LREC*, pages 422–429, 2012.

[164] C. Tam, R. Brutti, K. Lai, and J. Pustejovsky. Annotating situated actions in dialogue. In J. Bonn and N. Xue, editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 45–51, Nancy, France, June 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.dmr-1.5/.

[165] H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

[166] R. Titung and C. O. Alm. FUSE - FrUstration and surprise expressions: A subtle emotional multimodal language corpus. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7544–7555, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.666/.

[167] J. Torres-Foncesca, C. Henry, and C. Kennington. Symbol and communicative grounding through object permanence with a mobile robot. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 124–134, Edinburgh, UK, Sept. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sigdial-1.14.

[168] Y.-L. Tuan, Y.-N. Chen, and H.-y. Lee. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1194. URL https://aclanthology.org/D19-1194.

[169] T. Udagawa and A. Aizawa. A natural language corpus of common grounding under continuous and partially-observable context. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7120–7127. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33017120. URL https://doi.org/10.1609/aaai.v33i01.33017120.

[170] T. Udagawa and A. Aizawa. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011, 2021. doi: 10.1162/tacl_a_00409. URL https://aclanthology.org/2021.tacl-1.59.

[171] T. Udagawa, T. Yamazaki, and A. Aizawa. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.67. URL https://aclanthology.org/2020.findings-emnlp.67.

[172] N. Ueda, H. Habe, A. Yuguchi, S. Kawano, Y. Kawanishi, S. Kurohashi, and K. Yoshino. J-CRe3: A Japanese conversation dataset for real-world reference resolution. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9489–9502, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.829/.

[173] H. Wang, M. Hu, Y. Deng, R. Wang, F. Mi, W. Wang, Y. Wang, W.-C. Kwan, I. King, and K.-F. Wong. Large language models as source planner for personalized knowledge-grounded dialogues. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.641. URL https://aclanthology.org/2023.findings-emnlp.641.

[174] Y. Wang, Z. Zheng, X. Zhao, J. Li, Y. Wang, and D. Zhao. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In A. Rogers, J. Boyd-Graber,

and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.276. URL https://aclanthology.org/2023.acl-long.276/.

[175] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[176] N. Weir, R. Thomas, R. d'Amore, K. Hill, B. Van Durme, and H. Jhamtani. Ontologically faithful generation of non-player character dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9242, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.520. URL https://aclanthology.org/2024.emnlp-main.520/.

[177] A. Welivita, C.-H. Yeh, and P. Pu. Empathetic response generation for distress support. In S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.59. URL https://aclanthology.org/2023.sigdial-1.59/.

[178] X.-F. Wen, W. Wei, and X.-L. Mao. Sequential topic selection model with latent variable for topic-grounded dialogue. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1209–1219, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.87. URL https://aclanthology.org/2022.findings-emnlp.87.

[179] J. White, G. Poesia, R. Hawkins, D. Sadigh, and N. Goodman. Open-domain clarification question generation without question examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 563–570, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.44. URL https://aclanthology.org/2021.emnlp-main.44.

[180] C. H. Wu, Y. Zheng, X. Mao, and M. Huang. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2368–2382, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.183. URL https://aclanthology.org/2021.emnlp-main.183.

[181] C.-S. Wu, A. Madotto, W. Liu, P. Fung, and C. Xiong. QAConv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.370. URL https://aclanthology.org/2022.acl-long.370.

[182] S. Wu, Y. Li, P. Xue, D. Zhang, and Z. Wu. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.43.

[183] S. Wu, W. Hsu, and M. L. Lee. EHDChat: A knowledge-grounded, empathy-enhanced language model for healthcare interactions. In J. Hale, K. Chawla, and M. Garg, editors, *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 141–151, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sicon-1.10. URL https://aclanthology.org/2024.sicon-1.10/.

[184] T.-L. Wu, S. Kottur, A. Madotto, M. Azab, P. Rodriguez, B. Damavandi, N. Peng, and S. Moon. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.345. URL https://aclanthology.org/2023.acl-long.345/.

[185] F. Xia, B. Li, Y. Weng, S. He, K. Liu, B. Sun, S. Li, and J. Zhao. MedConQA: Medical conversational question answering system based on knowledge graphs. In W. Che and E. Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 148–158, Abu Dhabi, UAE, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.15. URL https://aclanthology.org/2022.emnlp-demos.15/.

[186] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang, V. Zhong, B. Wang, C. Li, C. Boyle, A. Ni, Z. Yao, D. Radev, C. Xiong, L. Kong, R. Zhang, N. A. Smith, L. Zettlemoyer, and T. Yu. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.39. URL https://aclanthology.org/2022.emnlp-main.39.

[187] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, B. Liu, and P. Yu. User memory reasoning for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5288–5308, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.463. URL https://aclanthology.org/2020.coling-main.463.

[188] L. Xu, Q. Zhou, J. Fu, M.-Y. Kan, and S.-K. Ng. CorefDiffs: Coreferential and differential knowledge flow in document grounded conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 471–484, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.38.

[189] B. Xue, W. Wang, H. Wang, F. Mi, R. Wang, Y. Wang, L. Shang, X. Jiang, Q. Liu, and K.-F. Wong. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.525. URL https://aclanthology.org/2023.findings-emnlp.525/.

[190] S. Yamashita, K. Inoue, A. Guo, S. Mochizuki, T. Kawahara, and R. Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.85/.

[191] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. ISBN 2374-3468.

[192] J. Yu, S. Wu, J. Chen, and W. Zhou. LLMs as collaborator: Demands-guided collaborative retrieval-augmented generation for commonsense knowledge-grounded open-domain dialogue systems. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13586–13612, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.794. URL https://aclanthology.org/2024.findings-emnlp.794/.

[193] Y. Yu, A. Eshghi, G. Mills, and O. Lemon. The BURCHAK corpus: a challenge data set for interactive learning of visually grounded word meanings. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2001. URL https://aclanthology.org/W17-2001.

[194] M. Zare, A. Wagner, and R. Passonneau. A POMDP dialogue policy with 3-way grounding and adaptive Sensing for learning through communication. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6767–6780, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.504. URL https://aclanthology.org/2022.findings-emnlp.504/.

[195] S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernández, and D. Schlangen. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, 2016.

