Maintaining Common Ground in Dynamic Environments

Takuma Udagawa[1] **and Akiko Aizawa**[1,2]
The University of Tokyo, Tokyo, Japan[1]
National Institute of Informatics, Tokyo, Japan[2]
{takuma_udagawa,aizawa}@nii.ac.jp

Abstract

Common grounding is the process of creating and maintaining mutual understandings, which is a critical aspect of sophisticated human communication. While various task settings have been proposed in existing literature, they mostly focus on creating common ground under a static context and ignore the aspect of *maintaining* them overtime under *dynamic* context. In this work, we propose a novel task setting to study the ability of both creating and maintaining common ground in dynamic environments. Based on our minimal task formulation, we collected a large-scale dataset of 5,617 dialogues to enable fine-grained evaluation and analysis of various dialogue systems. Through our dataset analyses, we highlight novel challenges introduced in our setting, such as the usage of complex *spatio-temporal expressions* to create and maintain common ground. Finally, we conduct extensive experiments to assess the capabilities of our baseline dialogue system and discuss future prospects of our research.

1 Introduction

Common grounding is the process of creating, repairing, and updating mutual understandings (i.e., *common ground*), which is a critical aspect of sophisticated human communication (Clark, 1996). Humans can *create* substantial common ground by expressing various information in natural language, which can be clarified or *repaired* to resolve misunderstandings at essential levels of detail. Furthermore, as the situation changes and relevant information gets outdated, humans can *update* their common ground accordingly by discarding old information and acquiring new ones. Such ability plays a vital role in sustaining collaborative relationships and adapting to emerging problems in nonstationary, real-world environments.

However, despite the wide variety of tasks proposed in existing literature (Fang et al., 2015; Zarrieß et al., 2016; De Vries et al., 2017; Udagawa and Aizawa, 2019; Haber et al., 2019), they mostly focus on creating common ground under *static* (time-invariant) context and ignore their *dynamic* aspects. While some recent dialogue tasks deal with dynamic information, they often lack suitable evaluation metrics (Pasunuru and Bansal, 2018), *context updates* in the course of the dialogue (Alamri et al., 2019), or diverse dynamics of the environment itself (De Vries et al., 2018; Suhr et al., 2019; Narayan-Chen et al., 2019; Thomason et al., 2019; Moon et al., 2020). Therefore, it remains unclear how well existing dialogue systems can adapt to the diversely changing situations through advanced common grounding (§2).

To address this problem, we propose a novel dialogue task based on three design choices (§3):

First, we formulate a novel *sequential collaborative reference task* as a temporal generalization of the collaborative reference task proposed in He et al. (2017) and Udagawa and Aizawa (2019). In our formulation, the goal of the agents is generalized to track and select the common entity *at multiple timesteps*, while the agents' observations change dynamically between each timestep. This setting requires both *creation* and *maintenance* of common ground, while enabling clear evaluation based on the length of successful timesteps.

Secondly, we focus on synthesizing the *entity movements*, as popularized in the recent video understanding benchmarks (Girdhar and Ramanan, 2020; Yi et al., 2020; Bakhtin et al., 2019). By leveraging such synthetic dynamics, we can minimize undesirable biases, maximize diversity, and enable fully controlled evaluation and analysis.

Finally, we build upon the OneCommon Corpus (Udagawa and Aizawa, 2019) to introduce natural difficulty of common grounding with minimal task complexity. To be specific, we represent entity attributes and their temporal dynamics based on *continuous* real values to introduce high ambiguity and uncertainty. In addition, we consider

Transactions of the Association for Computational Linguistics, vol. 9, pp. 995–1011, 2021. https://doi.org/10.1162/tacl_a_00409
Action Editor: Michel Galley. Submission batch: 1/2021; Revision batch: 4/2021; Published 9/2021.
© 2021 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

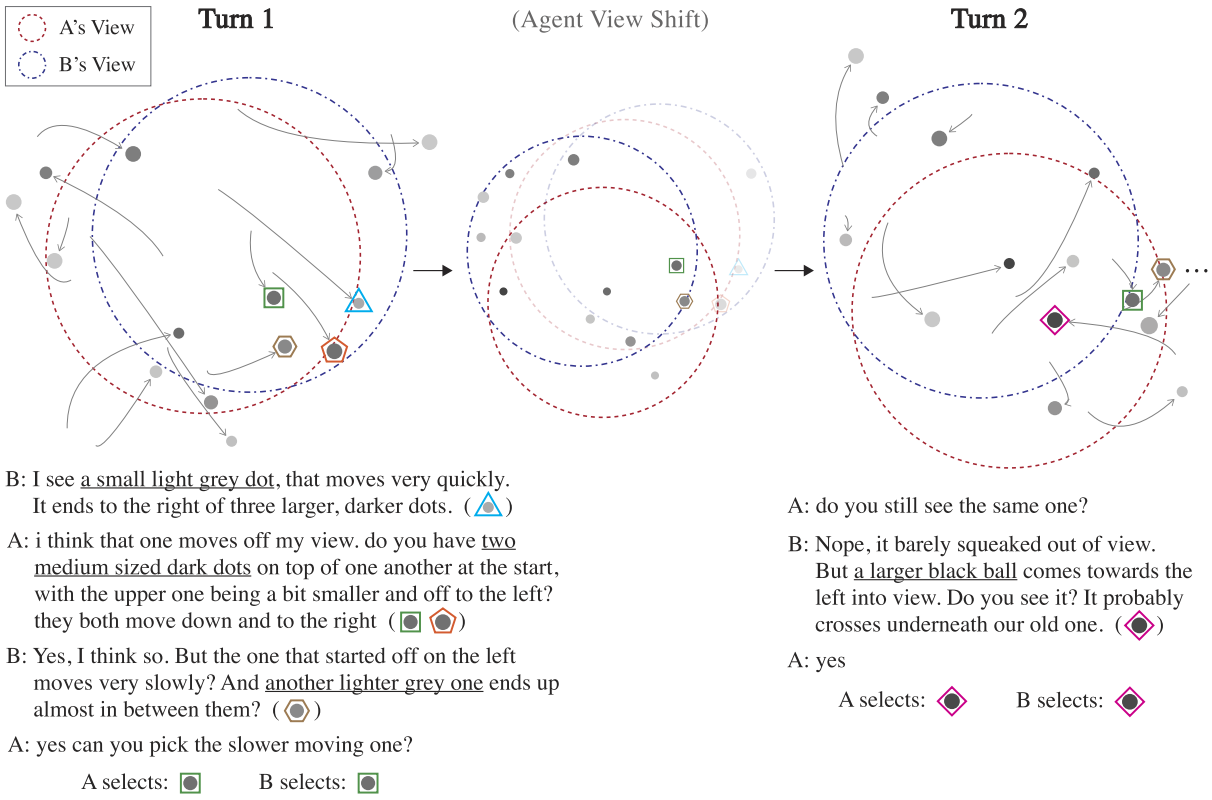


Figure 1: Example dialogue of our sequential collaborative reference task (§3). Each agent has a partial view of a 2-D plane with synthetic entities (grayscale dots of various sizes). *During* each turn, the entities move randomly on the 2-D plane. *At the end* of each turn, the agents communicate with each other to find and select one of the same, common entities. *After* each turn (if the selections match), both agents' views shift randomly and the next turn begins. Note that the colored polygons (indicating the referents of the underlined expressions) are shown for illustration purposes only and not visible to the agents nor provided in the current dataset.

a *partially observable* setting where each agent only has a partial view of the environment, which introduces various misunderstandings and partial understandings that need to be resolved.

