Thank you for the positive, constructive and in-depth reviews. We found the suggestions and comments to be very helpful. Below, we summarize the main questions and comments raised by each reviewer and provide responses.

[R1] **Drawback of SNP.** We agree. We will add a discussion on this. **Transition model in the regression task.** In Appendix D.1, we describe how length-scale l, kernel-scale σ, Δl, and $\Delta\sigma$ are chosen. To perform transition, we execute $l + \Delta l$ and $\sigma + \Delta\sigma$ and add a small Gaussian noise. **NLL in the regression task** is estimated by MC sampling, the same way as used in the Attentive Neural Processes (ANP) paper. We tested it on a held-out set of 1600 examples. **Time in regression task.** Normalized time $t' = 0.25 + 0.5 \times (t/T)$ is appended to the original query x to obtain $\tilde{x} = (x, t')$ and used an MLP($\tilde{x}, y$) to encode the query together with the target y for context encoding. **Motion model and canvas size in the 2D task.** Shapes start at random positions on a *96×96-sized canvas* with a speed of 13 pixel per time-step towards a randomly chosen direction. The bouncing behaviour is modeled the same way as in the moving MNIST dataset. **Action in 3D tasks** is uniformly randomly picked. If an action leads the object outside the arena, the action is re-picked until it doesn't. **How action and time is encoded in GQN baseline in 3D tasks.** We use a *forward*-RNN to encode context and actions for generation using $r_t = \text{RNN}(r_{t-1}, C_t, a_t)$. For inference/training, a backward-RNN similar to this is used to encode actions, context and targets of the *entire* episode. At t, action sequence is encoded as $\tilde{a}_t$ by a forward-RNN as $\tilde{a}_t = \text{RNN}(\tilde{a}_{t-1}, a_t)$. Query to GQN at time t is the concatenation $(x, \tilde{a}_t)$. Deploying an RNN encoding of the action sequence, we believe this is somewhat a stronger baseline than the vanilla GQN. **Performance Metric.** We thank for pointing out this. There was some confusion. What we actually used is sample-based NLL estimation. We found our argument connecting MSE to NLL needs a fix. The recall-MSE should be recall-NLL. The **linear PD loss annealing** was simply our initial trial that we found to work well empirically. We agree that it is worth to try your suggestion of controlling $\tilde{T}$ instead of α. **PD-α annealing.** $\alpha = 0$ in early training and set to 1 after reconstruction loss saturates. Using probability 0.2-0.5 of picking posterior transition in $\tilde{\mathcal{T}}$ worked well in practice. **Why PD and no-PD behave differently for the two 3D tasks.** For now, we hypothesize that using PD could be more effective when the task is more complex because reducing the gap between posterior and prior without PD could be easier for simple tasks. For the 3D multi-object case, because the latents need to model the dynamics of multiple objects, the information gap between $z_{<t}$ and C_t could be larger than that of single-object case, and this could make using PD more effective. In the table below, we measured two $\mathbb{KL}$s. As shown in the first row $\mathbb{KL}$, there is not much difference between using PD and not using it because $z_{<t}^{\text{posterior}}$ contains pretty abundant information. But for the second row $\mathbb{KL}$, we see that the gain by using PD becomes clearer as the task becomes more complex in the order of Multi-Object > Color-Cube > Color-Shapes. We agree that we need more investigation to understand PD better, it will be helpful to have a **toy task to analyze posterior collapse**. We hope to include this in the camera-ready version.

Task	Multi-Object		Color-Cube		Color-Shapes	
Loss Type	No PD	PD	No PD	PD	No PD	PD
$\mathbb{KL}(q(z_t \mid z_{<t}^{\text{posterior}}, C_t, D_t) \parallel p(z_t \mid z_{<t}^{\text{posterior}}, C_t))$	4.78	3.18	1.86	0.83	0.73	0.56
$\mathbb{KL}(q(z_t \mid z_{<t}^{\text{prior}}, C_t, D_t) \parallel p(z_t \mid z_{<t}^{\text{prior}}, C_t))$	57.45	3.49	3.51	1.06	1.05	0.67

[R2] We thank for the positive and insightful review. We treasure all the points that would make our paper clearer and more precise. We agree on all of them. To judiciously use space, we address the remaining comments below. **Do we explore empty C_t for $t > T$?** Yes, for 2D and 3D tasks, we show context only up to $t = 5$ and we demonstrate the temporal generalization up to $t = 20$ or 30. **Posterior notation.** We thank for pointing out this. We followed the argument and we will make it clearer in the camera-ready. **$P_\theta \equiv Q_\phi$?** We will clarify that "in practice $\phi = \theta$". **Sum of C_t and D_t.** We meant the sum of the respective vector encodings but we agree it is more apt to say "$C_t \cup D_t$". **PD's α sensitivity.** We agree this needs more study. We responded with some details in line 20 above. **Details about the 1D task.** In sub-tasks (a) and (b) we train the model under those settings before validating. For regression tasks, dynamics are actionless. Our training time-horizon was $T = 20$ for tasks (a) and (b) and $T = 50$ for task (c). **Choosing C_t, D_t for TGQN.** At each time t, we take 20 random camera angles in $[0, 2\pi)$ and we use a part of it as context and leave the remaining as target. In each of the first 5 time-steps, we randomly decide the context set sizes uniformly in ranges $[1, 3]$ and $[1, 5]$ for 2D and 3D tasks, respectively. For 2D task, we pick the patch location (viewpoint) uniformly on the canvas. **Uncertainty demonstration.** Due to limited space, we initially could not fit it in the main body but we would find a way to emphasize it more. We will also properly emphasize the fact that "**SNP's main motivation is *not* to model a stochastic process that is a sequence.**"

[R3] Thanks for the positive review and the reference to *defensive importance sampling*. It helps build a better motivation for the PD loss. **Can we query any viewpoint (in query space) in the spirit of GQN?** Yes, the regularly spaced viewpoints in diagrams are only for illustration to ease the reader in following object motion. **In face-uncertainty, does consistency hold per scene?** Yes, we will add illustration for this in uncertainty demonstration. However, more study is needed on consistency across time-steps since each t has its own latent. **Code.** We will make it available.