We thank the reviewers for their helpful feedback and positive view of the work. We will aim to incorporate your suggestions in future versions of the paper.
Re: dataset release. We are planning to release the new datasets in the next few weeks.

Reviewer 1

• **Add a speed comparison**. E-GQN usually reaches a given loss significantly (often 10x) faster in wall clock time (in addition to having a lower final loss). On the two datasets (rfc, jaco) where final performance is similar, E-GQN is somewhat slower due to processing $\sim 30\%$ fewer samples per second. We will add a speed analysis to the final paper.
• **Provide visualization of the attention mechanism.** If accepted, we will add one to the supplemental materials.
• **Add an ablation study**. Since the aim of our work is to design & study a new attention mechanism, we hold the remainder of the architecture constant, so the GQN numbers show the performance without the attention mechanism.

Reviewer 2

• **Discuss how approaches similar in spirit to EGQN could benefit training of conditional models in general.** Good question. Geometric structure could also be exploited in settings with moving cameras. Temporal locality may also be exploitable. We will expand on these in the conclusion.
• **Have the authors analyzed EGQN's scene representation?** We agree this would be interesting, but considered it out of scope for this paper as our main aim was improving model performance on more challenging datasets.
• **Discuss weaknesses or in general cons of the proposed model.** Thanks for pointing this out. The main weaknesses of the method are inapplicability when there is little content overlap between context and target frames (which we briefly mention in the paper), additional memory requirement relative to GQN, and slower training in terms of samples per second (though usually faster training overall as mentioned above). We will expand in the paper.

Reviewer 3

• **Would be good to show other evaluation measures.** We updated the evaluation to include standard deviation, RMSE, and ELBO in addition to existing evalution for all 7 datasets. The new table is included below as **Figure 1**.
• **Exemplary images are not given for all datasets.** Thanks for pointing this out. We agree more images would help readers better understand the models' performance. We uploaded additional images from all datasets to the supplemental website for the paper, and also included a few of them below as **Figure 2**.
• **How was the min Y value in Fig 4 computed?** We compute the minimum ELBO by assuming the KL term to be 0 and the mean of the output distribution to be the true target image. Per Section 4.2, we use a different output variance hyperparameter for our datasets vs. those from the original GQN paper, which causes the different scaling of the ELBO. We will add more detail to the description in the paper.
• **Would be nice to compare to other methods that build on the GQN, e.g. CGQN.** CGQN addresses the problem of multiple simultaneous predictions being inconsistent with one another. Since we focused on improving the quality of individual predictions, we considered it out of scope. Combining CGQN and E-GQN could be interesting future work.

Dataset	Mean Absolute Error (pixels)		Root Mean Squared Error (pixels)		ELBO (nats / dim)	
	GQN	E-GQN	GQN	E-GQN	GQN	E-GQN
rrc	7.40 ± 6.22	$\mathbf{3.59 \pm 2.10}$	14.62 ± 12.77	$\mathbf{6.80 \pm 5.23}$	0.5637 ± 0.0013	$\mathbf{0.5629 \pm 0.0008}$
rfc	12.44 ± 12.89	$\mathbf{12.05 \pm 12.79}$	$\mathbf{26.80 \pm 21.35}$	27.65 ± 20.72	$\mathbf{0.5637 \pm 0.0011}$	0.5639 ± 0.0012
jaco	4.30 ± 1.12	$\mathbf{4.00 \pm 0.90}$	8.58 ± 2.94	$\mathbf{7.43 \pm 2.32}$	0.5634 ± 0.0007	$\mathbf{0.5631 \pm 0.0005}$
sm7	3.13 ± 1.30	$\mathbf{2.14 \pm 0.53}$	9.97 ± 4.34	$\mathbf{5.63 \pm 2.21}$	0.5637 ± 0.0009	$\mathbf{0.5628 \pm 0.0004}$
oab	10.99 ± 5.13	$\mathbf{5.47 \pm 2.54}$	22.11 ± 8.00	$\mathbf{10.39 \pm 4.55}$	1.2587 ± 0.0018	$\mathbf{1.2569 \pm 0.0011}$
disco	18.86 ± 7.16	$\mathbf{12.46 \pm 9.27}$	32.72 ± 6.32	$\mathbf{22.04 \pm 11.08}$	1.2635 ± 0.0055	$\mathbf{1.2574 \pm 0.0007}$
rro	10.12 ± 5.15	$\mathbf{6.59 \pm 3.23}$	19.63 ± 9.14	$\mathbf{12.08 \pm 6.52}$	1.2573 ± 0.0011	$\mathbf{1.2566 \pm 0.0009}$

Figure 1: Performance of GQN and E-GQN. Note: ELBO scaling is due to different choices of output variance.

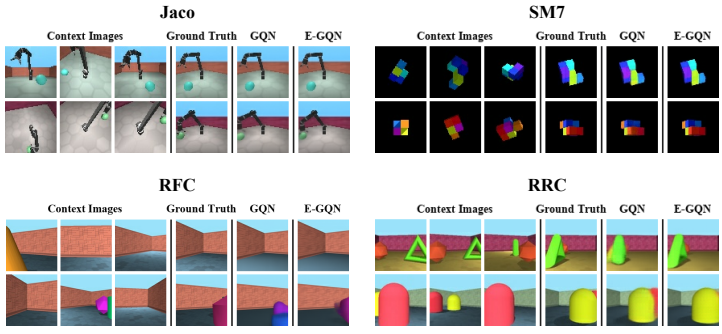

Figure 2: Randomly chosen samples from GQN and E-GQN