Based on this task design, we collected a large-scale dataset of 5,617 dialogues (including over 65K utterances) through careful crowdsourcing on Amazon Mechanical Turk (§4).

We show an exemplary dialogue of our task in Figure 1. Since the environment is dynamic, humans rely on various *spatio-temporal expressions* to express entity states at different timesteps (''started off on the left'', ''ends to the right'') or how they changed dynamically (''moves very quickly'', ''come towards the left'') to create common ground. Furthermore, in later turns, humans often leverage their *previous common ground* (''still see the same one?'', ''crosses underneath our old one'') to update their common ground more reliably and efficiently. We conduct detailed analyses of the dataset to study such strategies in §5.

In our experiments (§6), we train a neural-based dialogue system based on Udagawa and Aizawa (2020). Through our extensive evaluation and analysis, we assess the current model's strengths as well as important limitations and demonstrate huge room left for further improvement.

Overall, our main contributions are:

- Proposal of a novel dialogue task to study common grounding in dynamic environments.
- Large-scale dataset of 5,617 dialogues to develop and test various data-driven models.[1]
- Detailed dataset analyses that highlight novel challenges introduced in our setting.
- Extensive evaluation and analysis of a simple yet strong baseline dialogue system.

[1]Our code and dataset are publicly available at `https://github.com/Alab-NII/dynamic-onecommon`.

Dataset	Environment (Context Type)			Context Update	Context Source	Evaluation of Common Grounding
	Continuous	Partially Observable	Dynamic			
Twitch-FIFA (Pasunuru and Bansal, 2018)	✓	✗	✓	✓	Synthetic	N/A
AVSD (Alamri et al., 2019)	✓	✓	✓	✗	Real	Indirect
SIMMC (Moon et al., 2020)	✓	✗	✗	✓	Synthetic+Real	Indirect
MutualFriends (He et al., 2017)	✗	✓	✗	✗	Synthetic	Create
GuessWhat?! (De Vries et al., 2017)	✓	✗	✗	✗	Real	Create
Photobook Dataset (Haber et al., 2019)	✓	✓	✗	✓	Real	Create
OneCommon (Udagawa and Aizawa, 2019)	✓	✓	✗	✗	Synthetic	Create
Dynamic-OneCommon (Ours)	✓	✓	✓	✓	**Synthetic**	**Create+Maintain**

Table 1: Comparison with the major datasets. Environments are considered *dynamic* if they involve rich, spontaneous dynamics and contexts to be *updated* if new information is provided in the course of the dialogue.

2 Related Work

The notion of common ground was originally introduced in Lewis (1969) and Stalnaker (1978) and theoretically elaborated in fields such as psycholinguistics (Clark and Brennan, 1991; Brennan et al., 2010). While formal approaches (rule/logic-based) exist to computationally model the process of common grounding (Traum, 1994; Van Ditmarsch et al., 2007; Poesio and Rieser, 2010), capturing their full complexities in realistic, situated conversations remains a formidable problem.

From an empirical perspective, various dialogue tasks have been proposed to develop and evaluate data-driven models of common grounding. Most of the existing literature focuses on closed domain, goal-oriented settings to measure the ability both quantitatively and objectively (Fang et al., 2015; Zarrieß et al., 2016; De Vries et al., 2017). Recent works, summarized as the *grounded agreement games* in Schlangen (2019), introduce symmetric speaker roles to encourage more bilateral interaction. Udagawa and Aizawa (2019) also raise *continuous* and *partially observable* context to be essential for requiring advanced common grounding (§3.1). Finally, Haber et al. (2019) propose a multi-round image identification task, where different combinations of images are provided to each agent at every round. While this setting is useful for studying *subsequent references* affected by the existing common ground (Brennan and Clark, 1996; Takmaz et al., 2020), the observations in each round are static, temporarily independent images. Hence, all of these tasks focus on creating common ground under *static* context and lack evaluation metrics for *maintaining* common ground in dynamic environments.

We also note that some recent dialogue tasks require dealing with dynamic information, although common grounding usually takes place *implicitly* and may be difficult to measure directly. For instance, Alamri et al. (2019) proposed Q&A-based dialogues grounded in video contexts. However, the information given to each agent remains fixed throughout the dialogue, requiring *creation* but minimal *update* of common ground. Many recent works also focus on dialogues grounded in external environments (De Vries et al., 2018; Suhr et al., 2019; Narayan-Chen et al., 2019; Thomason et al., 2019; Moon et al., 2020). These settings often involve dynamic change of the *perspectives*, but they usually assume the environments themselves to be stationary and do not change spontaneously (without direct intervention). In contrast to these works, we introduce both *context updates* in the course of the dialogue and *diverse dynamics* of the external environment to require advanced common grounding.[2] We summarize our comparison with the major existing datasets in Table 1.

Finally, our work is relevant to the emerging literature on spatio-temporal grounding in computer vision and NLP. This includes video QA (Lei et al., 2018; Yu et al., 2019; Castro et al., 2020), video object grounding (Zhou et al., 2018; Chen et al., 2019; Sadhu et al., 2020), and video captioning (Krishna et al., 2017a), all of which are essential subtasks in our dialogue. However, existing resources often contain exploitable biases and lack visual/linguistic diversity as well as reliable evaluation metrics (especially in language generation) (Aafaq et al., 2019). It is also challenging to probe model behaviors without the controllability of the video contexts (Girdhar and Ramanan,

[2]While Pasunuru and Bansal (2018) collected live-stream dialogues grounded in soccer video games, the non-goal-oriented, unconstrained nature of their setting makes evaluation and analysis of common grounding very challenging.

2020). We have addressed such concerns based on our task design (§3.2) and expect our resource to be useful for promoting this line of research as well.

3 Task Formulation

In this section, we review the collaborative reference task from OneCommon Corpus (OCC in short) and formulate our *sequential* counterpart as its temporal generalization.

3.1 Collaborative Reference Task

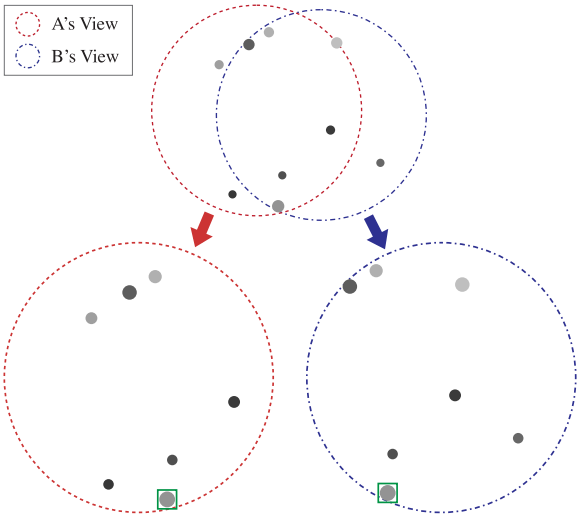
Based on Udagawa and Aizawa (2019), a *collaborative reference task* is a multi-agent cooperative game with entities $E = \{e_1, e_2, \ldots, e_m\}$ and agents $A = \{a_1, a_2, \ldots, a_n\}$. Each agent $a_j \in A$ has an observation of entities $obs_j(E)$ and can exchange information with other agents in natural language. At the end of the game, each agent selects one of the observable entities, and the game is *successful* if and only if all the agents selected the same entity.[3] This can be considered as a general framework for evaluating accurate *mutual recognition* of a common entity, which is often a critical step in general common grounding.

