We thank the reviewers for their thoughtful comments and suggestions. We performed several new experiments and analyses to address the comments and will make the suggested changes to the main text. We also thank all reviewers for taking the time to point out minor errors. Below, we address the reviewers' comments individually.

R1, R6: Additional analyses/ablations for $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{sep}}$. We agree with Reviewer 1 that much of the novelty of our work lies in the losses and training approach. We performed new analyses to show that $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{sep}}$ are crucial to the performance and stability of the model, both in terms of video metrics (especially FVD, Fig. A) and coordinate tracking accuracy (Fig. B), on which downstream tasks depend. We will add these analyses to the main text.

R1: Temporal consistency and "jumping" keypoints. We initially experimented with using predictions from the dynamics model as "prior" for the keypoint detector, but achieved better performance without enforcing temporal consistency explicitly. Keypoints can indeed "jump" between frames, but we show in a new analysis (Fig. D) that the VRNN partially smooths over such jumps: We displaced the location of one keypoint by $0.5 \times$ image width in the direction of the image center for one frame (Basketball dataset). The keypoint location inferred by the VRNN jumps by less than $0.5 \times$ image width in the perturbed frame and quickly recovers. Jumping thus seems to be a minor issue.

R1: Did you observe training issues when combining a large K with $\mathcal{L}_{\text{sep}}$? Note that the optimal σ_{sep} (spatial Gaussian radius of $\mathcal{L}_{\text{sep}}$) is very small ($\sigma_{\text{sep}} = 2 \times 10^{-3} \times$ image width for Human3.6M). At this σ_{sep}, the loss does not interfere with initial training even for large K, but still prevents keypoints from collapsing onto the same image feature.

R1: What is the size of the feature vector in CNN-VRNN? We made sure to match the size of the feature vectors of the models, such that the CNN-VRNN had $K \times 3$ dimension at the narrowest point. Therefore, in principle, the CNN-VRNN had the capacity to exactly recapitulate the Struct-VRNN structure.

R1: Usefulness of KP structure for RL. Our claim has since been confirmed by Kulkarni et al. (arXiv 1906.11883v1).

R5: How is spatial structure imposed and why is it not sensitive to initialization? See Jakab et al. [12] for how the keypoint detector imposes spatial structure. A naïve application of [12] to video indeed suffers from sensitivity to initialization (see Figs. A and B, "no $\mathcal{L}_{\text{sparse}}/\mathcal{L}_{\text{sep}}$ loss"). By adding $\mathcal{L}_{\text{sparse}}$ and $\mathcal{L}_{\text{sep}}$, we achieve high robustness.

R5, R6: Comparison to adversarial methods. We note that we do compare to an adversarial method ("EPVA-GAN", Fig. 3, bottom right). A GAN loss could also be added to our model as a complementary objective; this is orthogonal to our contributions. We agree that comparison to SAVP would be interesting, but we could not obtain results in time for the rebuttal. We will include them in the final paper.

R5: Why train keypoint detector and dynamics model separately? We initially tried to train the model jointly ($\varphi^{\text{det}} \rightarrow$ VRNN $\rightarrow \varphi^{\text{rec}}$), but found that the model learned an unstructured latent code, rather than spatially meaningful keypoints. Presumably it was easier for φ^{rec} to reconstruct the image from an unstructured code, than for the VRNN to learn the keypoint structure. Isolating the keypoint detector from the dynamics model solves this problem.

R6: Why not apply B.o.M. sampling and $\mathcal{L}_{\text{sparse}}$ to CNN-VRNN? We did apply both to CNN-VRNN, but this yields no gains because sample evaluation and sparsity are less meaningful in an unstructured space than in keypoint space.

R6, R7: Is sample diversity an advantage? Are all samples good? We agree with Reviewers 6 and 7 that we need to expand the discussion of sample diversity. Fig. E below shows that even the samples with the lowest VGG cosine similarity to ground truth are of high visual quality. For videos, see Sections 2 and 3 on the anonymous website (link in original submission). We will add more examples and videos to the final paper. We emphasize that frame-wise similarity to GT (e.g. VGG sim, PSNR and SSIM) does not meaningfully measure video prediction quality. For real data, at test time, there is no single "ground truth". Instead, there is an astronomical number of plausible futures that are all consistent with the conditioning frames. We believe that most previous models dramatically underestimate this diversity; our model comes closer to it. This is backed up by FVD, which is designed to measure sample diversity.

R7: More fine-grained evaluation of object tracking. We performed a new analysis of per-object tracking performance on Basketball (Fig. C). We identified two different failure modes: The basketball (yellow traces) has relatively large tracking errors across all 10 model initializations, presumably because the dynamics of the ball are hard to learn. On the other hand, Player 3 (pink) is tracked well in some and poorly in other model initializations, presumably because the keypoint detector fails to recognize the light-colored object. We will describe these failure modes in the main text.

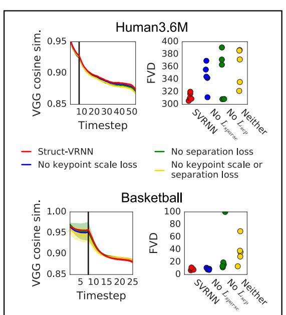

Figure A: Video metrics for loss ablations. Each dot is one model initialization.

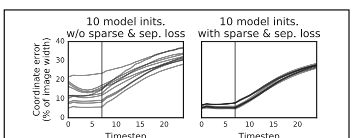

Figure B: Coordinate regression error for 10 model initializations (Basketball dataset).

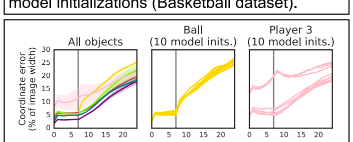

Figure C: Different objects (e.g. Ball and Player 3 in Basketball) have different error modes.

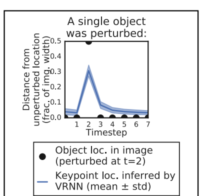

Figure D: The dynamics model partially smooths over "jumping" keypoints. We displaced the location of one object by 0.5 × img. width. The location inferred by the VRNN jumps by less than 0.5 × img. width in the perturbed frame and quickly recovers (Basketball dset).

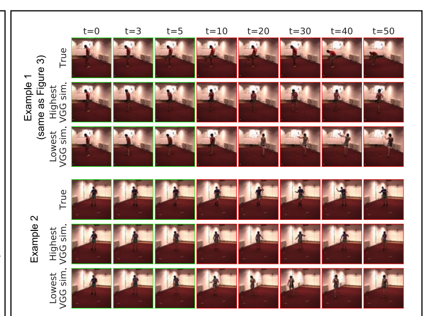

Figure E: Even the samples with the lowest VGG cosine similarity to ground truth are visually realistic.