We thank the reviewers for their insightful and constructive comments.

[R3] “... larger networks provide better generalization": We not only show that larger networks provide better generalization capacity per se, but also provide rigorous studies about (1) the minimal inductive bias necessary to achieve high quality video prediction, (2) quantifying the gains resulting from each architectural component, and (3) quantifying the gains resulting from gradual increase in capacity. We show the progression and performance increase going from an encoder/decoder CNN (CNN); to adding a recurrent component (LSTM); to adding a recurrent stochastic component in the architecture (SVG). In addition, our study progressively increases the difficulty of the datasets to highlight how each of the models being studied perform at each level of difficulty (i.e., action conditioned prediction, action-free with static background, action-free with moving background). As highlighted by R1 and R4, although the idea that increasing capacity can be beneficial for model performance may not be a surprise, our paper is the first to successfully and comprehensively demonstrate and quantify this for video prediction over five different metrics. A lot of effort goes into discovering domain-specific architectures (e.g. using optical flow, segmentation masks, and other forms of inductive bias) – and we hope our work encourages the field to rethink about these aspects of scalability.

[R3] “Datasets are relatively causality-explicit, and thus, not much uncertainty in prediction": While we agree that action conditioning limits the uncertainty for the BAIR experiments, there is still partial observability in the object interactions. The model has to hallucinate the unseen parts of the objects and also any stochasticity in the interaction which cannot be fully determined by observing the pixels (e.g., table friction). On the other hand, Human 3.6M and KITTI contain larger amounts of stochasticity. First, the actions in the Human 3.6M dataset are highly stochastic, that is, the human randomly decides to do different actions regardless of the label (e.g., "sitting" action randomly goes from sitting to getting up to walking). This makes the prediction not fully determined by the observations so the model has to choose one of the possible futures and predict it. Second, the driving data from KITTI is also highly stochastic due to strong partial observability. Given input frames from the driving scene, models need to be able to hallucinate the road and vehicles that are hidden by the horizon line caused. Having said that, we also agree that it is interesting to evaluate on datasets with higher uncertainties (though not as well established in the video prediction literature) and will try to include such results in the final version.

[R3] Prediction accuracy depending on the context length: We ran experiments with history length of 5 and 10 frames. We evaluated with the same data as in the submission (30 frames total), thus, we evaluate by predicting 20 frames into the future so we can align the future frames for comparison. Due to space limitations, we cannot provide full sequence plots for the frame-wise evaluation, and so, we provide the average over all time steps. Also, due to time constraints, we trained the baseline (smallest) model with M=1 and K=1. We will add results for the biggest models in the final version. For similar reasons, we couldn’t run experiments on the robot dataset. However, since there is action conditioning on the robot dataset, context frames may be less influential. Overall, we observe that most of the metrics improve with more context frames—i.e., 7 out of 8 evaluation settings except for the case of FVD on Human3.6M (each row in the table corresponds to a combination of evaluation metric and dataset). We further expect that larger-sized models will perform better with longer context size and will report more comprehensive results in the final version.

		CNN models		LSTM models		SVG’ models	
Dataset	Metric	history=5	history=10	history=5	history=10	history=5	history=10
Human 3.6M	PSNR **(higher/better)**	22.351	22.522	22.927	23.108	22.841	**23.399**
	SSIM **(higher/better)**	0.873	0.877	0.886	**0.894**	0.887	0.891
	Cos. Sim. **(higher/better)**	0.882	0.881	0.898	**0.903**	0.899	0.902
	FVD **(lower/better)**	848.714	890.270	616.474	572.628	**565.952**	693.561
KITTI driving	PSNR **(higher/better)**	11.325	11.585	13.988	**14.522**	14.262	14.516
	SSIM **(higher/better)**	0.261	0.263	0.37	0.405	0.389	**0.408**
	Cos. Sim. **(higher/better)**	0.465	0.475	0.597	0.617	0.600	**0.621**
	FVD **(lower/better)**	2921.798	2871.245	2063.228	2127.124	2151.003	**2021.726**

[R1, R4] Comparison with SOTA Architectures: SAVP is a competitive video prediction model that combines many of the previously proposed methods (optical flow, adversarial losses, masks) but it also requires significant hyperparameter tuning. Although SAVP achieved strong results on (relative easy) BAIR Robot Pushing and KTH datasets, it has not been demonstrated on more complex datasets (e.g., BAIR Towel-Pick and Human3.6M are much more challenging than BAIR Robot Pushing and KTH, respectively). In our initial experiments based on the authors’ implementation of SAVP, our large-scale models outperformed SAVP. We will further verify this with additional hyperparameter tuning for SAVP and report the results, but as of now, there is no evidence that SAVP (without scaling up) can be competitive to our best performing large-scale models on these challenging datasets. Scaling up SAVP could be interesting future work, but it may be nontrivial due to the complexity of the architecture and hyperparameter tuning.

[R1] Model capacity comparisons in main text: Thanks, we will fit the primary capacity results in the main paper.