One main feature of OCC is that it represented all entity attributes (color, size and location on a 2-D plane) based on *continuous* real values. Unlike discrete/categorical attributes, this introduces high ambiguity and uncertainty to be expressed in symbolic natural language. In addition, OCC introduced *partial-observability*, where each agent only has a partial view of the 2-D plane, which requires collaborative resolution of various misunderstandings. We show an example of a successful dialogue from OCC in Figure 2.

However, this current task formulation assumes each observation to be *static* and can only evaluate the ability of *creating* common ground.

3.2 Sequential Collaborative Reference Task

To address this limitation, we generalize each observation to be *dynamic* and collaborative reference to be *sequential*. Specifically, each agent $a_j \in A$ now receives observation $obs_j(E, t)$ at each timestep $t \in [t_0, \infty)$, and the agents' goal is to communicate in natural language to select the same entity at multiple timesteps $t_1, t_2, \ldots \in (t_0, \infty)$.[4] At each selection timestep t_k ($k \in \mathbb{N}$), a_j must select one entity observable at t_k but has all previous observations up to t_k, $\{obs_j(E, t) | t \in [t_0, t_k]\}$. The game ends when the selections no longer match at timestep $t_{k'}$ ($k' \in \mathbb{N}$): Therefore, the success at t_1 measures the ability of *creating* common ground, and the length of successful timesteps (LST) $k' - 1$ measures the ability of *maintaining* them. This is a general framework for evaluating both creation and maintenance of mutual entity recognition in dynamic environments.



A: I see three in a line going up and to the right. The middle one is the largest and darkest
B: I don't see that. I have one large, medium gray dot that's under a small, darker gray dot
A: Is the larger dot slightly to the left
B: yes, slightly, let's choose the larger one
A selects:
B selects:

Figure 2: Example dialogue from OneCommon Corpus (OCC). We can see that the human players are able to detect misunderstandings and make flexible clarifications to reduce ambiguity and uncertainty.

Based on this task formulation, we propose a minimal task setting extending OCC and incorporate dynamic change of the entity *locations*.

We refer to each time range $[t_{k-1}, t_k]$ as *turn* k. During each turn, we change the location of each entity $e_i \in E$ based on a simple parameterized movement, where the *trajectory* is determined

[3] In contrast to the typical reference tasks (De Vries et al., 2017), agent roles are *symmetric* and they can agree upon any of the common entities (as long as it's the same).

[4] We assume $t_{k-1} < t_k$ for all $k \in \mathbb{N}$.

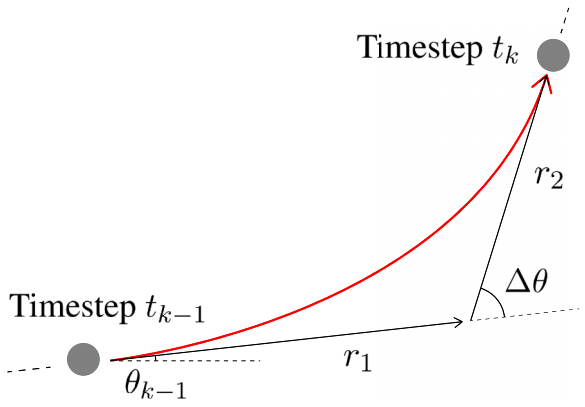


Figure 3: Illustrated movement of each entity in turn k.

by a quadratic Bézier curve (Bézier, 1974).[5] See Figure 3 for an illustration, where r_1, r_2 are parameters of *distance* and θ_{k-1}, $\Delta\theta$ represent *angles*. We sample r_1, r_2, $\Delta\theta$ from fixed uniform distributions each turn and update θ_k as $\theta_k \leftarrow \theta_{k-1} + \Delta\theta$ (θ_0 is initialized randomly). This way, we can generate diverse, unbiased, coherent, and fully controllable dynamics of the environment.

To enable fair comparison with OCC, we limit the number of agents to 2 and set the circular agent views to have the same diameter as OCC. At each selection timestep t_k, we ensure that each agent has 7 observable entities with only 4, 5, or 6 of them in common, which is also identical to OCC. Finally, we sample all entity attributes (color, size, and initial location) from the same uniform distributions as OCC with minimal modifications.[6] Therefore, we expect the (distribution of) observations at t_k to be similar and enable mostly fair comparison with OCC (in §5 and §6).

To ensure task difficulty, we also shift the *perspective* of each agent after each successful turn (see Figure 1) so that the overlapping regions differ every turn. The same dot is prohibited from staying in common for over 3 consecutive selection timesteps, requiring frequent updates of common ground. Finally, we limit the maximum number of turns to 5 for practical purposes (hence the maximum LST is 5 in each game).

[5]Its *speed* is proportional to the length of the trajectory.

[6]To be specific, we set the minimum distance between entities (at t_k) and the possible range of entity size to be slightly different to avoid entity overlapping during movements.

4 Dataset Collection

To collect large-scale, high-quality dialogues, we conducted careful crowdsourcing on Amazon Mechanical Turk. The Web application is based on the CoCoA framework (He et al., 2017), and we used Scalable Vector Graphics (SVG) to animate entity movements and parallel shifts of the agent perspectives. Before working on our task, crowd workers were required to take a brief tutorial on the task setting, dialogue interface, and instructions. Sample screenshots of our dialogue interface and tutorial are shown in Figure 4. Note that animations up to the current turn could be replayed anytime for the ease of playing the game.[7]

To ensure worker quality, we required crowd workers to have more than 500 completed HITs and acceptance rates higher than 99%. To encourage success, we rewarded \$0.25 for every successful turn plus additional bonuses for longer LST achieved (up to \$0.25 if LST $= 5$). Finally, we manually reviewed all submitted works and excluded dialogues which clearly violated the instructions (e.g., relying on premature guessing or other ineffective strategies[8]). We did not exclude dialogues based on task failures (even if LST $=$ 0), as long as they were based on valid strategies.

To solicit linguistic/strategic variety, we generally used a unique environment for each game. However, if the task was unsuccessful (i.e., LST $=$ 0), we allowed the environment to be reused in another game. This way, we can expect to eventually collect successful (LST > 0) dialogues for the relatively difficult environments as well.

Overall, we collected 5,804 dialogues, and after the reviewing process, we were left with 5,617 qualified dialogues. We refer to this dataset as **Dynamic-OneCommon Corpus (D-OCC)**. Note that our dataset is currently in English, but the dataset collection procedure is language-agnostic and can be applied in any other languages.

5 Dataset Analysis

Next, we conduct detailed analyses of the dataset to study human common grounding strategies under dynamic context. Whenever possible, we give comparative analyses with OCC to highlight the effect of dynamic factors introduced in D-OCC.

[7]This also allows us to ignore the disadvantage of imperfect human memories in comparison to machines.

[8]Typical examples include strategies relying solely on color, size, and absolute positions in the agent's view.

