Supplementary Materials

1. Reward Progression along Training Steps

Figure 1. Training Results. Reward changes over training steps. (model 1&3) VLM+Feature Reward, (model 2&4) VLM+KeyPoints Reward

Figure 1 shows that the reward score increases while training, which indicates the models are well following the intention of our reward functions and the framework. We stopped the training when the validation images show the signs of overfitting, such as blurry or unrecognizable background.

2. Qualitative Progression along Training Steps

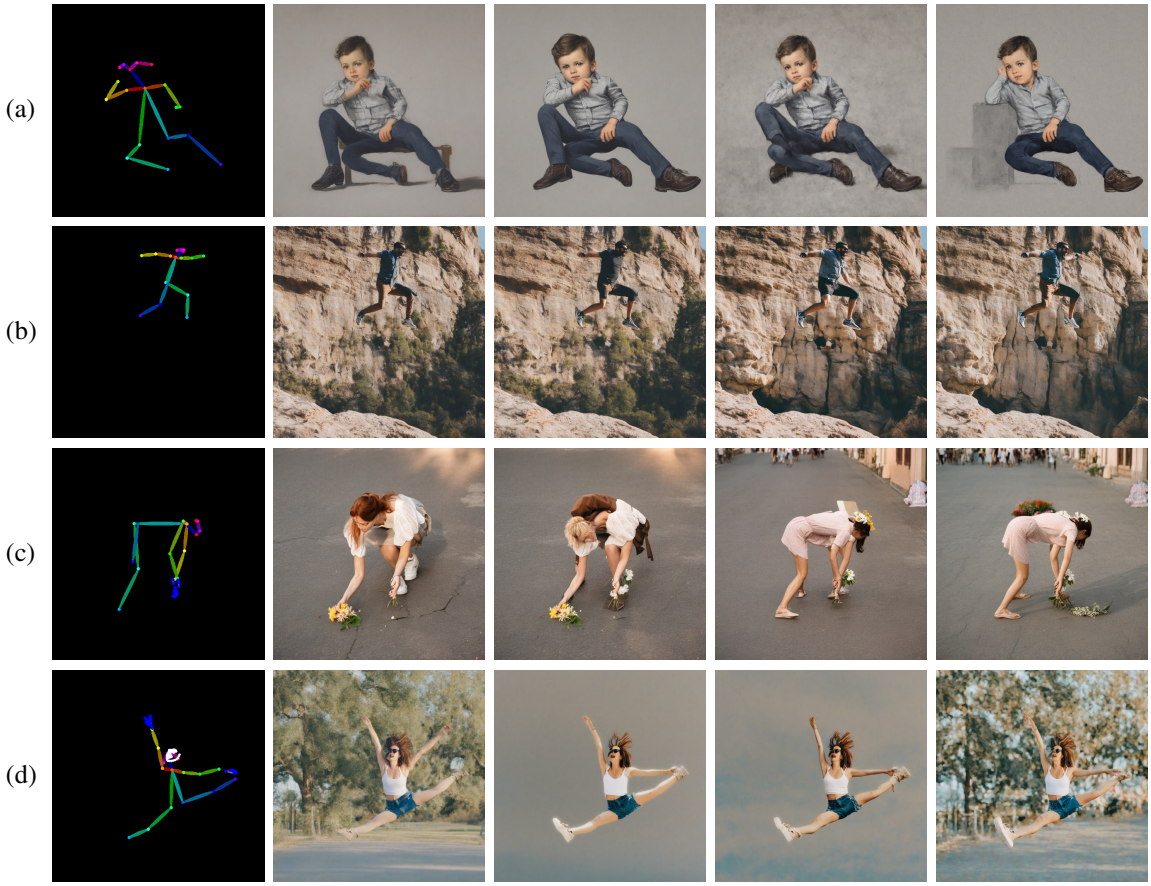

Figure 2. Progression of the models along steps. (first column) Skeleton image conditions, (rest columns) The pose is getting well followed to the conditioning image while training. (row (a) $\sim$ (d)) The generated images by the proposed model 1 $\sim$ 4.

Figure 2 shows the progression of the models with regard to the pose accuracy while training. The models end up generating the images following the conditioning images correctly without omissions or additions of the body parts. For example, In Fig. 2-row (a), which is the images of the model 1, extra leg is generated in the earlier steps but it disappears and generates the legs properly following the conditioning images well. In row (b), the model 2 omits left arm at first but the arm appeared in the third image. It generates the left arm correctly at last. In row (c) and (d), the images generated by model 3 and 4, the direction of the body or the arm is following the conditioning images while it was wrong in the earlier steps.

