Language-Conditioned Graph Networks for Relational Reasoning (Supplementary Material)

A. Implementation details

In our implementation, we use $d_{txt} = 512$ as the dimensionality of the textual vectors (such as h_s, q, and c_t), and $d_{ctx} = 512$ as the dimensionality of the context features x_i^{ctx} of each entity i.

On the GQA dataset, we first reduce the dimensionality of the input local features x_i^{loc} (convolutional grid features, object detection features or GT objects and attributes in Table 2 of the main paper) to the same dimensionality $d_{loc} = 512$ with a single fully-connected layer (without non-linearity). During training, we train with a sigmoid cross entropy loss and use the Adam optimizer [1] with a batch size of 128 and a learning rate of 3×10^{-4}.

On the CLEVR dataset and the CLEVR-Ref+ dataset, we first apply a small two-layer convolutional network on the ResNet-101-C4 features to output a $14 \times 14 \times 512$ feature map, so that the feature dimensionality at each location on the feature map is also reduced to $d_{loc} = 512$. We use the Adam optimizer [1] with a batch size of 64 and a learning rate of 10^{-4}. On the CLEVR dataset, we train with a softmax loss for answer classification. On the CLEVR-Ref+ dataset, we train with a softmax loss to select the target location p and an L2 loss (*i.e.* mean squared error loss) for the bounding box offset u.

To facilitate reasoning about spatial relations such as "left" and "right", we also add spatial information to the local features. On the GQA dataset, when using object detection features or GT objects and attributes, we concatenate the local features with the bounding box coordinates of the corresponding objects. When using convolutional grid features (on GQA, CLEVR and CLEVR-Ref+), for each convolutional grid location (h, w), we concatenate the sinusoidal positional encoding [2] of h and w to the convolutional channel output at (h, w).

The shapes of the parameters in our LCGN model are shown in Table A.1. All our models are trained using a single Titan Xp GPU.

Parameter	Shape	Shared across t
(textual command extraction)		
W_1	$1 \times d_{txt}$	yes
$W_2^{(t)}$	$d_{txt} \times d_{txt}$	no
$W_3^{(t)}$	$d_{txt} \times d_{txt}$	yes
(language-conditioned message passing)		
W_4	$d_{ctx} \times d_{loc}$	yes
W_5	$d_{ctx} \times d_{ctx}$	yes
W_6	$d_{ctx} \times (d_{loc} + 2d_{ctx})$	yes
W_7	$d_{ctx} \times (d_{loc} + 2d_{ctx})$	yes
W_8	$d_{ctx} \times d_{txt}$	yes
W_9	$d_{ctx} \times (d_{loc} + 2d_{ctx})$	yes
W_{10}	$d_{ctx} \times d_{txt}$	yes
W_{11}	$d_{ctx} \times 2d_{ctx}$	yes
W_{12}	$d_{loc} \times (d_{loc} + d_{ctx})$	yes
(the single-hop answer classifier for VQA)		
W_{13}	$1 \times d_{loc}$	n/a
W_{14}	$d_{loc} \times d_{txt}$	n/a
W_{15}	$d_{ans} \times 512$	n/a
W_{16}	$512 \times (d_{loc} + d_{txt})$	n/a
(GroundeR for REF)		
W_{17}	$1 \times d_{loc}$	n/a
W_{18}	$d_{loc} \times d_{txt}$	n/a
W_{19}	$4 \times d_{loc}$	n/a

Table A.1. The parameter shapes in our LCGN model. All parameters are shared across different time steps t, except for $W_2^{(t)}$.

B. Quantitative analysis on edge weights

We perform quantitative analysis on the learned edge weights $\{w_{j,i}\}$, and measure how much they vary across different questions on the same image using the CLEVR dataset (where all images have exactly 10 associated questions). For each receiver node i, we associate it with a max-connected sender node $j^* = \arg\max_j \{w_{j,i}\}$. Then, we count for each receiver node i in the image how many unique j^* there are (across the 10 questions) – this number would be between 1 and 10 for each image; the higher number, the more variance in $w_{j,i}$ across questions. On average, each source node i is connected to 6.396 unique j^* across the 10 questions on the same image, showing that the learned edge weights $\{w_{j,i}\}$ are largely dependent on the input questions.