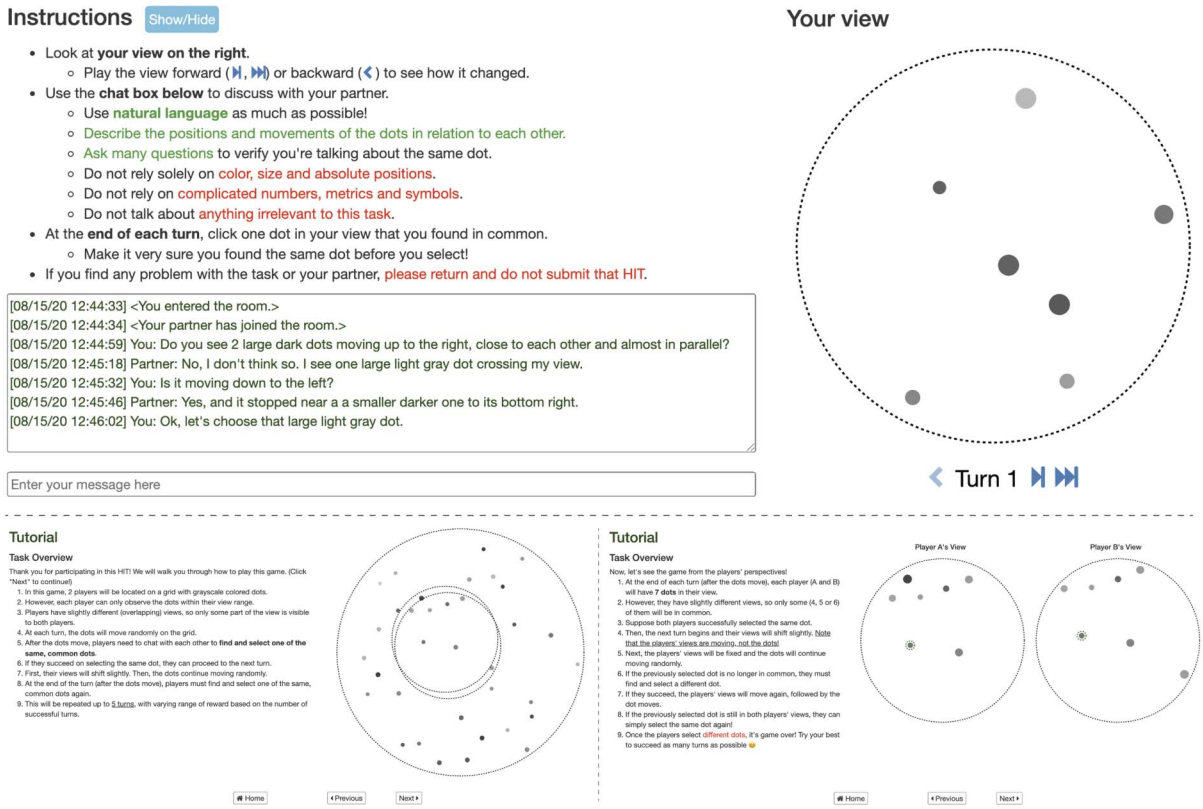


Figure 4: (Top) Our dialogue interface. During the game, animations up to the current turn could be replayed anytime using the forward/backward buttons. (Bottom) Sample screenshots from our tutorial on the *task setting*.

5.1 Overall Statistics

First, we summarize the overall statistics of OCC and D-OCC in Table 2.

In total, OCC and D-OCC have a comparable number of dialogues. However, dialogues can be much longer in D-OCC, since collaborative reference is repeated multiple times. On average, utterance lengths are slightly shorter in D-OCC; this can be mostly attributed to the increased (relative) frequency of short utterances like acknowledgments and *shortened* subsequent responses (e.g., ‘‘same again?’’ = ‘‘select the same black dot again?’’).[9] Note that long, complex utterances are also common in our dataset, as seen in Figure 1. Overall, we found 462 unique workers participated in D-OCC, which indicates reasonable diversity at the *player* level as well.

In terms of LST, the overall average was 3.31 with over half (53.5%) of the dialogues succeeding all 5 turns. This suggests that humans can solve the task reliably through sophisticated common grounding. After filtering dialogues with poor/careless workers (whose average LST < 2), we observed a slight improvement up to 3.57. If we only focus on the top 10 workers (with at least 10 tasks completed), average LST was significantly higher reaching 4.24. These results indicate that (at least potentially) much higher human ceiling performance can be achieved. Note that if we include the last unsuccessful turn in 46.5% of the dialogues, the average of all completed turns was slightly longer (3.77) in our dataset.

Statistics	OCC	D-OCC
Total dialogues	6,760	5,617
Uttrances per dialogue	4.8	11.7
Tokens per utterance	12.4	10.3
Duration per dialogue (minutes)	2.1	5.7
Unique workers	N/A	462
Avg. LST	–	3.31
Avg. completed turns	–	3.77
Unique tokens	3,621	3,895
Occupancy of rare tokens (%)	1.4	1.0
Overlap of all tokens (%)	29.4	
Overlap w/o rare tokens (%)	53.0	

Table 2: Statistics of OCC and D-OCC datasets.

[9] In fact, utterances with fewer than 5 tokens were almost twice as frequent in D-OCC (33.8%) than OCC (17.6%).

Reference	Examples	Freq.	Cohen's κ
Current State	It's to the right of where the grey one ended up for me after moving up and left. Now I have another triangle / Does it land next to two smaller gray dots? Does it have a lighter one below and to the left when they stop? Two similar shades close to each other (*implicit*)	23.8%	0.91
State Change	a small dark one traveling southwest / 2 other dots following it Do you have two dark med-size dots move slowly apart as they drift right? I have a large pale grey that moves down but starts out curving to the right and then takes a sharp turn to the south east	32.7%	0.97
Previous State	I still see the larger gray one that was next to it in the previous turn. I have the smaller dot that started out below it to the left. Before it moves, is there a lighter gray dot down and to the right of it?	5.5%	0.79

Table 3: Spatio-temporal expressions. Keywords (such as *tense*, *events*, and *motion verbs*) are underlined.

Finally, we found that both datasets have a relatively small vocabulary size as well as the occupancy of *rare tokens* (used less than 10 times in the dataset).[10] This indicates minimal complexity at the *lexical* level, as observed in Udagawa and Aizawa (2019). We also found that the two datasets have a large vocabulary overlap, which is expected as D-OCC extends the setting of OCC.

5.2 Spatio-Temporal Expressions

At the utterance level, we observed an extensive usage of *spatio-temporal expressions*, which are characteristic in dynamic environments. To study the frequency of such expressions, we manually annotated 100 dialogues in D-OCC with LST ≥ 2 (focusing on the more successful strategies).

Specifically, we detect whether each utterance contains 3 types of spatio-temporal expressions:[11]

- Reference to **current state** describes location of entities at the end of the current turn (i.e., timestep t_k if the utterance is in turn k).
- Reference to **state change** describes temporal change of entity locations (i.e., movements).
- Reference to **previous state** describes entity locations at previous timestep t (where $t < t_k$).

We show examples and estimated frequencies of spatio-temporal expressions in Table 3. We also computed the agreement of our annotation based on 50 dialogues with 3 annotators, which we found to be reliable based on Cohen's κ (Cohen, 1968).

Based on this result, we found that reference to *state change* is the most widely used strategy, which could be simple as ''moves northwest'' or more complex as in Table 3. Reference to *previous state* is much less frequent compared to other types but still observed in many dialogues. Note that humans distinguish *previous* and *current* states in various ways, including temporal expressions (''was'', ''now''), motion verbs (''started out'', ''landed''), and implicit/default reasoning.

