We thank each reviewer for taking the time to thoughtfully comment on our work and we're glad that they recognize its usefulness and novelty; *R2* says *"It will be of great help to the improvement of the generalization ability of the [...] models in understanding NL."* and *R3* writes that we *"presented multiple research problems that haven't been addressed by previous work [...]."* ***R4*** says *"gSCAN is [...] more realistic and requires stronger reasoning [...]."*. We address the concerns and suggestions of 3 reviews below, as ***R1*** finds *"no significant weaknesses"*.

The use of synthetic benchmarks (*R3* and *R4*). For systematic generalization tests we believe it to be important to limit the scope of the data (***R3***) enough such that one can pinpoint where errors come from. Furthermore, on discrepancy with real-world environments (***R4***): the issues studied in gSCAN would feature prominently in realistic NLU tasks, such as teaching autonomous agents to perform tasks by demonstration. gSCAN is distinguished by evaluating 7 types of compositionality (mostly from a single training set), whereas most benchmarks focus on a single type. gSCAN's grounding facilitates context-sensitive ("small") and modification-based ("while spinning") forms of compositionality that go beyond existing tests. That said, since the vocabulary of gSCAN is limited, for future work we will make it possible to test whether a larger vocabulary (which is trivial extension of the current setup to N object colors and M shapes) increases performance.

Response to *R2*. R2 asks an important question: NLU is hard to define for a computer, can sequence matching accuracy be used as a proxy for it? We agree with R2 that NLU is hard to define, but we believe that within gSCAN accuracy is a proxy for understanding. Generating a valid action sequence is only possible if both the command and world state are understood. Our goal with gSCAN is to be as comprehensive as possible about different kinds of compositionality, and as controlled as possible about the rest; we don't claim that it is a general-purpose NLU benchmark. R2 also wonders if gSCAN removes artifacts from SCAN. We are confident of this, as this is precisely the motivation of designing gSCAN. The methods that are SOTA on SCAN (like [1], [13], [26], [31], [33]) all exploit artifacts like interchangeability of primitives (e.g., jump, walk, run, look can be used in the same context). In gSCAN, this is not the case (due to for example the use of adverbs that have a transformative effect on the action sequence and grounding), and therefore methods that solve SCAN, fail on gSCAN. R2 wonders: don't these results just show that we need more compositional machinery? We agree with R2, and we hope gSCAN inspires this in the research community. R2 also points out some things that are unclear in the experiments section. They ask what SD in table 1 and <SOS> in Figure 3 refer to. Thanks for raising this; SD refers to standard deviation between runs, and <SOS> is the 'start-of-sequence' token that initializes the decoder. On the "lots of inconsistencies in the experiment section", we don't understand what R2 means; please clarify in the final review so we can address. R2 says the caption is wrong for Table 1, but we can't find a mistake; The claim that the models do not fail on split C and F is based on the higher accuracies (especially GECA on C). It's true the models also perform well on the random split (A), which we left unsaid but will add to the caption. Finally, we thank R2 for pointing out 2 missing links in Fig. 3, we will update them accordingly.

Respose to *R3*. R3 says a there is no significant contribution to the model, and is concerned that more models are needed to draw conclusions. The two models we used are representative of many SOTA models and have machinery that has proven important in NLU (seq. networks, conv. nets, double attention, data aug.). We believe a strength of gSCAN is that it is currently unclear how to design a model that will perform better on the tasks. We hope gSCAN sparks more research into compositional architectures that are better at reasoning from little data.

Response to *R4*. To clarify the difference between split C and E: C is about generalizing to unseen combinations of familiar colors and shapes. In the training set red squares are seen as target object, but never referred to with the color identifier (only "square" or "the big/small square"). Split E asks whether a model can learn that the same expression can refer to different objects based on the context. R2 wonders why GECA performs well on C, but not on E: the analysis in the appx. C can shed light on this. GECA adds a lot of red squares to the training set. For split E, GECA doesn't add anything of use, as it does not identify the issue that the referrent "a small circle" is never used for a circle of size 2. A limitation R4 then points out is: the results might be affected by distribution bias of the fixed held-out attributes. This is a good point that we thought about. We did an analysis on the distribution of the training set. E.g., yellow squares are referred to 16,725 times in the training set without the referring expression containing 'yellow' (more of these stats in the paper Sec. 5). That said, the training distribution indeed still differs from the test distribution, as is often the case in systematic generalization and few-shot learning: this is what we hope a model can be robust to. Finally, thanks for pointing out the failing link for [2].

Clarification about related or previous works. ***R4*** wonders how our work differs from work that focuses on contextual generalization, like [8]. It indeed seems at first that this is a related paper in terms of the dataset scope. However, [8] is specifically designed for human-in-the-loop training and sample-efficiency, with no linguistic generalization involved. R2 and R4 ask us to compare to such related work, so we are happy to add a paragraph highlighting [8] and other work. ***R3*** says the paper can be improved for readers unfamiliar with related work. This is useful information for us, and we are happy to incorporate more background from the original SCAN paper. It would be great if R3 could clarify which parts are especially unclear and what of the original SCAN paper helped for clearing this up.