[196] W. Zeng, K. He, Z. Wang, D. Fu, G. Dong, R. Geng, P. Wang, J. Wang,

C. Sun, W. Wu, and W. Xu. Semi-supervised knowledge-grounded pre-training for task-oriented dialog systems. In Z. Ou, J. Feng, and J. Li, editors, *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 39–47, Abu Dhabi, Beijing (Hybrid), Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.seretod-1.6. URL https://aclanthology.org/2022.seretod-1.6/.

[197] H. Zhan, S. Maruf, I. Zukerman, and G. Haffari. Going beyond imagination! enhancing multi-modal dialogue agents with synthetic visual descriptions. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 420–427, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.36. URL https://aclanthology.org/2024.sigdial-1.36/.

[198] H. Zhang, Z. Liu, C. Xiong, and Z. Liu. Grounded conversation generation as guided traverses in commonsense knowledge graphs. *arXiv preprint arXiv:1911.02707*, 2019.

[199] J. Zhang, K. Qian, Z. Liu, S. Heinecke, R. Meng, Y. Liu, Z. Yu, H. Wang, S. Savarese, and C. Xiong. DialogStudio: Towards richest and most diverse unified dataset collection for conversational AI. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2299–2315, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.152/.

[200] K. Zhang, Y. Kang, F. Zhao, and X. Liu. LLM-based medical assistant personalization with short- and long-term memory coordination. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.132. URL https://aclanthology.org/2024.naacl-long.132/.

[201] R. Zhang and C. Eickhoff. SOCCER: An information-sparse discourse state tracking collection in the sports commentary domain. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4325–4333, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.342. URL https://aclanthology.org/2021.naacl-main.342.

[202] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1205. URL https://aclanthology.org/P18-1205/.

[203] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[204] X. Zhang, R. Divekar, R. Ubale, and Z. Yu. GrounDialog: A dataset for repair and grounding in task-oriented spoken dialogues for language learning. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.26. URL https://aclanthology.org/2023.bea-1.26/.

[205] X. Zhang, R. Divekar, R. Ubale, and Z. Yu. GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, 2023.

[206] Y. Zhang, P. Ren, W. Deng, Z. Chen, and M. Rijke. Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3543–3555, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.248. URL https://aclanthology.org/2022.acl-long.248.

[207] C. Zhao, S. Gella, S. Kim, D. Jin, D. Hazarika, A. Papangelis, B. Hedayatnia, M. Namazifar, Y. Liu, and D. Hakkani-Tur. "what do others think?": Task-oriented conversational modeling with subjective knowledge. In S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 309–323, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.28. URL https://aclanthology.org/2023.sigdial-1.28/.

[208] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui. Retrieval-augmented generation for ai-generated content: A survey, 2024. URL https://arxiv.org/abs/2402.19473.

[209] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.272. URL https://aclanthology.org/2020.emnlp-main.272.

[210] X. Zhao, T. Fu, C. Tao, and R. Yan. There is no standard answer: Knowledge-grounded dialogue generation with adversarial activated multi-reference learning. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1878–1891, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.123. URL https://aclanthology.org/2022.emnlp-main.123/.

[211] Y. Zheng, G. Chen, X. Liu, and J. Sun. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.621.

[212] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4623–4629, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi: 10.24963/ijcai.2018/643.

[213] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.635. URL https://aclanthology.org/2020.acl-main.635.

[214] H. Zhou, M. Huang, Y. Liu, W. Chen, and X. Zhu. EARL: Informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2395, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.184. URL https://aclanthology.org/2021.emnlp-main.184.

[215] P. Zhou, H. Cho, P. Jandaghi, D.-H. Lee, B. Y. Lin, J. Pujara, and X. Ren. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.714. URL https://aclanthology.org/2022.emnlp-main.714/.

[216] W. Zhu, K. Mo, Y. Zhang, Z. Zhu, X. Peng, and Q. Yang. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264, 2017. URL http://arxiv.org/abs/1709.04264.

A Appendix

A.1 Paper Selection Process

The initial selection of papers is based on the ACL Anthology, the selection was done using a keyword match, i.e. we considered all papers published between 2015 and 2025 whose title or abstract contain both terms "ground(ing)" and "dialogue". In total, we retrieved and annotated 448 papers with respect to topic (models, datasets, evaluation, theory, supplementary), modality (knowledge, visual, multimodal, persona, other), scope (contextual, domain, commonsense, mixed), and type (static, dynamic, mixed). Note that among the keyword-selected papers there were also some that do not directly relate to the topic of grounding in dialogue and the match could happen e.g. if the paper talks about dialogue processing using ground-truth labels (i.e. term "ground" matches but has a different meaning). Such papers received an additional annotation "irrelevant" and were excluded from the final statistics. We also excluded supplementary papers that are relevant for the grounding in dialogue but focus on very specialized topics (e.g. self-anthropomorphism in robots [99]) or papers introducing dialogue researchers participating in YRRSDS round table [121, 76, 112]. The number of grounding papers decreased from 448 to 384 after filtering out spurious matches and supplementary papers. The initial selection was extended with the papers published in the venues other than the ACL Anthology based on the citations and background knowledge of the authors.

A.2 Topic, Modality, Scope and Type Distribution

Figure 4 shows the topic distribution of the selected papers and indicates a substantial imbalance with the majority of papers (61%) dedicated to the modeling approaches, while 23% of the papers introduce new datasets, and only 10% focus on evaluation. It is important to note that new modeling approaches are typically accompanied with the evaluation results, but the evaluation is usually quite limited and includes only 1-2 baselines and the proposed approach on a few selected datasets. More rigorous benchmarking and model comparison is still missing in the current research on grounding in dialogue.

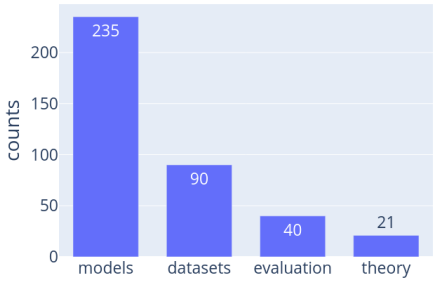

Figure 4. Topic distribution of the papers on grounding in dialogue.

This survey annotates the papers according to the following core modalities: knowledge, visual, multimodal, persona and other. Figure 5 shows that the majority of the papers (60%) is about knowledge grounding, visual grounding is represented at 16%, multimodality is discussed in 12% of the papers, and the rest is almost equally spread among the mixed topics and persona-grounding.