We also found that expressions are often *nuanced* and *pragmatic*, which are characteristic under continuous and partially observable context (Udagawa and Aizawa, 2019). Nuances are typically expressed by the *degree modifiers* to convey subtle differences in location, movements, confidence, and so forth. Following Paradis (2008), we categorize them into 2 main types (and 5 subtypes): *scalar modifiers* used for concepts in a range of scale (*diminishers*, *moderators*, *boosters*) and *totality modifiers* used for concepts with definite boundaries (*approximators*, *maximizers*). See Table 4 for examples and the estimated occurrences of such modifiers in OCC and D-OCC.[12] Based on these results, we can verify that there are comparable numbers of various degree modifiers in D-OCC as well, which are used effectively to cope with complex ambiguity and uncertainty.

[10] Occupancy is computed based on the proportion of total frequencies (TF), i.e., *TF of rare tokens* / *TF of all tokens*.

[11] Note that a single utterance may contain none or multiple types of such expressions, and expressions of color, size, or possession are not considered as spatio-temporal expressions.

[12] Following the prior analysis in OCC, we manually curated keyword-based dictionaries of such modifiers (based on unigrams and bigrams) while removing polysemous words (such as *little*, *about*, *too*, etc).

Degree Modifiers		OCC	D-OCC	Examples (# Keywords)	Usage in D-OCC
Scalar	Diminishers	9.2	8.9	a bit, faintly, slightly (10)	**slightly** curves up
	Moderators	1.3	0.9	fairly, rather, somewhat (6)	**fairly** quickly
	Boosters	9.8	6.1	very, really, extraordinary (27)	**extremely** slowly
Totality	Approximators	10.2	6.4	almost, maybe, probably (34)	**almost** collides with
	Maximizers	4.3	4.2	exactly, completely, definitely (37)	**perfectly** straight

Table 4: Average occurrences of degree modifiers per 100 utterances (estimated based on keywords).

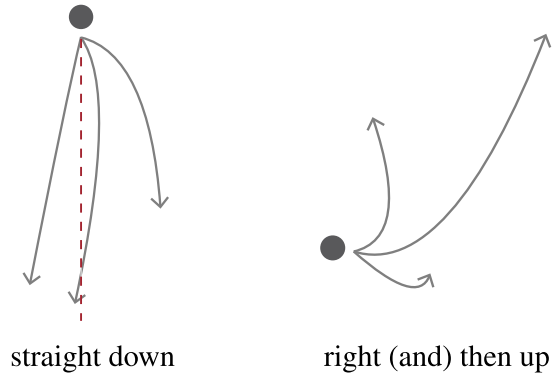


Figure 5: Pragmatic expressions of movements.

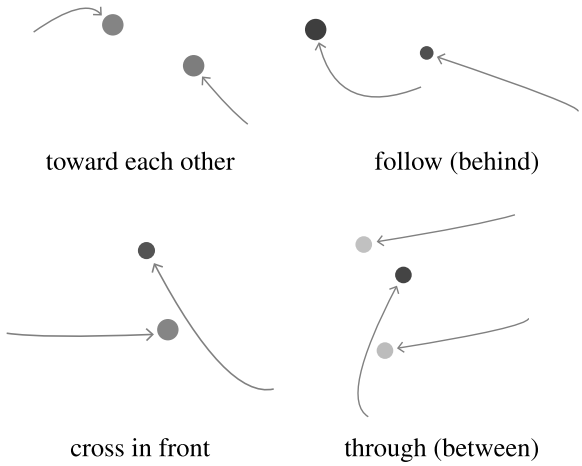


Figure 6: Expressions of multiple entity interactions.

In Figure 5, we show examples of *pragmatic* expressions that require pragmatic (non-literal) interpretations (Monroe et al., 2017). For instance, trajectories of the expression ‘‘straight down’’ may not indicate vertical lines in the literal sense (e.g., could be curving or leaning to the left). Similarly, the expression of ‘‘(moving) right and (then) up’’ may be used for diverse movements ending up in various locations (e.g., even below the initial location!). While such expressions more or less deviate from literal semantics, they are pragmatically sufficient to convey the speaker's intention (i.e., identify the target among the distractors) (Grice, 1975); alternatively, the speaker may need to choose different expressions for the same movement depending on the context (distractors).

We also show exemplary expressions of multiple entity interactions in Figure 6, which demonstrate interesting pragmaticality as well. For instance, ‘‘toward each other’’ may be used for trajectories moving in *orthogonal* (rather than opposite) directions for the most of the time.

Overall, our analyses of spatio-temporal expressions reveal advanced language understanding and generation required in D-OCC, regardless of the task/lexical simplicity.

5.3 Turn-Level Strategies

Finally, we study and compare human strategies at different timesteps (in different turns). Table 5 shows detailed statistics of the dataset in the initial turn and later turns, where *creation* and *maintenance* of common ground are required, respectively. Note that we also distinguish later turns based on whether the previous selection (i.e., *previous target*) stays in common (✓) or leaves at least one agent's view (✗): Former cases can *retain* the same common ground but the latter cases require an *update* of common ground.

First, if we focus on the 1st turn, we can verify that success rates are consistently higher in D-OCC than OCC, especially in difficult cases when the number of shared entities is smaller. This indicates that humans can create common ground more accurately by leveraging dynamic information (e.g., entity movements) unavailable in OCC.

In later turns, we found that human performance is near perfect with shorter dialogues in ✓ cases (when the previous target stays in common). This is natural because they can simply retain common

Dataset	Turn	Previous Target	Success Rate (%) #Shared=4	#Shared=5	#Shared=6	Utterances per Turn	Tokens per Utterance
OCC	1st	–	65.8	77.0	87.0	4.8	12.4
D-OCC	1st	–	73.4	82.0	87.6	3.2	11.0
	≥2st	✓	95.4	97.0	97.8	2.3	5.9
		✗	81.7	88.4	91.6	3.5	11.7

Table 5: Turn-level statistics of OCC and D-OCC. ✓ denotes cases where the previous target stays in common and ✗ denotes it left at least one agent's view. Note that # shared entities are 4, 5, or 6 at selection timesteps (§3.2).

Previous Target	Examples	Freq.
Stay (✓)	I still see the same dot / I still have all three dots from the line before Left my screen, but may have come back traveling left to right?	36.8%
Leave (✗)	I lost the last one / I lost the light one but still see the darker one that was on its left. both are gone for me / similar size black dot that barely moves? (*implicit*)	63.2%

Table 6: Comparison of utterances when the previous target stays in common (✓) or not (✗).

ground and repeat the same selection. Notably, human performance is consistently higher than the 1st turn *even in* ✗ cases (when the previous target is no longer in common), which verifies that humans can leverage previous common ground to *update* common ground more reliably as well.

We show example utterances of ✓ and ✗ cases in Table 6. Note that the previous target may temporarily leave the view and come back in ✓ cases, which occasionally makes even *retainment* of the same common ground non-trivial. In ✗ cases, humans either inform about the lost entities explicitly or *implicitly*, for example, by ignoring old entities and starting to focus on the new ones.

6 Experiments

Finally, we conduct extensive experiments to assess our baseline model's capability of common grounding in dynamic environments.

6.1 Evaluation

To study the model's capability from various aspects, we design 3 (sub)tasks based on D-OCC.