C. Additional visualization examples

Figures C.1 and C.2 show additional visualization examples for the VQA task on the GQA dataset and the CLEVR dataset, respectively. Figure C.3 shows additional examples for the REF task on the CLEVR-Ref+ dataset.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 1

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

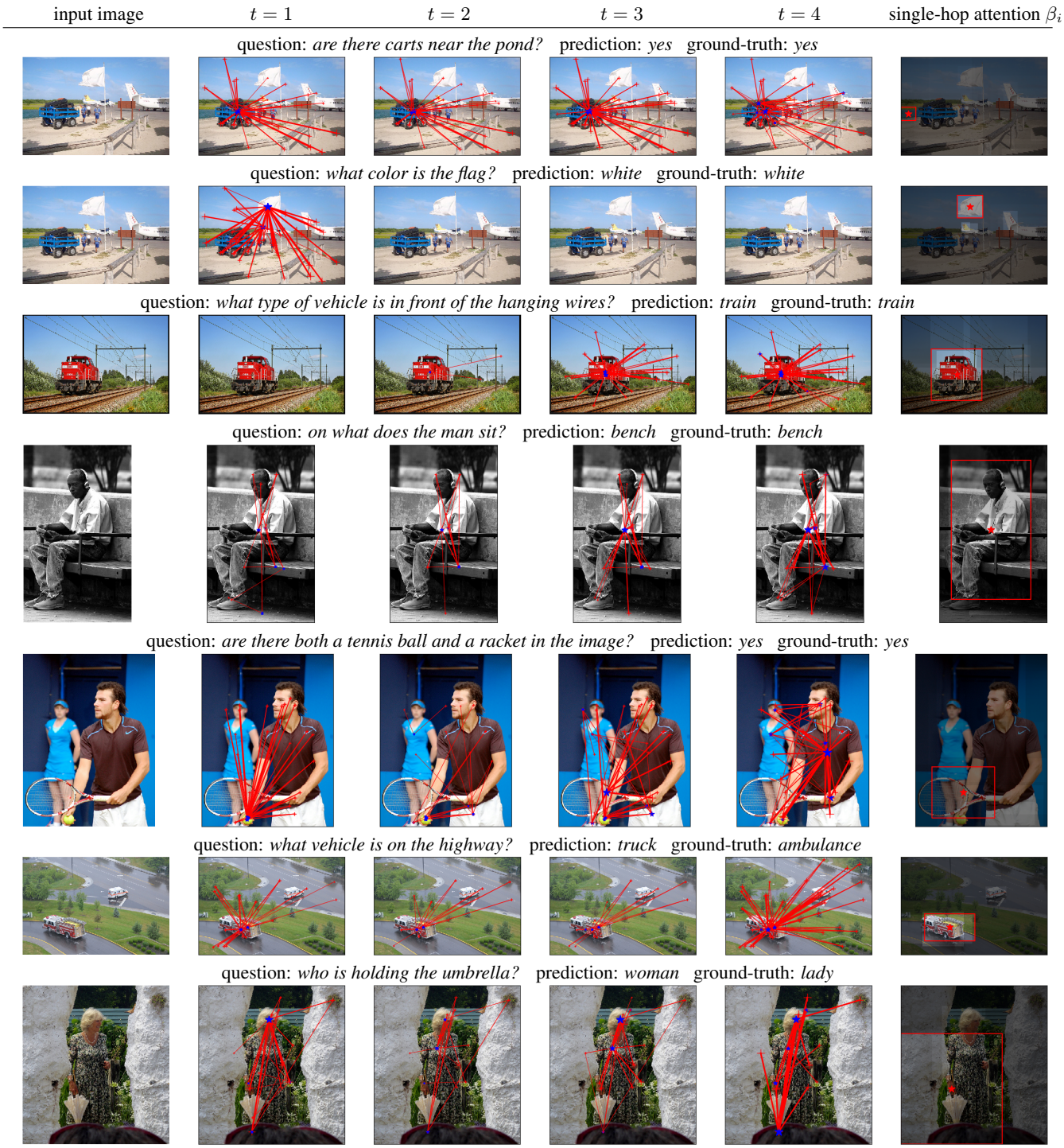

Figure C.1. Additional examples from our LCGN model on the validation split of the GQA dataset for VQA. In the middle 4 columns, each red line shows an edge $j \to i$ along the message passing paths (among the N detected objects) where the connection edge weight $w_{j,i}^{(t)}$ exceeds a threshold. The blue star on each line is the sender node j. In these example, the objects of interest receive messages from other objects through those connections with high weights (the red lines). The red star (along with the box) in the last column shows the object with the highest attention β_i in the single-hop VQA classifier in Sec. 3.2 of the main paper. The last two rows show two failure examples on the GQA dataset. Some failure cases are due to ambiguity in the answers in the GQA dataset (*e.g.* "woman" vs. "lady" in the last example).

Figure C.2. Additional examples from our LCGN model on the validation split of the CLEVR dataset for VQA. The middle 4 columns show the connection edge weights $w_{j,i}^{(t)}$ similar to Figure C.1, where the blue stars are the sender nodes. The last column shows the attention β_i in the single-hop VQA classifier in Sec. 3.2 of the main paper over the $N = 14 \times 14$ feature grid. In these examples, the relevant objects in the question usually first propagate messages within the convolutional grids of the same object (possibly to form an object representation from the CNN features), and then the object of interest tends to collect messages from other context objects. The last two rows show two failure examples on the CLEVR dataset.