Grounding scope an be characterized as contextual, domain-specific, commonsense or mixed (see Figure 9). It is interesting to see that contextual common ground is the most commonly researched

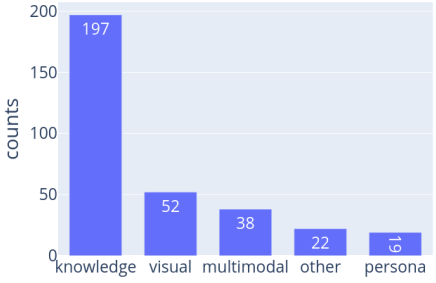

Figure 5. Modality distribution of the papers on grounding in dialogue.

topic, while domain-specific knowledge is also often considered. A significant proportion of papers (27%) has mixed scope and only 5 papers (2%) focus exclusively on the commonsense grounding.

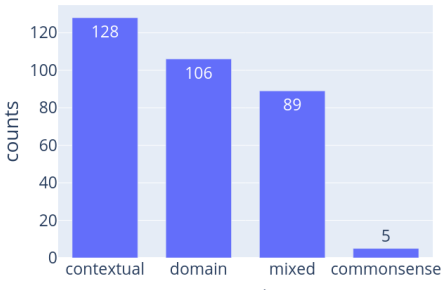

Figure 6. Scope distribution of the papers on grounding in dialogue.

The distribution of grounding types: static vs. dynamic (see Figure 7) makes it clear that static grounding is much better researched than dynamic or mixed cases (46% vs. 30 and 24% correspondingly). This can be likely attributed to the fact that static grounding is easier to model because it often requires an access to a knowledge base or a document collection and the grounding knowledge does not change throughout the conversation. Dynamic grounding requires the dialogue agent to be pro-active, being able to identify ambiguous cases and resolve misunderstanding, e.g. by asking clarification questions.

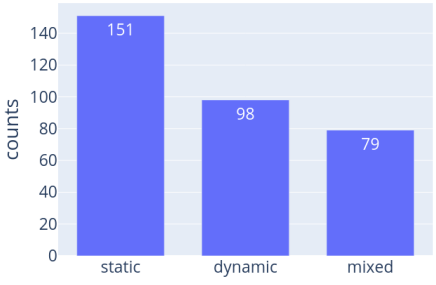

Figure 7. Type distribution of the papers on grounding in dialogue.

A.3 Modality, Scope and Type of Datasets

In this section we summarize the annotation results based on the papers describing the datasets. Figure 8 shows that knowledge modality is much more prevalent than others (51%). Only 16% of the

datasets are visual and 17% are multimodal. Also, the grounding scope has unbalanced distribution with 49% datasets related to contextual grounding and 29% domain-specific ones. Interestingly, static and dynamic grounding are almost equally represented in the existing datasets (which is different from the type distribution for modeling approaches as shown in Section A.4)

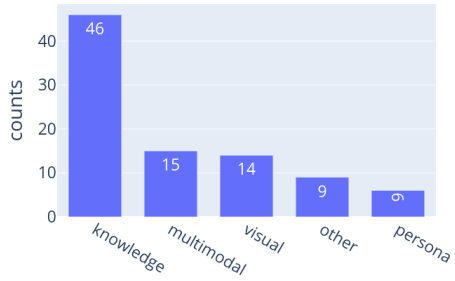

Figure 8. Modality distribution of the *datasets* for grounding in dialogue.

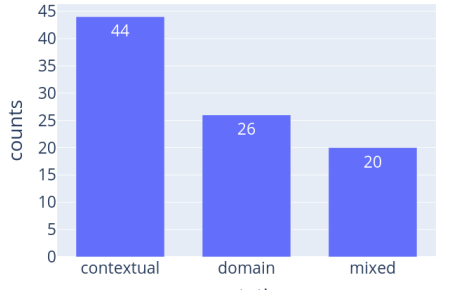

Figure 9. Scope distribution of the *datasets* for grounding in dialogue.

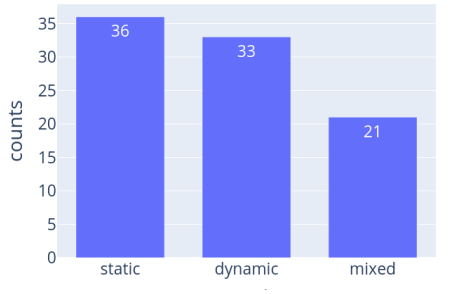

Figure 10. Type distribution of the *datasets* for grounding in dialogue.

A.4 Modality, Scope and Type of Models

If we consider only those papers that describe modeling approaches and plot their distribution per modality, we can see in Figure 11 that knowledge-based approaches are the most common ones. Multimodal and visual grounding receive less attention (11 and 17% each) and the least researched modality is persona-based grounding with only 6% of all papers addressing this topic. These statistics emphasize that there is a lack of research on grounding in the modalities that go beyond knowledge, especially when multiple modalities and persona-related features should be taken into consideration.

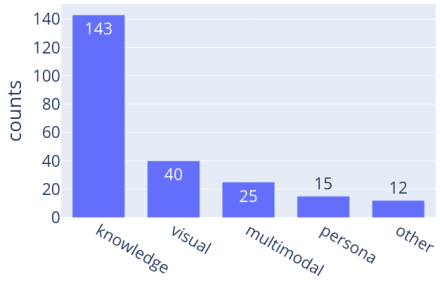

Figure 11. Modality distribution of the *models* for grounding in dialogue.

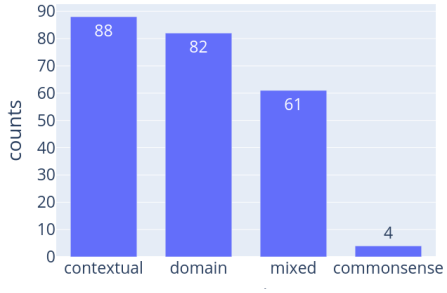

Figure 12. Scope distribution of the *models* for grounding in dialogue.

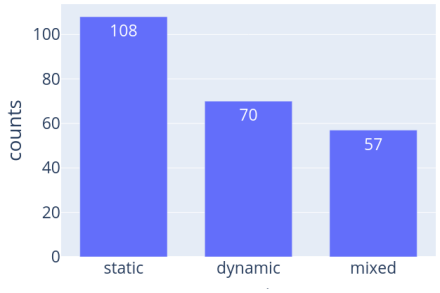

Figure 13. Type distribution of the *models* for grounding in dialogue.