First, we evaluate the model's ability of *recognizing* common ground based on the **target selection task**, originally proposed for OCC. This is an important subtask of (sequential) collaborative reference, where the model is given one player's observation and the (ground-truth) dialogue history to predict which target was selected by the player. Since there can be multiple selections in D-OCC, the model makes predictions at the end of each turn k (at timestep t_k). The number of entities observable at t_k is fixed at 7 for both OCC and D-OCC (§3.2), so this is a simple classification task evaluated based on accuracy.

Secondly, we estimate the model's ability of *creating* and *maintaining* common ground based on the **selfplay dialogue task**, where each model plays the full sequential collaborative reference task against an identical copy of itself. While this evaluation has the advantage of being scalable and automatic, succeeding on this setting is only *necessary* for human-level common grounding and not *sufficient*, since the model may only be able to coordinate with itself (and not with real humans).

Thirdly, we conduct **human evaluation** to test the model's ability of playing sequential collaborative reference against real human workers on AMT. Due to the high cost of this evaluation, we only focus on the top 3 variants of our baseline ranked by average LST in the selfplay dialogue task.

6.2 Model Architecture

For a fair comparison with prior work, we implement our baseline model following the OCC

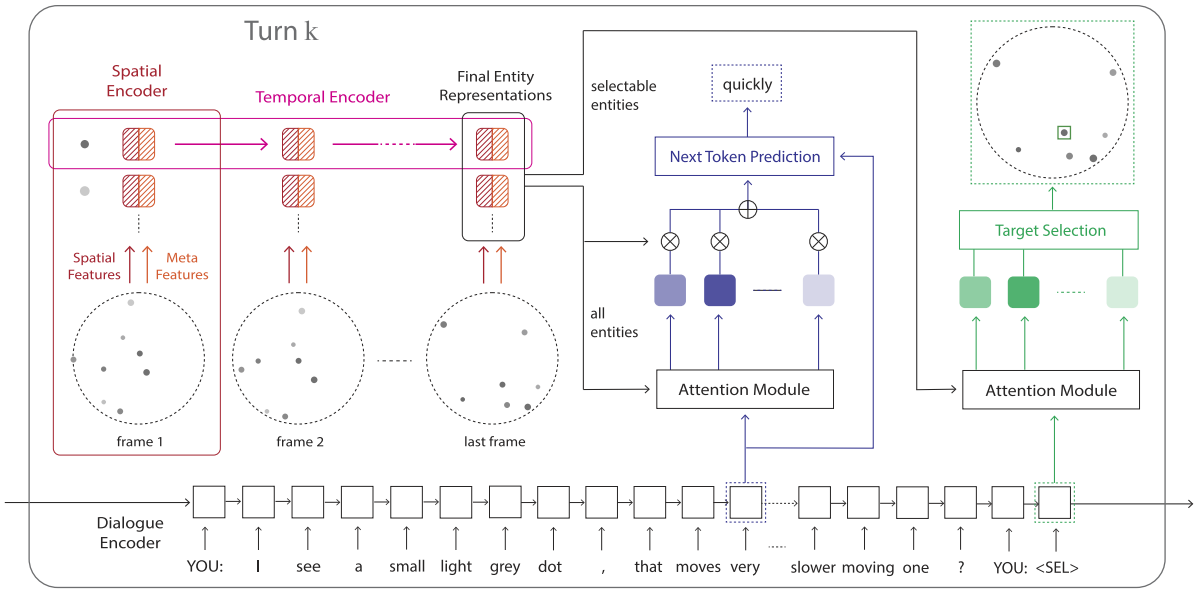


Figure 7: Our baseline model architecture. Information flow in turn k is illustrated. When generating model utterances (in selfplay dialogue and human evaluation), we sample next tokens with the temperature set to 0.25.

models in Udagawa and Aizawa (2020). The overall model architecture is shown in Figure 7.

To encode the dialogue tokens throughout the turns, we use a unidirectional GRU (Cho et al., 2014). To encode the observation during turn k, we first split the animation of entity movements into 10 frames and the agent view shift into 5 frames. Then, we process each observation frame based on the *spatial* encoder, followed by the *temporal* encoder to integrate these outputs.

The spatial encoder is used to extract *spatial features* and *meta features* from each observation frame. Spatial features represent the spatial attributes of each entity (color, size, and location in the frame), which are encoded using an MLP and a relation network (Santoro et al., 2017). The relation network is used to represent the spatial attributes relative to a subset of entities $\tilde{E} \subset E$, which could be *all entities* observable in turn k (E_{all}) or *selectable entities* visible at t_k (E_{sel}). Hence, the spatial features of e_i are computed as:

$$\mathrm{MLP}(\mathbf{e}_i) \odot \sum_{\substack{e_j \in \tilde{E}, \\ j \neq i}} \mathrm{MLP}(\mathbf{e}_i - \mathbf{e}_j) \qquad (1)$$

where $\mathbf{e}_i$ is the vector representation of entity e_i and $\odot$ is the vector concatenation.[13]

[13]To be precise, $\mathbf{e}_i$ is a 4-dimensional vector representing color, size, and 2-D location. If the entity is not observable in the frame, we use the default value of (0, 0) for the location.

Meta features are binary information of each entity representing whether (or not) the entity (i) is visible in the frame, (ii) is visible at timestep t_k, (iii) was visible at timestep t_{k-1}, and (iv) was selected in the previous turn (i.e., is the *previous target*). Meta features are also encoded using an MLP, and we take the sum of spatial/meta features as the (entity-level) output of the spatial encoder.

Finally, we use the temporal encoder based on a GRU to encode the outputs of the spatial encoder. The final state of the temporal encoder is considered as the final representation of each entity.

Based on the outputs of these encoders, we use two attention modules (based on MLPs) to compute attention scores for each entity. The first attention module is used to weight the final representations of all entities E_{all} conditioned on the current dialogue state: then, the weighted sum of E_{all} is concatenated with the dialogue state to predict the next dialogue token (Xu et al., 2015). The second module is used to predict the target entity, where we simply take the (soft)max of attention scores for the selectable entities E_{sel} in turn k.

Note that there are only two main differences between our baseline and the best OCC model (TSEL-REF-DIAL) from Udagawa and Aizawa (2020): First, in TSEL-REF-DIAL, the final representation of each entity is its *spatial features*, that is, the meta features and temporal encoder are not used (which are only meaningful in D-OCC).

Second, TSEL-REF-DIAL is also trained on the *reference resolution* task (using an additional attention module), which is only available in OCC. Due to this architectural similarity, we can virtually *pretrain* our model on OCC by initializing the shared model parameters based on TSEL-REF-DIAL and then fine-tune the whole model on D-OCC.[14]

6.3 Experiment Setup

All modules of our baseline (MLPs and GRUs) are single-layered with 256 hidden units, except for the attention modules, which are 2-layered. Dropout rate of 0.5 is applied at each layer during training, and we use the Adam optimizer (Kingma and Ba, 2015) with the initial learning rate set to 0.001. After manual tuning on the validation set, we weight the losses from next token prediction and target selection with the ratio of 2:1.

In terms of data splits, we use 500 dialogues with LST ≥ 2 for testing target selection, another 500 for validation, and the rest for training.[15] Note that we use all unsuccessful turns (where the players failed to agree upon the same entity) as well, assuming they are still based on valid strategies. For selfplay dialogue and human evaluation, we collect 2,000 and 200 dialogues in unseen environments, respectively. Each experiment is repeated 5 times with different random seeds (including data splits), except for human evaluation.

Finally, we conduct extensive ablations to study the effect of various model architectures, including *pretraining*, spatial attributes (*color*, *size*, and *location*), and the meta feature (*previous target*). In addition, we also ablate the *dynamic* information of the observation by only using the last frame in each turn as the input for the temporal encoder.

