We thank all the reviewers for their insightful questions, comments and commendations (novelty, clarity, performance).
Common Clarifications: **(CC1)** *Evaluation with other datasets (VQA-CP, GQA)* @R1, R2, R4: The main focus and scope of the paper is to demonstrate the efficacy of MGN in generalizing to unseen combinations of known linguistic constructs. Using datasets like CLEVR and CLOSURE allows us fine-grained control over the evaluation and analysis of our primary objective. Additionally, we tried evaluating on GQA but the size of the dataset and unconstrained language vocabulary proved to be computationally expensive. We believe our promising results on CLEVR/CLOSURE should generalize to GQA as well, and believe that future work can build upon our results.

(CC2) *Task 1 motivation, complexity* @R3, R4: Motivation behind Task 1 is efficacy, not complexity [L46-49]. We use the caption truthiness task to show the representational efficacy of MGN embedding for classification tasks, in addition to VQA. While we presented the 1-object case for clarity and space, we also obtain 99% accuracy on multi-object (2, 3 objects) caption correctness task, and the original CLEVR-CoGenT (8 objects) VQA task.

(CC3) *Prior Work* @R1, R4: *VQA-CP, VQA2.0*: These datasets and corresponding models test *generalization* across distributional shifts in train-test class labels. GVQA (from VQA-CP) builds on stacked attention networks (SAN). SAN, LXMERT, and ViLBERT use a two-stream architecture to embed each modality separately, then fuse them using attention-based interactions. Thus, they are orthogonal to MGN in problem setting and architecture. CLEVR already incorporates non-uniform train-test labels, and SAN performs reasonably well (73.2%) on CLEVR (baseline in NS-VQA). However, SAN and, by extension, GVQA architectures do not evaluate for, and generalize poorly on, unseen object attributes (CLEVR-CoGenT) and linguistic structural pattern (CLOSURE) combinations. While the NSM model also builds a scene graph from the image similar to our approach, they convert the question text into a sequence of instructions for graph traversal. In contrast, we convert text into a graph as well and perform multi-modal fine-grained matching. We show results using MAC (from the GQA authors, L291) with both CLEVR and GQA results. MGN matches MAC on CLEVR tasks, and performs better on CLOSURE [Table 2].

@R1 **Weaknesses (W1)** *Novelty*: The novelty of our contribution arises from shortcomings of cited work, and MGN addresses compositional generalization using multimodal graph fusion. Even if components of our model share similarities to prior work, the entire system is itself novel. **(W2)** *Other datasets*: Please see [*CC1, CC3*]. *Object detector supervision*: Applies only to MAC and FiLM baselines. Modifying MAC knowledge base input to feed detector supervision would be a significant model change, due to its motivation and design. The 'PG' baselines (Table 2) incorporate FiLM in their model design, and NS-VQA builds on NMN's functional program modules approach. Thus, they are fair proxies for FiLM, NMN with object detector supervision. **Prior Work**: We specifically discuss neural approaches like NS-VQA [L63], in 'Related Work' (Section 2 [L56-67]). Also see *CC3*. **Feedback**: We will add in further elaboration on GIN and Sinkhorn normalization. Fig. 6 and Table 4 in appendix A, B provide more details about the symbolic program generation. The dataset used for Task 1 is aligned with standard CLEVR-CoGenT dataset [L212] aimed at VQA – we modify templates to cater to questions with binary answer labels for classification. For CLOSURE, our model sees unannotated language as input – the color codings and tags in Fig. 4 are illustrative. We have corrected formatting errors.

@R2 **(W1)** *Results on real images*: Please see [*CC1*]. **(W2)** *Comparison fairness using parsers*: The image parser is pretrained (on 4K CLEVR images) [L129, Fig. 2] using ground-truth object attributes and 3D coordinates (identical to NS-VQA). The language parser is not trained, constructs text (s) $\rightarrow$ object graphs (G_s) using rules-based entity recognizer [L126]. **(W3)** *CLOSURE performance*: To the best of our knowledge, the results are SOTA (Table 2, and Fig 5), and no other approaches have applied GNN and neuro-symbolic methods to CLOSURE. **(W4)** *Broader Impact*: We will update with discussions on negative impacts and ethical considerations.

@R3 **(W1)** *Task 1 complexity*: Please see [*CC2*]. **(W2)** *NS-VQA comparison unfair due to RL finetuning*: We respectfully disagree. NS-VQA uses identical RL fine-tuning, and all NS-VQA results/comparisons are with RL fine-tuning. **Clarity** Minor clarifications - 5a: corrected in the camera-ready version. 5b: Yes. A_t is fixed adjacency matrix of a disjoint graph G_t [L123-125]. 5c: Yes. In [L161], H_s, H_t is supposed to be H_{G_s}, H_{G_t} – we have corrected this typo in revision. 5d: In Fig. 3, third e.g., Green -> Cyan.

@R4 **(W1)** *Evaluation on synthetic data*: Addressed in [*CC1*]. **(W2)** *Related work*: Please see [*CC3*]. **Correctness**: **(C1, C3)**: Please see [*CC2*]. Task 1 is significantly harder than CLEVR since the model has to correctly answer all the statements (10 per image) to get each instance correct. We believe the models in Table 2 might do well here also, but our goal is to simply use task 1 as a demonstration of our model's ability (98% success). **(C2)** *Related systems*: Please see [*CC3*]. **Clarity**: In Fig 1, the G_t graph for 'small blue sphere' is intentionally unlabeled alluding to its irrelevance to the question, as specifically elucidated in L92. Appendix E in supplementary materials shows exactly what graphs would look like for a particular example. **Prior Work**: [*W2@R4*]. **Broader Impact** 8,11: [*W4@R2*].