A.5 Grounding Datasets

In the scope of this survey, we compiled a list of the datasets for grounding in dialogue categorized according to the modality, type, and scope (see Table 2-7). The datasets include several additional resources that were not published through the ACL Anthology. This information along with our paper annotations will be made available to the research community in a GitHub repository and we will provide a link to it in the non-anonymized version of the paper.

Table 2. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
SAGA22 [17]	The SAGA22 dataset is based on 148 transcribed videos, it is a dataset of teacher and student talk moves and annotated math tutoring sessions. Talk moves use dialogue acts grounded in Accountable Talk theory.	Knowl.	Dyn.	Cont.	upon request/unknown
Reannotated Spot the Difference and Meetup Datasets [122]	The Meetup and Spot the Difference datasets were (re-)annotated with Grounding Acts, Common Grounding units, and degrees of grounding.	Knowl.	Dyn.	Cont.	https://osf.io/qfcnm/?view_only=34e7259fe8fc4ade82d55ba7d5105ffe
The Common Ground Corpus [116]	The Common Ground Corpus is annotated on the top of the LDC CALLHOME American Speech corpus, which consists of collections of 120 unscripted dialogs between close friends or family members. The dialogs are available in both written form and audio. The Common Ground corpus is the first attempt at annotating common ground in a discourse, providing the annotations for beliefs and common ground updates.	Knowl.	Dyn.	Cont.	https://github.com/cogstates/2023-emnlp-common-ground
GrounDialog [204]	An annotated dataset of spoken conversations with repair and grounding patterns. The dataset contains 42 dialogues with 1569 turns.	Knowl.	Dyn.	Cont.	upon request/unknown
CoomonLayout [119]	The dataset is built for the CommonLayout task in which two workers lay out the same figure set into a common design through text chat. To perform the task, they discuss the idea of a final layout and move figures into the same position one by one. The dataset contains 984 dialogues and each dialogue has 28.8 utterances on average.	Knowl.	Dyn.	Cont.	upon request/unknown
Reflect [215]	Reflect is a dataset that annotates dialogues with explicit common ground (represented as inferences approximating shared knowledge and beliefs) and contains 9K diverse human-generated responses each following one common ground.	Knowl.	Dyn.	Cont.	https://inklab.usc.edu/Reflect/
SPOLIN [29]	Selected Pairs Of Learnable ImprovisatioN (SPOLIN) corpus is a collection of more than 26K English dialogue turn pairs, each consisting of a prompt and subsequent grounded response, where responses are not only coherent with dialogue context but also initiate the next relevant contribution.	Knowl.	Dyn.	Cont.	https://justin-cho.com/spolin
KNUDGE [176]	KNUDGE (KNowledge Constrained User-NPC Dialogue GEneration) is constructed from side quest dialogues drawn directly from game data of Obsidian Entertainment's The Outer Worlds, leading to real-world complexities in generation: (1) utterances must remain faithful to the game lore, including character personas and backstories; (2) a dialogue must accurately reveal new quest details to the human player; and (3) dialogues are large trees as opposed to linear chains of utterances. KNUDGE contains 159 dialogue trees.	Knowl.	Mix	Cont.	https://github.com/nweir127/KNUDGE
KETOD [27]	KETOD (Knowledge-Enriched Task-Oriented Dialogue) enriches task-oriented dialogues with chit-chat based on relevant entity knowledge. It contains >5K dialogues.	Knowl.	Mix	Cont.	https://github.com/facebookresearch/ketod
ChattyChef [91]	ChattyChef is a dataset of cooking dialogues, designed to support research on instruction-grounded conversational agents. ChattyChef contains 267 dialogues with 26 utterances per dialogue.	Knowl.	Dyn.	Dom.	https://github.com/octaviaguo/ChattyChef
EHD [183]	Empathetic Healthcare Dialogue (EHD) dataset can help with generating human-like empathetic responses within the healthcare domain. It contains a wide range of synthetic, multi-turn dialogues between doctors and patients that are not only emotionally supportive, but also clinically informative. EHD contains 33K dialogues, with an average of 12 utterances per dialogue.	Knowl.	Mix	Dom.	https://huggingface.co/datasets/ericw955/EHD
MathDial [111]	MathDial is a dataset of 3K one-to-one teacher-student tutoring dialogues grounded in multi-step math reasoning problems.	Knowl.	Mix	Dom.	https://github.com/eth-nlped/mathdial
ArgSciChat [147]	ArgSciChat is a dataset of 41 argumentative dialogues between scientists on 20 NLP papers. The dataset includes both exploratory and argumentative questions and answers in a dialogue discourse on a scientific paper.	Knowl.	Mix	Dom.	https://github.com/UKPLab/acl2023-argscichat
KdConv [213]	KdConv, a Chinese multi-domain dataset towards multi-turn Knowledge-driven Conversation with 86K utterances and 4.5K dialogues in three domains.	Knowl.	Mix	Dom.	https://github.com/thu-coai/KdConv
List2QA [161]	List2QA dataset is designed to evaluate the ability of QA systems to respond effectively using list information. The dataset is created from unlabeled customer service documents with language models and model-based filtering, it has >2K utterances.	Knowl.	Stat.	Dom.	upon request/unknown
MISeD – Meeting Information Seeking Dialogs dataset [57]	MISeD – Meeting Information Seeking Dialogs dataset is a dataset of information-seeking dialogues focusing on meeting transcripts for 225 meetings, comprising 432 dialogues, and 4161 query-response pairs.	Knowl.	Stat.	Dom.	https://github.com/google-research-datasets/MISeD
Verify-then-Generate [34]	1K student solutions and their stepwise reasoning chains in the domain of multi-step math problem-solving.	Knowl.	Stat.	Dom.	https://github.com/eth-lre/verify-then-generate
NewsDialogues [96]	A human-to-human Chinese dialogue dataset with 1K conversations with a total of 14.6K utterances and detailed annotations for target topics and knowledge spans.	Knowl.	Stat.	Dom.	https://github.com/SihengLi99/NewsDialogues