6.4 Results

We show the results for target selection in Table 7. Human performance is estimated by 3 annotators based on 50 dialogues with LST ≥ 2.

Based on these results, we can verify that all ablations hurt the performance of our baseline in some way. Pretraining on OCC is generally effective, and all spatial attributes contribute to the overall performance (especially color and size). When the meta feature of the correct previous target is available, all models perform remarkably well in ✓ cases (previous target stays in common), which is natural since humans often repeated the same selection. Finally, dynamic information also contributes to the baseline performance, despite the effect being rather marginal.

Model	Turn / Previous Target		
	1st / –	≥2nd / ✓	≥2nd / ✗
Baseline	**76.4±1.7**	96.6±0.3	67.4±0.5
– pretraining	74.6±2.7	96.3±0.7	66.9±1.1
– color	56.3±2.0	95.7±0.6	50.5±1.4
– size	58.4±1.3	95.7±0.9	52.2±0.5
– location	74.4±1.5	96.1±0.9	67.3±0.7
– previous target	76.1±1.7	83.3±1.1*	**67.8±0.6***
– dynamics	75.1±2.3	**96.7±1.0**	67.0±0.7
Human	97.0±1.1	98.2±0.5*	95.8±2.0*

Table 7: Results for the **target selection task** (* denotes cases where the correct previous targets were not provided during prediction).

However, there is huge room left for improvement in the 1st turn and even more so in ✗ cases (previous target no longer in common). These results indicate that recognizing the *creation* of common ground is still difficult, and recognizing how they are *updated* (rather than *retained*) remains even more challenging for the current baseline.

Next, we show the results for selfplay dialogue and human evaluation in Table 8. We also include the results of TSEL-REF-DIAL (trained on OCC *without* fine-tuning on D-OCC) as a reference.[16]

In selfplay dialogue, we can verify that the baseline model performs reasonably well, outperforming TSEL-REF-DIAL in the 1st turn of D-OCC (as well as OCC). However, it is worth noting that TSEL-REF-DIAL may be suffering from a minor covariate shift in D-OCC (c.f. §3.2), and without pretraining, our baseline still underperforms this best OCC model. We also found that all ablations of spatial attributes hurt performance, while the locational attributes became more critical in the full dialogue task. The meta feature of the previous target (selected by the model) is also critical, as the models seem to be relying heavily on this feature to both retain and update the target.

[14] For pretraining, we retrained TSEL-REF-DIAL with the shared word embedding for OCC and D-OCC.

[15] We ensured no overlaps in terms of the *environments* across data splits.

[16] When testing TSEL-REF-DIAL on D-OCC, we used the spatial features of the last observation frame as the input.

				Seflplay Dialogue				Human Evaluation	
Model	Dataset	Turn	Previous Target	Success Rate (%) #Shared=4	#Shared=5	#Shared=6	Avg. LST	Success Rate (%)	Avg. LST
Baseline	D-OCC	1st	–	46.8±1.8	63.8±1.8	80.2±2.3		**44.5**	
		≥2nd	✓	99.4±0.3	99.7±0.2	99.6±0.2	1.94±0.09	81.9	**1.00**
			✗	**48.5±2.2**	**64.6±2.8**	**81.5±1.5**		**44.4**	
– pretraining	D-OCC	1st	–	39.4±1.0	53.5±0.8	73.7±1.8			
		≥2nd	✓	98.6±2.4	98.8±1.8	99.4±1.0	1.35±0.09	N/A	N/A
			✗	30.3±5.7	42.1±6.3	65.4±4.9			
– color	D-OCC	1st	–	36.3±2.0	54.6±2.3	72.9±1.5			
		≥2nd	✓	99.7±0.1	99.7±0.0	99.6±0.1	1.50±0.10	N/A	N/A
			✗	42.1±3.5	56.7±4.2	72.4±4.6			
– size	D-OCC	1st	–	41.5±0.8	58.0±0.9	75.2±1.3			
		≥2nd	✓	99.8±0.1	99.7±0.1	99.8±0.2	1.58±0.07	N/A	N/A
			✗	39.6±3.5	55.3±3.6	69.9±1.5			
– location	D-OCC	1st	–	45.7±1.9	60.4±1.6	77.7±1.7		40.0	
		≥2nd	✓	99.8±0.1	99.7±0.0	99.7±0.1	1.68±0.09	**91.8**	0.81
			✗	40.8±3.6	54.6±2.5	73.9±4.2		36.3	
– previous target	D-OCC	1st	–	**49.2±1.3**	64.0±1.8	82.2±2.0			
		≥2nd	✓	85.8±2.7	87.5±1.6	91.2±1.3	1.45±0.05	N/A	N/A
			✗	29.2±1.5	41.9±1.9	64.5±1.0			
– dynamics	D-OCC	1st	–	**49.2±2.2**	**65.8±1.3**	**83.3±1.9**		37.0	
		≥2nd	✓	**99.9±0.1**	**99.9±0.1**	**99.8±0.1**	**2.02±0.07**	86.8	0.79
			✗	48.3±2.2	63.5±2.8	81.1±2.1		39.2	
TSEL-REF-DIAL	D-OCC	1st	–	41.0±1.2	58.7±1.1	76.0±1.8	–	N/A	–
	OCC	1st	–	45.9±1.6	62.7±2.2	79.7±1.0			
Human	D-OCC	1st	–	73.4	82.0	87.6		80.5	
		≥2nd	✓	95.4	97.0	97.8	3.31	96.7	3.31
			✗	81.7	88.4	91.6		86.6	

Table 8: Results for the sequential collaborative reference task (**selfplay dialogue** and **human evaluation**). Human performance is estimated based on the overall average of the crowd workers (c.f. Table 2 and 5).

However, we found that ablation of dynamic information does not degrade (actually improves) performance in selfplay dialogue. This indicates that the last frame of each turn (*current state*) is sufficient for the baseline to coordinate with itself, and it is unlikely to be leveraging sophisticated temporal information (*state change* or *previous state*) like the human strategies seen in §5.2. Also, while the models perform near perfectly in ✓ cases, the success rates drop or do not improve significantly in ✗ cases (compared with the 1st turn). This shows that current models can *retain* the same common ground easily but struggle in *updating* them using the previous common ground, unlike the human strategies seen in §5.3.

Finally, in human evaluation, we could verify that our baseline performs the best of the top 3 models in the selfplay dialogue task, but the success rates were much lower than observed in selfplay. This indicates that current models may not be using natural language in the same way humans use it (i.e., are not properly *grounded* [Bender and Koller, 2020]), although they do become closer to it when all the features are available.[17]

To summarize, our results in sequential collaborative reference show that the current baseline can leverage all spatial features and retain the same common ground, especially when provided explicitly as the meta feature. However, it may not be using temporal information effectively, and the creation and update of common ground still remain challenging in the dynamic environments, especially when conversing with real humans.

7 Discussion and Conclusion

In this work, we proposed a novel dialogue task to study the ability of creating, retaining and

[17] At the *superficial* level, all models could generate fluent utterances and complete the task with minimal confusion.

updating common ground in dynamic environments. The introduced dynamics are fully controllable in our setting to maximize diversity, minimize biases and enable reliable evaluation and analysis. Based on our dataset analyses and experiments, we demonstrated the advanced strategies of common grounding required and the open room for improvement in our newly developed Dynamic-OneCommon Corpus (D-OCC).