3. Additional Quantitative Evaluation Results: Comparison with Full Fine-tuning

Condition Type	Model	CLIP Score ↑	LPIPS ↓	OKS ↑
OpenPose w/o hands and face	ControlNet	29.2268	0.7236	0.5028
	ControlNet(Fine-tuning)	<u>29.5648</u>	0.7360	0.5135
	Model 1 (ours)	29.0550	<u>0.7230</u>	0.5420
	Model 2 (ours)	29.0678	0.7269	<u>0.6006</u>
OpenPose w/ hands and face	ControlNet	<u>29.1303</u>	<u>0.7137</u>	0.6046
	ControlNet(Fine-tuning)	28.7491	0.7168	0.6051
	Model 3 (ours)	29.0470	0.7257	<u>0.6837</u>
	Model 4 (ours)	29.0534	0.7167	0.6551

Table 1. Additional Quantitative Evaluation Result. We conducted additional quantitative evaluations by training the ControlNet+OpenPose model on our training dataset. The results of the fine-tuned model are presented in the second row of each condition type.

Table 1 shows the results that are carried out to demonstrate that our training framework outperforms full fine-tuning ControlNet models in terms of improving pose accuracy, which was the main objective of our paper. Full fine-tuning ControlNet model affected a slight increase in OKS score, while our models in both condition types exhibit noticeable improvements. This validates the effectiveness of our proposed training framework.

4. Additional Samples

We provide further qualitative comparisons between the baseline models and our proposed models using various pose conditioning images. In these results, we show how our models handle different poses and demonstrate improved accuracy in pose alignment compared to the baseline models.

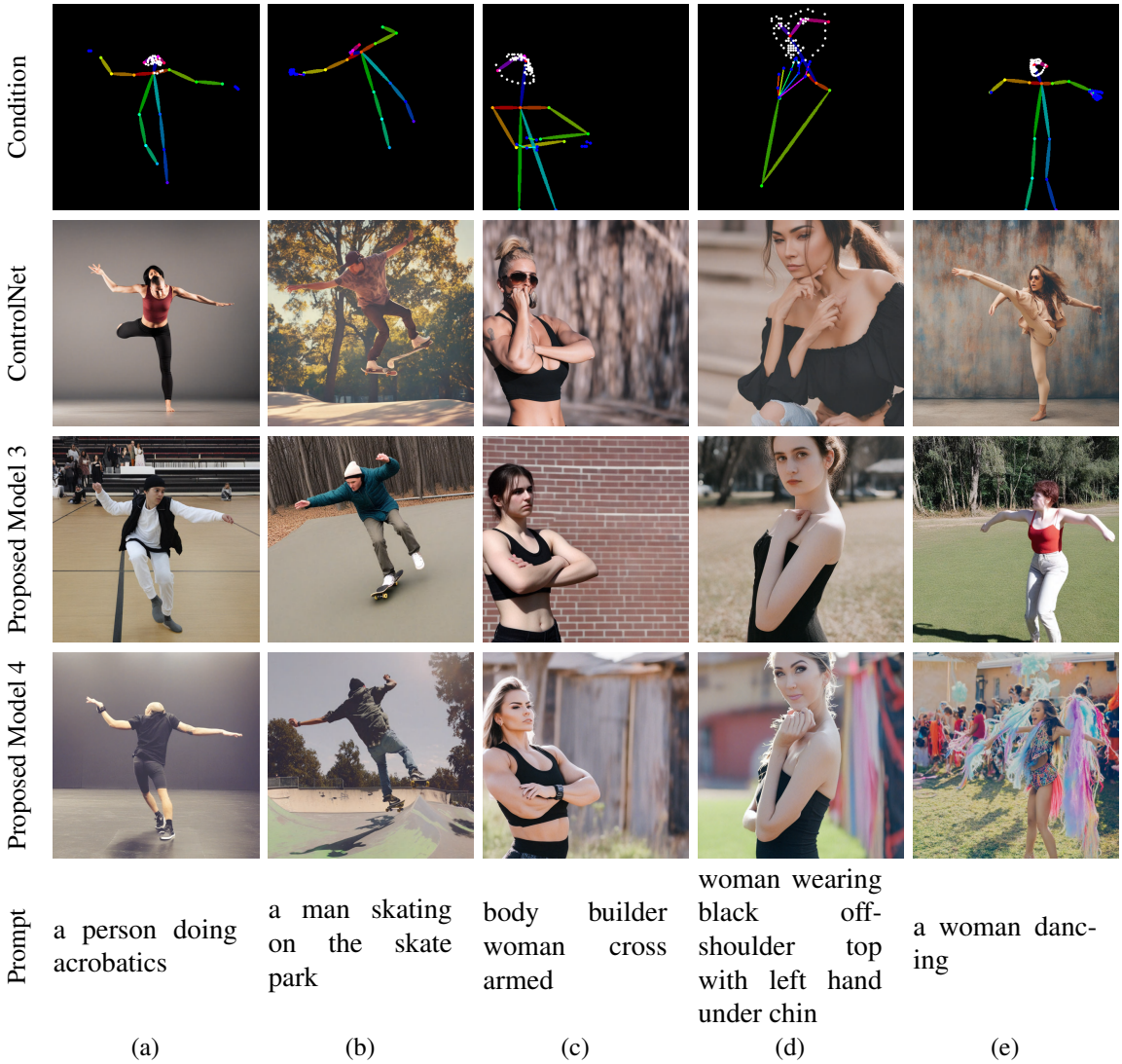

Figure 3. Additional Qualitative Results(w/ hands and face). The results are generated with condition images with hands and face keypoints. (first row) Skeleton image conditions, (second row) the generated images by the baseline model, ControlNet, (third and fourth row) the generated images by our model, the proposed model 3 and 4, (last row) the input prompts.