Table 3. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
CMDQA [109]	Chinese dialogue-based information-seeking question answering dataset CMDQA, which is mainly applied to the scenario of getting Chinese movie related information. It contains 10K QA dialogs (40K turns in total).	Knowl.	Stat.	Dom.	upon request/unknown
SPORTSINTERVIEW [160]	Dataset in the domain of sports interview, it contains two types of external knowledge sources as knowledge grounding, 150K interview sessions and 34K distinct interviewees.	Knowl.	Stat.	Dom.	upon request/unknown
Doc2Bot [47]	Dataset with over 100K turns based on Chinese documents from five domains.	Knowl.	Stat.	Dom.	https://github.com/Doc2Bot/Doc2Bot
MultiRefKGC [210]	A multi-reference Knowledge-Grounded Conversation (KGC) dataset based on conversations from Reddit with 130K dialogues.	Knowl.	Stat.	Dom.	https://github.com/TingchenFu/MultiRefKGC
CM-CQA [185]	A large-scale Chinese Medical CQA (CM-CQA) dataset based on 45 medical subdomains, 33615 entities, 8808 symptoms, 1294753 dialogues.	Knowl.	Stat.	Dom.	https://github.com/WENGSYX/LingYi
Social-Dialogues-Coreference [89]	Dataset for resolving third-person references in social dialogues (inner and outer-circle references), based on the episodes of the Friends series. It contains social dialogue and long-term connections between mentions that go beyound a single document.	Knowl.	Stat.	Dom.	https://github.com/cltl/inner-outer-coreference
MultiDoc2Dial [45]	Conversations grounded in 488 documents, 4796 dialogues in total.	Knowl.	Stat.	Dom.	https://doc2dial.github.io/multidoc2dial/
TicketTalk [16]	A movie ticketing dialog dataset with 23,789 annotated conversations that range from completely open-ended and unrestricted to more structured in terms of the knowledge base, discourse features, and number of turns.	Knowl.	Stat.	Dom.	https://git.io/JL8an
Doc2Dial [44]	The dataset of goal-oriented dialogues that are grounded in the documents. 4500 annotated conversations grounded in over 450 documents from four domains.	Knowl.	Stat.	Dom.	http://doc2dial.github.io/
Background-aware movie dataset [120]	Background-aware conversation dataset about movies with 90K utterances from 9K conversations grounded in plots, reviewes, comments.	Knowl.	Stat.	Dom.	https://github.com/nikitacs16/Holl-E
Multi-turn and multi-domain dataset [43]	The dataset of 3031 dialogues that are grounded through knowledge bases and span three distinct tasks in the in-car personal assistant space: calendar scheduling, weather information retrieval, and point-of-interest navigation.	Knowl.	Stat.	Dom.	http://nlp.stanford.edu/projects/kvret/kvret_dataset_public.zip
SOCCER [201]	2263 soccer matches including with time-stamped natural language commentary accompanied by discrete events such as a team scoring goals, switching players or being penalized with cards.	Knowl.	Dyn.	Mix	https://github.com/bcbi-edu/p_eickhoff_SOCCER
FloDial [142]	FloDial has 2738 dialogs grounded on 12 different troubleshooting flowcharts.	Knowl.	Dyn.	Mix	https://dair-iitd.github.io/FloDial
OpenDialKG [124]	Open-ended Dialog and KG parallel corpus called OpenDialKG, where each utterance from 15K human-to-human role-playing dialogs is manually annotated with ground-truth reference to corresponding entities and paths from a large-scale KG with 1M+ facts.	Knowl.	Dyn.	Mix	https://github.com/facebookresearch/opendialkg
FEDI [137]	FEDI, the first English task-oriented and document-grounded dialogue dataset annotated with implicit user feedback, emotions and demographic information.	Knowl.	Mix	Mix	https://github.com/UKPLab/FEDI
Situated Actions in Dialogue [164]	Action and Abstract Meaning Representation annotations for first-person point-of-view videos (based on the Fibonacci Weights Task dataset and Epic Kitchens dataset).	Knowl.	Mix	Mix	upon request/unknown
Japanese Move Recommendations with external and speaker-derived grounding [83]	Annotated knowledge-grounded dialogue corpus Japanese Movie Recommendation Dialogue that contains >5K dialogues. Each entity is annotated with its information source, either derived from external knowledge (database-derived) or the speaker's own knowledge, experiences, and opinions (speaker-derived).	Knowl.	Mix	Mix	upon request/unknown
Task2Dial [156]	A dataset of document-grounded task-based dialogues, where an Information Giver (IG) provides instructions (by consulting a document) to an Information Follower (IF). The dataset contains dialogues with an average 18.15 number of turns grounded in 353 documents.	Knowl.	Mix	Mix	http://www.huggingface.co/datasets/cstrathe435/Task2Dial
QAConv [181]	A question-answering (QA) dataset that uses conversations as a knowledge source and offers 34608 QA pairs with both human-written and machine-generated questions.	Knowl.	Mix	Mix	https://github.com/salesforce/QAConv
A Dataset for Conversational Curiosity [146]	14K dialogues (181K utterances) where users and assistants converse about geographic topics like geopolitical entities and locations. This dataset is annotated with pre-existing user knowledge, message-level dialog acts, grounding to Wikipedia, and user reactions to messages.	Knowl.	Mix	Mix	http://curiosity.pedro.ai/
BridgeKG [151]	Annotated human conversations across five knowledge domains, 26 information-seeking conversations and 669 dialogue turns.	Knowl.	Stat.	Mix	https://github.com/philotron/Bridge-KG
DialogStudio [199]	Collection with diverse data from open-domain dialogues, task-oriented dialogues, natural language understanding, conversational recommendation, dialogue summarization, and knowledge-grounded dialogues.	Knowl.	Stat.	Mix	https://github.com/salesforce/DialogStudio