In future work, we plan to utilize and enrich this dataset in several ways. For instance, we can conduct various *causal analysis*, for example, by changing certain feature of entities (such as movement) and studying the differences in model behavior, which is essential yet difficult to conduct in many existing datasets (c.f. §2). Another promising direction is to add fine-grained annotation of *reference resolution* (Udagawa and Aizawa, 2020), as (partially) illustrated in Figure 1. We can also annotate *spatio-temporal expressions*, for example, by following the procedure in Udagawa et al. (2020). Such annotations would allow us to gain deeper understandings of the *intermediate* process of common grounding: For instance, we can study whether the developed models recognize and use the spatio-temporal expressions appropriately and consistently in a *human-like* way (i.e., not only imitate at the superficial level, as observed in §6.4).

In order to improve the model performance, we are considering several approaches. One approach is to make the model learn from task success (and failure) through *reinforcement learning*. Due to the symmetric agent roles in our task, this is straightforward to conduct through selfplay (Lewis et al., 2017; Yarats and Lewis, 2018), and we can expect the models to avoid ineffective strategies like underspecification and premature guessing. We also expect the incorporation of *pragmatic reasoning* to be a fruitful area of future research. One representative approach is the *Rational Speech Act* (*RSA*) framework (Goodman and Frank, 2016), which has been applied in both continuous (Monroe et al., 2017) and partially observable domains (Hawkins et al., 2021). However, application in *dynamic* domains would involve additional complexities that need to be taken into account, such as the dependencies on previous common ground. Finally, we are planning to study wider variety of model architectures and pretraining datasets, including video-processing methods (Carreira and Zisserman, 2017; Wang et al., 2018), vision-language grounding models (Lu et al., 2019; Le et al., 2020), and large-scale, open domain datasets (Krishna et al., 2017b; Sharma et al., 2018). Note that the entity-level representation of the observation (required in our baseline) can be obtained from raw video features, for example, by utilizing the object trackers (Bergmann et al., 2019; Wang et al., 2020).

Finally, we'd like to discuss the main limitation of our current work, namely, the *ecological validity* (De Vries et al., 2020) of D-OCC. Since we focused on the simplest task setting under *continuous*, *partially observable* and *dynamic* context, direct application of our work in realistic settings may not be straightforward. However, the generic strategies required in our setting are fundamental in many real-world applications. For an illustration, imagine a navigation task in a *dynamic* environment, such as finding a lost child in an urban city. Since the *target entity* (the child) may not stay in one place, routing directions can no longer be fixed and need to be *updated* accordingly (as in ''now head more to the west'' or ''go back to the previous block''). Furthermore, the *landmark entities* may not be stationary either and could be ephemeral (as in ''following the group of travelers'' or ''in the middle of the crowd''). Lastly, if the child is not conspicuous with confusable distractors (e.g., with many pedestrians around), the descriptions need to be precise and distinguishing (as in ''wearing *a little* darker shirt'' or ''walking *right* towards the station'').

In order to study such (nuanced and pragmatic) spatio-temporal expressions and references to previous common ground, we expect D-OCC to be an essential proving ground. In addition, our sequential collaborative reference task is defined generally (c.f. §3.2), so we can easily scale up the task complexity to study the desired dynamics under consideration: the exploration of different, potentially more complex dynamics is an important research area left as future work.

Overall, we expect our task design, resource, and analyses to be fundamental for developing dialogue systems that can both create and maintain common ground in dynamic environments.

Acknowledgments

We are grateful to our action editor, Michel Galley, and the three anonymous reviewers for

their valuable suggestions that helped improve this paper. We also thank Saku Sugawara and Taichi Iki for their constructive feedback on earlier versions of this paper. This work was supported by JSPS KAKENHI grant number 21H03502.

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6):1–37. https://doi.org/10.1145/3355390

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567. https://doi.org/10.1109/CVPR.2019.00774

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. 2019. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, pages 5082–5093.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. 2019. Tracking without bells and whistles. In *International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2019.00103

Pierre Bézier. 1974. Mathematical and practical possibilities of UNISURF. In *Computer Aided Geometric Design*, pages 127–152. Elsevier. https://doi.org/10.1016/B978-0-12-079050-0.50012-6

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–93. https://doi.org/10.1037/0278-7393.22.6.1482

Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. 2010. Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of Learning and Motivation*, volume 53, pages 301–344. Elsevier. https://doi.org/10.1016/S0079-7421(10)53008-1

J. Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. LifeQA: A real-life dataset for video question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. 2019. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894. https://doi.org/10.18653/v1/P19-1183

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press. https://doi.org/10.1037/10096-006

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213. https://doi.org/10.1037/h0026256, Pubmed: 19673146

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435*.

Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512. https://doi.org/10.1109/CVPR.2017.475

Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 271–278. https://doi.org/10.1145/2696454.2696467

Rohit Girdhar and Deva Ramanan. 2020. CATER: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20:818–829. https://doi.org/10.1016/j.tics.2016.08.005, Pubmed: 27692852

H. Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. 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.

Robert X. D. Hawkins, H. Gweon, and Noah D. Goodman. 2021. The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, 45 3:e12926. https://doi.org/10.1111/cogs.12926, Pubmed: 33686646

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1766–1776. https://doi.org/10.18653/v1/P17-1162

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *International Conference on Computer Vision*, pages 706–715. https://doi.org/10.1109/ICCV.2017.83

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73. https://doi.org/10.1007/s11263-016-0981-7

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2020. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1846–1859.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379. https://doi.org/10.18653/v1/D18-1167

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453. https://doi.org/10.18653/v1/D17-1259

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338. https://doi.org/10.1162/tacl_a_00064

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415. https://doi.org/10.18653/v1/P19-1537

Carita Paradis. 2008. Configurations, construals and change: Expressions of DEGREE. *English Language and Linguistics*, 12(2): 317–343. https://doi.org/10.1017/S1360674308002645

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136. https://doi.org/10.18653/v1/D18-1012

Massimo Poesio and Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89. https://doi.org/10.5087/dad.2010.001

Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10417–10427. https://doi.org/10.1109/CVPR42600.2020.01043

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, pages 4967–4976.

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. *arXiv preprint arXiv:1908.11279*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565. https://doi.org/10.18653/v1/P18-1238

Robert C Stalnaker. 1978. Assertion. *Syntax and Semantics*, 9:315–332.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2119–2130. https://doi.org/10.18653/v1/D19-1218

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4350–4368. https://

doi.org/10.18653/v1/2020.emnlp-main.353

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Department of Computer Science, University of Rochester.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7120–7127. https://doi.org/10.1609/aaai.v33i01.33017120

Takuma Udagawa and Akiko Aizawa. 2020. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9081–9089. https://doi.org/10.1609/aaai.v34i05.6442

Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. 2020. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765. https://doi.org/10.18653/v1/2020.findings-emnlp.67

Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2007. *Dynamic Epistemic Logic*, volume 337. Springer Science & Business Media. https://doi.org/10.1007/978-1-4020-5839-4

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. https://doi.org/10.1109/CVPR.2018.00813

Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. 2020. Towards real-time multi-object tracking. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-030-58621-8_7

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the International Conference on Machine Learning*, pages 5587–5595.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134. https://doi.org/10.1609/aaai.v33i01.33019127

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 125–131.

Luowei Zhou, Nathan Louis, and Jason J. Corso. 2018. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference*.