(R1, R2, R3) We will correct all typos, grammatical errors, and misleading notations (e.g. Eq. 4) in the revision. We will also clarify ambiguous terms and cross-dataset evaluations.

(R1) Side-to-side video comparison. We will add side-to-side video comparison in the revision.

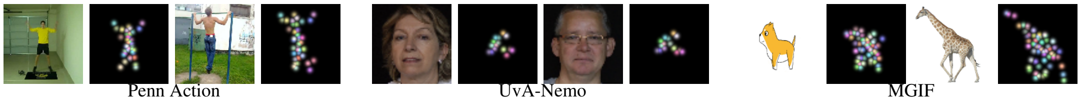
(R1) User study Details. We sampled a trajectory once for each video by randomly sampling the latent variable. During user study, users were shown four videos side by side, and asked to rank them according to the criteria we described in the paper. The order of four videos was randomly chosen for each vote.

(R1, R2) More quantitative evaluation. The table below shows more quantitative results. As our method is stochastic, we measure the quality of generated videos using Fréchet Video Distance (FVD) instead of metrics such as PSNR or SSIM. For the Penn action dataset, we additionally show the accuracy of human action recognition to evaluate the performance of motion generation using the two-stream CNN. We will add this table with discussion in the revision.

Method	Dataset	Real test data	Ours	[17]	[12]	[19]	Method	Dataset	Ours	[12]	[19]
Action recognition (%)	Penn Action	83.33	68.89	47.14	40.00	15.55	FVD	UvA-Nemo	162.4	265.2	666.9
FVD (lower is better)	Penn Action	-	1509.0	2187.5	3324.9	4083.3	FVD	MGIF	409.1	1079.6	683.1

(R1) Effect of adversarial loss in the motion generator. To show the effectiveness, we measured the action recognition accuracy of our network, and our network without adversarial loss. The results are 68.89 and 66.67 respectively, which indicates that the adversarial training is effective in generating more realistic motion.

(R1) Visualization of learned keypoints. Figure below shows examples of learned keypoints on each dataset where each colored dot indicates a specific keypoint. In UvA-Nemo dataset, keypoints are usually distributed around features of face since most of the videos have movements in local regions around the mouth and eyes. As can be seen in the figure (leftmost), our keypoint detector is not aware of body orientation. In the example, the upper-left green point represents "right hand" on the first sample, while it represents "left hand" on the second sample. We will add this discussion in the revision.



(R2) Multiple objects & Learning foreground/background. This is a good point. Handling multiple moving objects is still challenging for our method and all other prior methods. We showed failure examples in this case in Fig. 6 of the supplementary material. Our method also fails in generating scenes with drastic change of background since it focuses on learning motion of a foreground object. We agree that considering multiple objects with dynamically changing background is desirable, and we believe that our work can inspire follow-up research for learning complex scenes. We will discuss it with more failure cases in the revision.

(R2) End-to-end training. Our training consists of two stages, and end-to-end training is unstable. In the future, we plan to improve our method to be end-to-end trainable.

(R2) Identical artifacts. The uniform artifacts are mainly caused by imperfect mask prediction. In this case, foreground contents in the input image are identically copied to all frames.

(R2) Sync with the ground-truth. As our method is stochastic, sampled videos may not sync with the ground-truth.

(R2) Adversarial loss. We tried advanced losses, but we found a vanilla GAN was the best in terms of visual quality.

(R2) Evaluation of the detected keypoints. Please note that direct evaluation is difficult as our learned keypoints do not have ground-truth. We show the improvement of keypoint detection over the original work [20] in **(R3) Ablation study on the image translator.** The SIFT-based dynamic features may be applicable, but we believe that our network predicts more informative keypoints by learning proper locations to synthesize body parts from data.

(R2) Effect of class condition. We compare FVD scores of our method and our method without class condition, which are 1493.2 ± 22.9, and 1520.2 ± 32.7, respectively. Even though removing class condition slightly affects the performance, the gap is negligible compared to the baseline results.

(R3) Ablation study on the image translator. In Fig. 1 of the supplementary material, we compared results of (i) the original work [20], (ii) our network (+reference keypoints, -mask), and (iii) our network (+reference keypoints, +mask) on the Penn action dataset. Comparing (i) and (ii), our network works better on learning keypoints. We conjecture that explicitly learning the analogical relationship of the keypoints and the images enables our network to learn the keypoints detection easier without disentangling moving background. Comparing (ii) and (iii), our full network produces more plausible images as the masking enforces our network to focus on synthesizing foreground object, which is beneficial for the network to fool the image discriminator. We will add detailed discussion in the revision.

(R3) More generation results. We kindly remind you of the generation results of different scenarios shown in Fig. 4 and 5 of the supplementary material. We will add more results with discussion in the revision.

(R3) The UvA-NEMO dataset. Our purpose of using this dataset is to evaluate performance on learning class-agnostic motion. Nevertheless, we agree with your point and we will add results on the MUG dataset in the revision.