Table 4. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
SK-TOD [207]	Subjective-Knowledge Task-Oriented Dialogue (SK-TOD) dataset contains subjective knowledge-seeking dialogue contexts and manually annotated responses grounded in subjective knowledge sources. SK-TOD has >9K instances consisting of subjective user requests and subjective knowledge-grounded responses.	Knowl.	Stat.	Mix	https://github.com/alexa/dstc11-track5
HPD [25]	Harry Potter Dialogue (HPD) dataset in English and Chinese is annotated with vital background information, including dialogue scenes, speakers, character relationships, and attributes. It has >1K dialogues.	Knowl.	Stat.	Mix	https://nuochenpku.github.io/HPD.github.io
RSD [61]	Response Selection Data (RSD) dataset where responses from multiple response generators produced for the same dialog context are manually annotated as appropriate (positive) and inappropriate (negative). The data has 100K interactiona and 2.5 million turns.	Knowl.	Stat.	Mix	upon request/unknown
COMET [88]	A new task-oriented dialog dataset COMET, which contains 11.5K user-assistant dialogs (totalling 103K utterances), grounded in simulated personal memory graphs.	Knowl.	Stat.	Mix	https://github.com/facebookresearch/comet_memory_dialog
Augmented MultiWOZ 2.1 [81]	An augmented version of MultiWOZ 2.1, which includes new out-of-API-coverage turns and responses grounded on external knowledge sources. The dataset contains >10K dialogues with >9K augmented turns.	Knowl.	Stat.	Mix	upon request/unknown
MGConvRex [187]	A new Memory Graph (MG) - Conversational Recommendation parallel corpus called MGConvRex with 7K+ human-to-human role-playing dialogs, grounded on a large-scale user memory bootstrapped from real-world user scenarios.	Knowl.	Stat.	Mix	upon request/unknown
Annotated Weights Task Dataset [75]	A dataset of multimodal interactions in a shared physical space with speech transcriptions, prosodic features, gestures, actions, and facets of collaboration (based on the Weights Task).	Multi	Dyn.	Cont.	https://github.com/csu-signal/Common-Ground-detection
J-CRe3 [172]	A Japanese Conversation dataset for Real-world Reference Resolution (J-CRe3) that contains video and dialogue audio of real-world conversations between two people acting as a master and an assistant robot at home. The dataset is annotated with crossmodal tags between phrases in the utterances and the object bounding boxes in the video frames. These tags include indirect reference relations, such as predicate-argument structures and bridging references as well as direct reference relations.	Multi	Dyn.	Cont.	https://github.com/riken-grp/J-CRe3
LoCoMo [114]	LoCoMo, a dataset of very long-term conversations, each encompassing 600 turns and 16K tokens on avg., over up to 32 sessions. The dialogues are grounded on personas and temporal event graphs.	Multi	Dyn.	Cont.	https://snap-research.github.io/locomo
Chinese Whispers [85]	The corpus with 34 interactions, where each subject first assembles and then instructs how to assemble IKEA furniture. The dataset has speech, eye-gaze, pointing gestures, and object movements, as well as subjective interpretations of mutual understanding, collaboration and task recall.	Multi	Dyn.	Cont.	https://www.kth.se/profile/diko/page/material
Spatial AMR and Grounded Minecraft Dataset [14]	A multimodal corpus consisting of 170 3D structure-building dialogues between a human architect and human builder in Minecraft. The data contain sentence-level and document-level annotations designed to capture implicit information, the coordinates and the spatial framework annotation ground the spatial language in the dialogues.	Multi	Dyn.	Cont.	https://github.com/cu-clear/Spatial-AMR/
OneCommon [171]	OneCommon Corpus for visual conversational grounding with 600 dialogues annotated with spatial expressions that capture predicate-argument structure, modification and ellipsis.	Multi	Dyn.	Cont.	https://github.com/Alab-NII/onecommon
The Niki and Julie Corpus [7]	The Niki and Julie corpus contains more than 600 dialogues between human participants and a human-controlled robot or virtual agent, engaged in a series of collaborative item-ranking tasks designed to measure influence. Some of the dialogues contain deliberate conversational errors by the robot, designed to simulate the kinds of conversational breakdown that are typical of present-day automated agents. Data collected include audio and video recordings, the results of the ranking tasks, and questionnaire responses; some of the recordings have been transcribed and annotated for verbal and nonverbal feedback.	Multi	Dyn.	Cont.	upon request/unknown
REX Corpora [163]	A collection of multimodal corpora of referring expressions, the REX corpora. The corpora include time-aligned extra-linguistic information such as participant actions and eye-gaze on top of linguistic information, also the dialogues were collected with various configurations in terms of the puzzle type, hinting and language. The REX corpora contain 226 dialogues.	Multi	Dyn.	Cont.	upon request/unknown
GreThE [128]	GreThE, the Greek Theatrical Emotion dataset, a publicly available data collection for speech emotion recognition in Greek theatrical plays. The dataset contains 500 utterances that have been annotated in terms of their emotional content (valence and arousal).	Multi	Mix	Cont.	https://github.com/magcil/GreThE
Memory Dialog [125]	A corpus of memory grounded conversations, which comprises human-to-human role-playing dialogues given synthetic memory graphs with simulated attributes and connections to real entities (e.g. locations, events, public entities).	Multi	Mix	Cont.	upon request/unknown