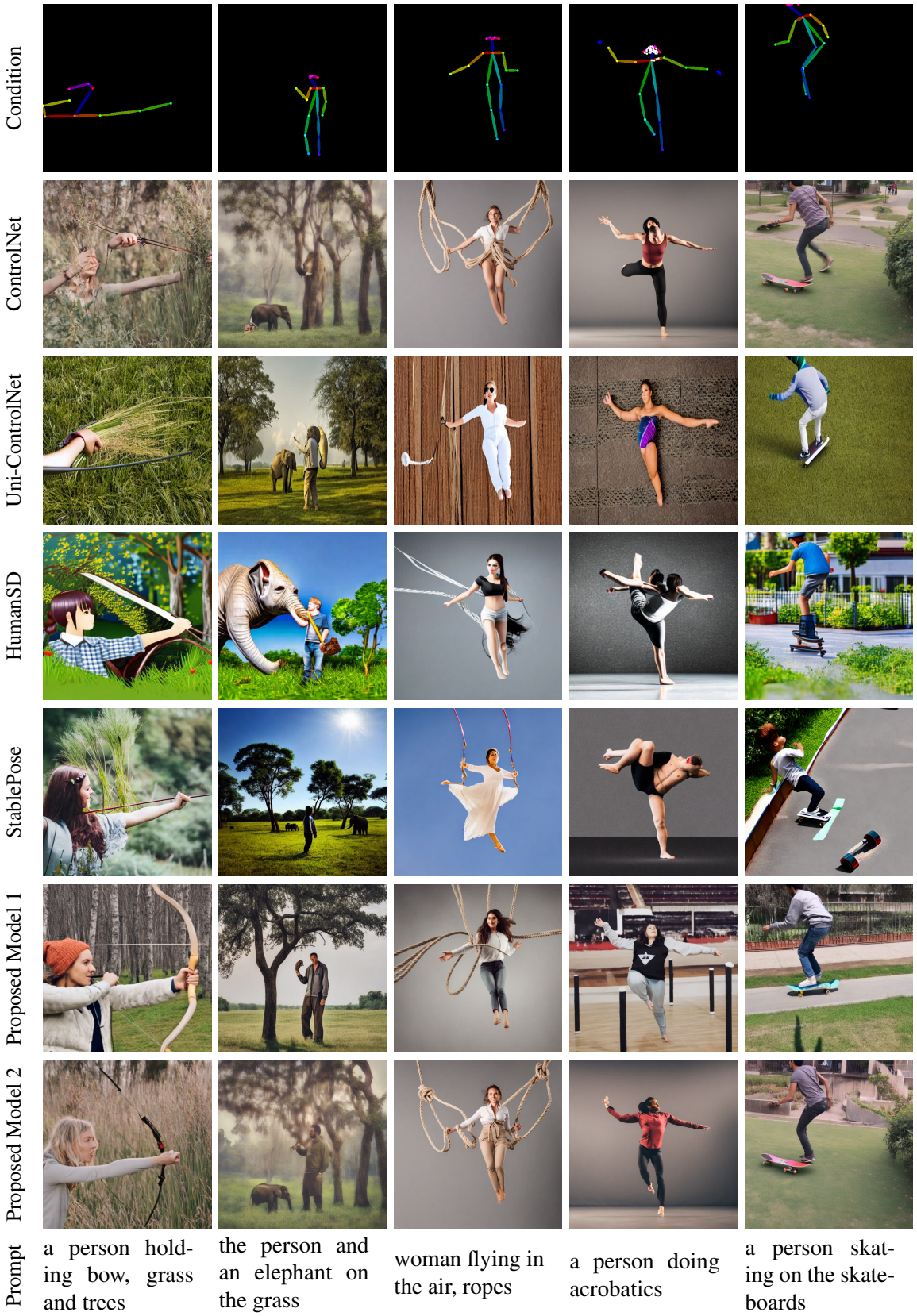

Figure 4. Additional Qualitative Results(w/o hands and face). The results are generated based on conditioning images without hands and face keypoints. (first row) input conditioning images, (second to fifth row) the generated images by the baseline models, ControlNet, Uni-ControlNet, HumanSD and StablePose, (sixth and seventh row) the results from our model 1 and 2, (last row) the input prompts.As inputs, OpenPose images are used for ControlNet, Uni-ControlNet, and ours, and MMPose images are used for HumansSD and StablePose.