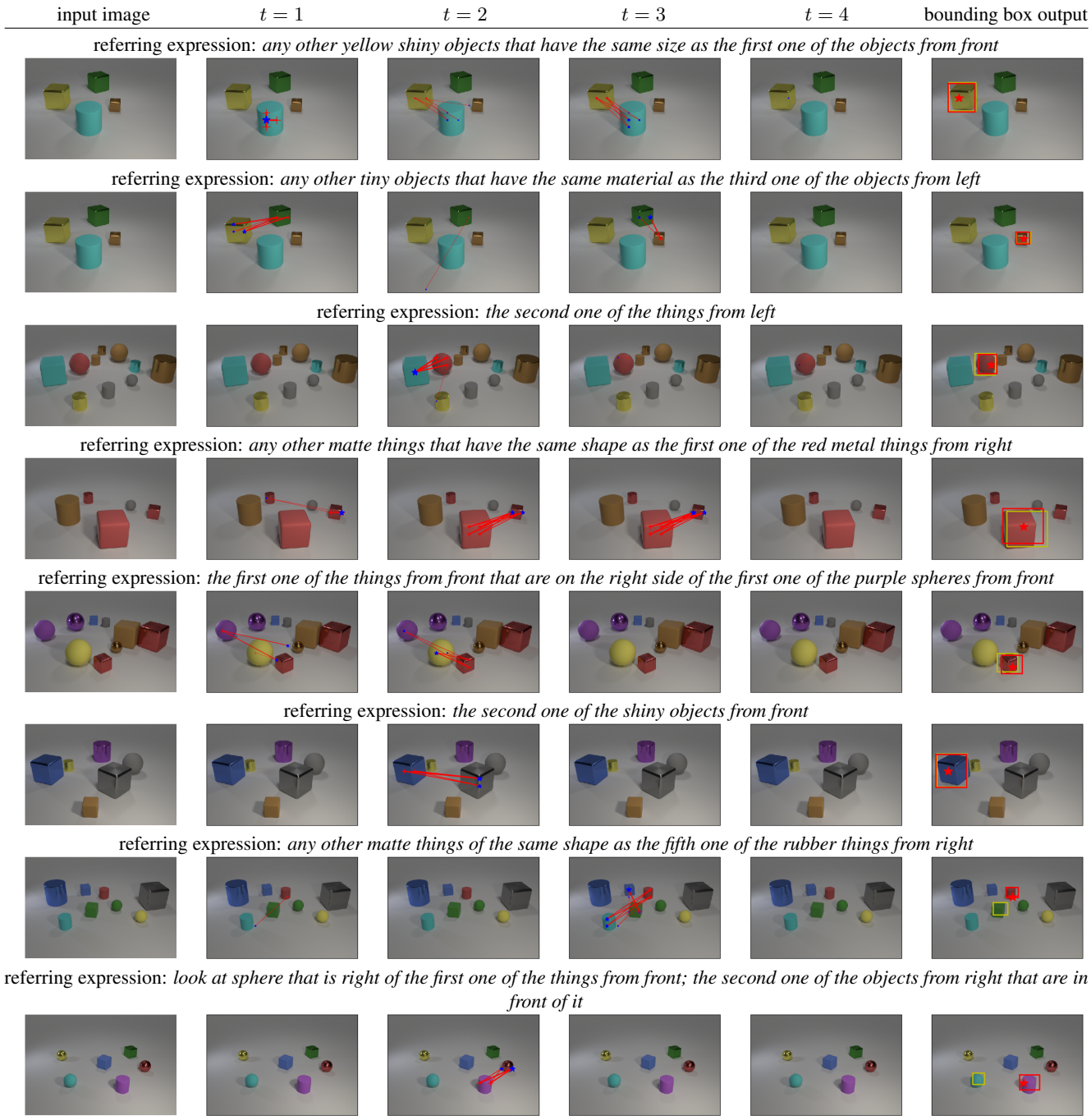

Figure C.3. Additional examples from our LCGN model on the validation split of the CLEVR-Ref+ dataset for REF. The middle 4 columns show the connection edge weights $w_{j,i}^{(t)}$ similar to Figure C.1, where the blue stars are the sender nodes. The last column shows the selected target grid location p on the $N = 14 \times 14$ spatial grid (the red star) in the GroundeR model in Sec. 3.2 of the main paper, along with the ground-truth (yellow) box and the predicted box (red box from bounding box regression u in GroundeR). In these examples, the objects of interest tend to collect messages from other context objects. The last two rows show two failure examples on the CLEVR-Ref+ dataset.