Table 5. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
FUSE [166]	FrUstration and Surprise Expressions (FUSE) is a multimodal corpus for expressive task-based spoken language and dialogue, focusing on language use under frustration and surprise.	Multi	Stat.	Cont.	https://fusecorpus.github.io/FUSE/
MPCHAT [2]	MPCHAT is the first multimodal persona-based dialogue dataset which extends persona with both text and images to contain episodic memories. It contains 15K dialogues sourced from Reddit.	Multi	Stat.	Cont.	https://github.com/ahnjaewoo/MPCHAT
NICE [24]	Neural Image Commenting with Empathy (NICE) dataset consists of almost two million images and the corresponding human-generated comments with a set of human annotations. The dataset can be used to generate dialogues grounded in a user-shared image with increased emotion and empathy while minimizing offensive outputs.	Multi	Stat.	Cont.	https://nicedataset.github.io/
SIMMC 2.0 [87]	A dataset for Situated and Interactive Multimodal Conversations, SIMMC 2.0, which includes 11K task-oriented user-assistant dialogs (117K utterances) in the shopping domain, grounded in immersive and photo-realistic scenes.	Multi	Mix	Dom.	https://github.com/facebookresearch/simmc2
HybriDialogue [129]	A dialogue dataset, HybriDialogue, which consists of crowdsourced natural conversations grounded on both Wikipedia text and tables. The conversations are created through the decomposition of complex multi-hop questions into simple, realistic multiturn dialogue interactions.	Multi	Stat.	Dom.	https://github.com/entitize/HybridDialogue
KOMODIS [48]	Knowledgable and Opinionated MOvie DIScussions (KOMODIS) is a labeled dialogue dataset in the domain of movie discussions, where every dialogue is based on pre-specified facts and opinions. It contains >7K dialogues and >103K utterances.	Multi	Stat.	Dom.	https://github.com/fabiangal/komodis-dataset
SIMMC [126]	Situated Interactive MultiModal Conversations (SIMMC) is a dataset with 13K human-human dialogs (169K utterances) collected using a multimodal Wizard-of-Oz (WoZ) setup, on two shopping domains: (a) furniture – grounded in a shared virtual environment; and (b) fashion – grounded in an evolving set of images. Data include multimodal context of the items appearing in each scene, and contextual NLU, NLG and coreference annotations.	Multi	Dyn.	Mix	https://github.com/facebookresearch/simmc
RED [177]	Reddit Emotional Distress (RED) is a large-scale dialogue dataset that contains 1.3M peer support dialogues spanning across more than 4K distress-related topics.	Other	Dyn.	Cont.	https://github.com/yehchunhung/EPIMEED
MDMD [206]	A multi-label dialogue malevolence detection (MDMD) dataset where a dialogue response is considered malevolent if it is grounded in negative emotions, inappropriate behavior, or an unethical value basis in terms of content and dialogue acts. MDMD contains >8K utterances.	Other	Dyn.	Cont.	https://github.com/repozhang/malevolent_dialogue
Dynamic OneCommon [170]	A large-scale dataset of 5617 dialogues to enable fine-grained evaluation, using complex spatio-temporal expressions to create and maintain common ground over time in dynamic environments.	Other	Dyn.	Cont.	https://github.com/Alab-NII/dynamic-onecommon
HuRDL [53]	The Human-Robot Dialogue Learning (HuRDL) corpus is a dialogue corpus with 22 dialogues and 1122 turns collected in an online interactive virtual environment in which human participants play the role of a robot performing a collaborative tool-organization task. The data can be used to improve question generation in situated intelligent agents.	Other	Dyn.	Cont.	https://github.com/USArmyResearchLab/ARL-HuRDL
ESConv [103]	Emotion Support Conversation dataset (ESConv) with rich annotation (especially support strategy) in a help-seeker and supporter mode for 1K dialogues.	Other	Dyn.	Cont.	https://github.com/thu-coai/Emotional-Support-Conversation
CreST [52]	A corpus of spontaneous, task-oriented dialogue (CReST corpus), which was annotated for disfluencies and conversational moves that can facilitate grounding and coordination.	Other	Dyn.	Cont.	upon request/unknown
EmpatheticDialogues [143]	EmpatheticDialogues is a dataset of 25K conversations grounded in emotional situations, the data were gathered from 810 different participants.	Other	Mix	Cont.	https://parl.ai/
ProsocialDialog [78]	The ProsocialDialog dataset consists of 58K dialogues, with 331K utterances, and 497K dialogue safety labels accompanied by free-form rationales. It can be used for generating more socially acceptable dialogues grounded in social norms.	Other	Stat.	Dom.	https://hyunw.kim/prosocial-dialog
BSBT [80]	Blended Skill BotsTalk (BSBT), a large-scale multi-skill dialogue dataset comprising 300K conversations where agents are grounded to the specific target skills.	Other	Stat.	Dom.	https://github.com/convei-lab/BotsTalk
JIC [132]	Journal Intensive Conversations (JIC) is a journal-based conversational dataset with around 400,000 dialogues and a framework for generating personalized conversations using long-form journal entries from Reddit. The data capture common personality traits — openness, conscientiousness, extraversion, agreeableness, and neuroticism — ensuring that dialogues authentically reflect an individual's personality.	Persona	Mix	Cont.	https://github.com/Sayantan-world/Beyond-Discrete-Personas
KBP [173]	A personalized knowledge-grounded dialogue dataset Knowledge Behind Persona (KBP) is the first to consider the dependency between persona and implicit knowledge. It comes with >2K dialogues grounded in persona and knowledge.	Persona	Stat.	Cont.	https://github.com/ruleGreen/SAFARI

Table 6. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
LiveChat [50]	The LiveChat dataset is composed of 1.33 million real-life Chinese dialogues with almost 3800 average sessions across 351 personas and fine-grained profiles for each persona representing multi-party conversations.	Persona	Stat.	Cont.	https://github.com/gaojingsheng/LiveChat
PersonaMinEdit [180]	The PersonaMinEdit dataset is derived from PersonaChat with multiple human references for the edited response, it can be used to evaluate persona-grounded minimal editing.	Persona	Stat.	Cont.	https://github.com/thu-coai/grounded-minimal-edit
MaLP [200]	The dataset contains 11K dialogues, it is based on an open-source medical corpus and can help with building personalized medical assistants. The dataset is focusing on medical scenarios, including domain and commonsense information as well as personal details (e.g., chronic diseases, dialogue preferences).	Persona	Mix	Dom.	https://github.com/MatthewKKai/MaLP
PeaCoK [51]	A large-scale persona commonsense knowledge graph, PeaCoK, contains 100K human-validated persona facts. It formalizes five common aspects of persona knowledge: characteristics, routines and habits, goals and plans, experiences, and relationships.	Persona	Stat.	Mix	https://github.com/Silin159/PeaCoK
Persona-Chat [203]	Persona-Chat is a crowd-sourced dataset, collected via Amazon Mechanical Turk, where each of the pair of speakers condition their dialogue on a given profile, which is provided. The dataset is based on 1155 possible personas and provides 11K dialogues.	Persona	Stat.	Mix	https://github.com/facebookresearch/ParlAI/tree/master/projects/personachat
RealPersonaChat [190]	RealPersonaChat (RPC) corpus is based on collecting the actual personality traits and personas of interlocutors and having them freely engage in dialogue. This corpus contains 14K dialogues in Japanese and represents one of the largest corpora of dialogue data annotated with personas and personality traits.	Persona	Stat.	Mix	https://github.com/nu-dialogue/real-persona-chat
VSTAR [174]	Video-grounded Scene &Topic AwaRe dialogue (VSTAR) dataset is a large scale video-grounded dialogue understanding dataset based on 395 TV series. It contains annotations for scene and topic transitions. VSTAR contains 185K dialogues.	Visual	Dyn.	Cont.	https://vstar-benchmark.github.io/
SIMMC-2.0 [184]	SIMMC-2.0 is a video-grounded task-oriented dialog dataset that captures real-world AI-assisted user scenarios in virtual reality. It contains fine-grained and scene-grounded annotations for 4K dialogues.	Visual	Dyn.	Cont.	https://github.com/patrick-tssn/VSTAR
DVD [94]	A Diagnostic Dataset for Video-grounded Dialogue (DVD) was designed to contain minimal biases and has detailed annotations for the different types of reasoning over the spatio-temporal space of video. Dialogues were synthesized over multiple question turns, each of which was injected with a set of cross-turn semantic relationships. DVD was built from 11K CATER synthetic videos and contains 10 instances of 10-round dialogues for each video, resulting in more than 100K dialogues and 1M question-answer pairs.	Visual	Dyn.	Cont.	https://github.com/facebookresearch/DVDialogues
VFD [71]	A visually-grounded first-person dialogue (VFD) dataset with verbal and non-verbal responses. The VFD dataset provides manually annotated (1) first-person images of agents, (2) utterances of human speakers, (3) eye-gaze locations of the speakers, and (4) the agents' verbal and non-verbal responses. For the verbal response selection task, VFD dataset has almost 600K dialogues. For the non-verbal response selection task it contains around 160K dialogues.	Visual	Dyn.	Cont.	https://randd.yahoo.co.jp/en/softwaredata
PhotoBook [59]	The dataset was collected through a collaborative game prompting two online participants to refer to images utilising both their visual context as well as previously established referring expressions. This resulted in 2500 annotated dialogues.	Visual	Dyn.	Cont.	https://dmg-photobook.github.io/
CoDraw [79]	This dataset is based on a Collaborative image-Drawing game between two agents, called CoDraw. The game is grounded in a virtual world that contains movable clip art objects and involves two players: a Teller and a Drawer. The Teller sees an abstract scene containing multiple clip art pieces in a semantically meaningful configuration, while the Drawer tries to reconstruct the scene on an empty canvas using available clip art pieces. The two players communicate with each other using natural language. The CoDraw dataset contains 10K dialogs with 138K messages exchanged between human players.	Visual	Dyn.	Cont.	https://github.com/facebookresearch/CoDraw
CLEVR-Dialog [86]	CLEVR-Dialog is a large diagnostic dataset for studying multi-round reasoning in visual dialog. The dialog grammar is grounded in the scene graphs of the images from the CLEVR dataset. This combination results in a dataset where all aspects of the visual dialog are fully annotated. In total, CLEVR-Dialog contains 5 instances of 10-round dialogs for about 85K CLEVR images, totaling to 4.25M question-answer pairs.	Visual	Dyn.	Cont.	https://github.com/satwikkottur/clevr-dialog
Twitch-FIFA [136]	The Twitch-FIFA dataset is a video-context, many-speaker dialogue dataset based on live-broadcast soccer game videos and chats from Twitch.tv. It is based on 49 FIFA-18 game videos along with their users' chat. The dataset provides the triples with video context, chat context, and response data.	Visual	Dyn.	Cont.	https://github.com/ramakanth-pasunuru/video-dialogue

Table 7. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
GuessWhat?! [37]	The goal of the GuessWhat?! game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the task. The dataset consists of 150K human-played games with a total of 800K visual question-answer pairs on 66K images.	Visual	Dyn.	Cont.	https://guesswhat.ai/download
MeetUp! [66]	MeetUp! is a two-player coordination game where players move in a visual environment, with the objective of finding each other. To do so, they must talk about what they see, and achieve mutual understanding. The collected data includes 5695 annotated turns.	Visual	Dyn.	Cont.	https://github.com/clp-research/meetup
Visually Grounded Follow-up Questions [42]	A dataset of questions that require grounding both on the visual input and the dialogue history. The dataset is based on GuessWhat?! And focuses on the follow-up questions that require multimodal grounding, such questions can be extracted by identifying patterns of trigger-zoomer questions where trigger restricts the context and zoomers are spatial questions that requires triggers to be answered first.	Visual	Dyn.	Cont.	https://github.com/tianaidong/2021SpLU-RoboNLP-VISPA
PentoRef [195]	PentoRef is a corpus of task-oriented dialogues collected in systematically manipulated settings. The corpus is multilingual, with English and German sections, and overall comprises more than 20K utterances. The dialogues are fully transcribed and annotated with referring expressions mapped to objects in corresponding visual scenes, which makes the corpus a rich resource for research on spoken referring expressions in generation and resolution. The corpus includes several sub-corpora that correspond to different dialogue situations where parameters related to interactivity, visual access, and verbal channel have been manipulated in systematic ways.	Visual	Dyn.	Cont.	https://github.com/clp-research/pentoref
Image-Chat [153]	Image-Chat consists of 202K dialogues over 202K images using 215 possible style traits. It is a dataset of grounded human-human conversations, where speakers are asked to play roles given a provided emotional mood or style, as the use of such traits is also a key factor in engagingness	Visual	Mix	Cont.	http://parl.ai/projects/image_chat
VisdialConv [1]	VisdialConv is a subset of the VisDial validation set consisting of 97 dialogs, where the crowd-workers identified single turns (with dense annotations) requiring historical information. The crowd-workers were asked whether they could provide an answer to a question given an image, without showing them the dialog history.	Visual	Mix	Cont.	https://github.com/shubhamagarwal92/visdialconv-amt
IGC [127]	Image Grounded Conversations (IGC) is a dataset in which natural-sounding conversations are generated about a shared image. This is a multiple reference dataset of crowd-sourced, event-centric conversations on images, where visual grounding constrains the topic of conversation. It contains >4K conversations.	Visual	Mix	Cont.	https://www.microsoft.com/en-us/download/details.aspx?id=55324&751be11f-ede8
MMChat [211]	MMChat is a large scale Chinese multi-modal dialogue corpus (32.4M raw dialogues and 120.84K filtered dialogues). MMChat contains image-grounded dialogues collected from real conversations on social media.	Visual	Stat.	Cont.	https://github.com/silverriver/MMChat
Region-under-Discussion for Visual Dialog [117]	A subset of the Guesswhat?! questions for which their dialog history completely changes the responses. Natural language understanding grounded in vision.	Visual	Stat.	Cont.	https://github.com/mmazuecos/Region-under-discussion-for-visual-dialog
BURCHAK [193]	A human-human dialogue dataset for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner. The dataset contains 177 dialogues (each about one visual object) with a total of 2454 turns.	Visual	Dyn.	Mix	https://service.tib.eu/ldmservice/dataset/burchak-corpus