Supplementary Material
KPE: Keypoint Pose Encoding for Transformer-based Image Generation

We present additional result of our KPE pose guided text-to-image model. All images were generated by our model, using only keypoints (skeleton image is only for illustration purpose) and the text as shown in the figure's caption. This demonstrates our method can **consistently** generate images that are faithful to both the pose and text prompt. We use inference method as described in the main paper, that is to select the top 0.1% or 8 out of 8192 possible tokens and sampling randomly from them.

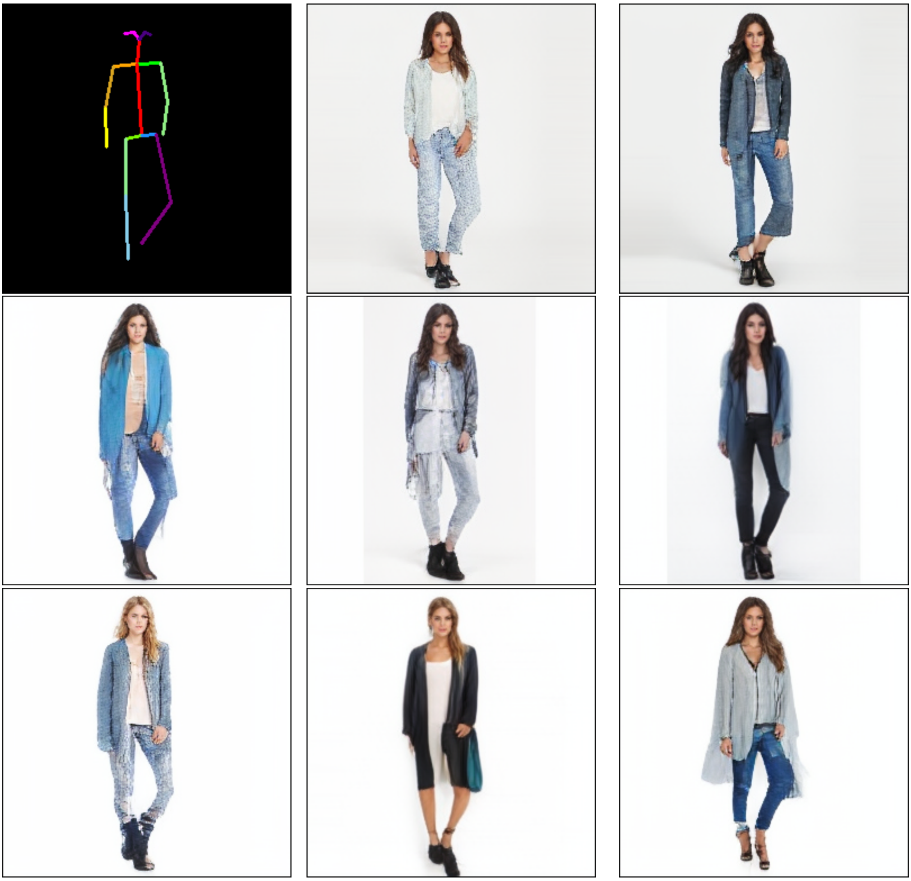

Fig. 7: 'the lady wore a blue long-sleeved cardigan.'

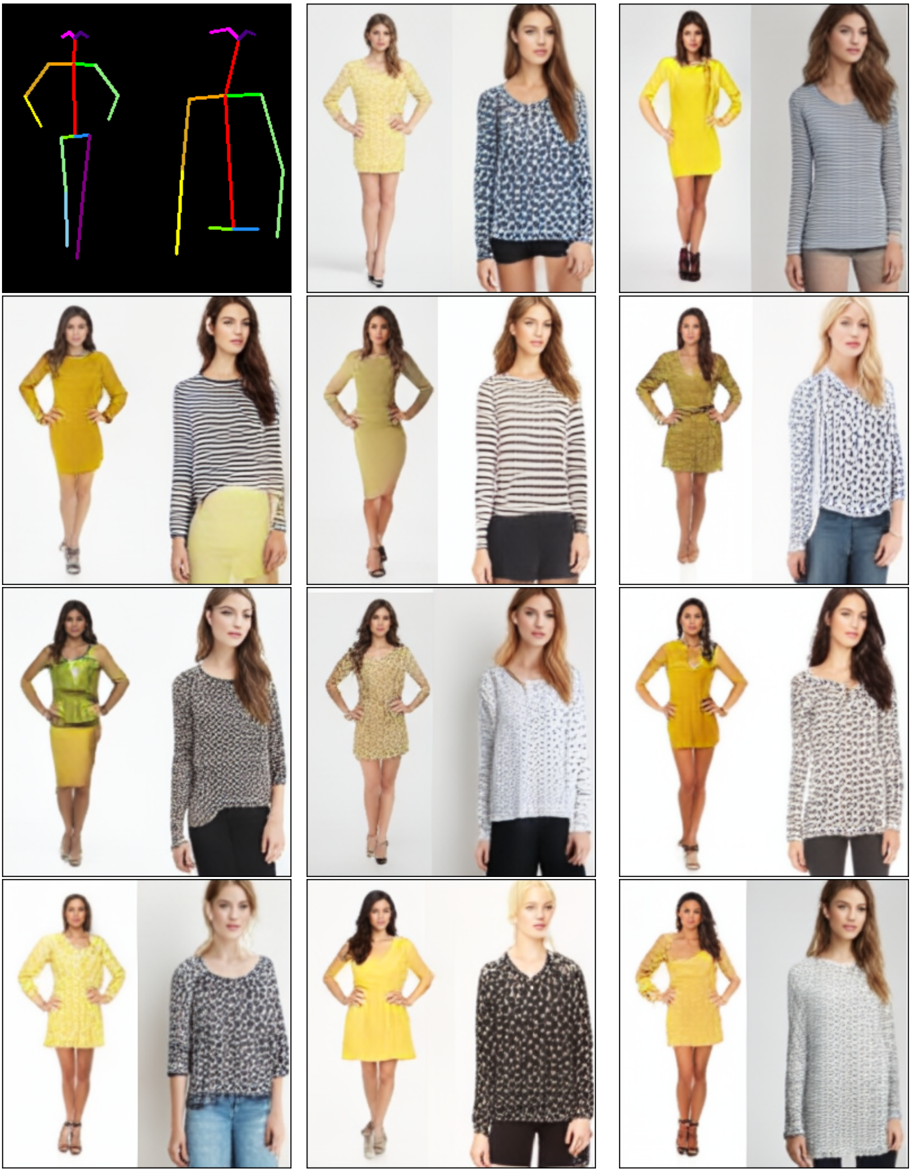

Fig. 8: ‘the lady is wearing a yellow long-sleeved dress.the lady is wearing a multi-color long-sleeved tee.’

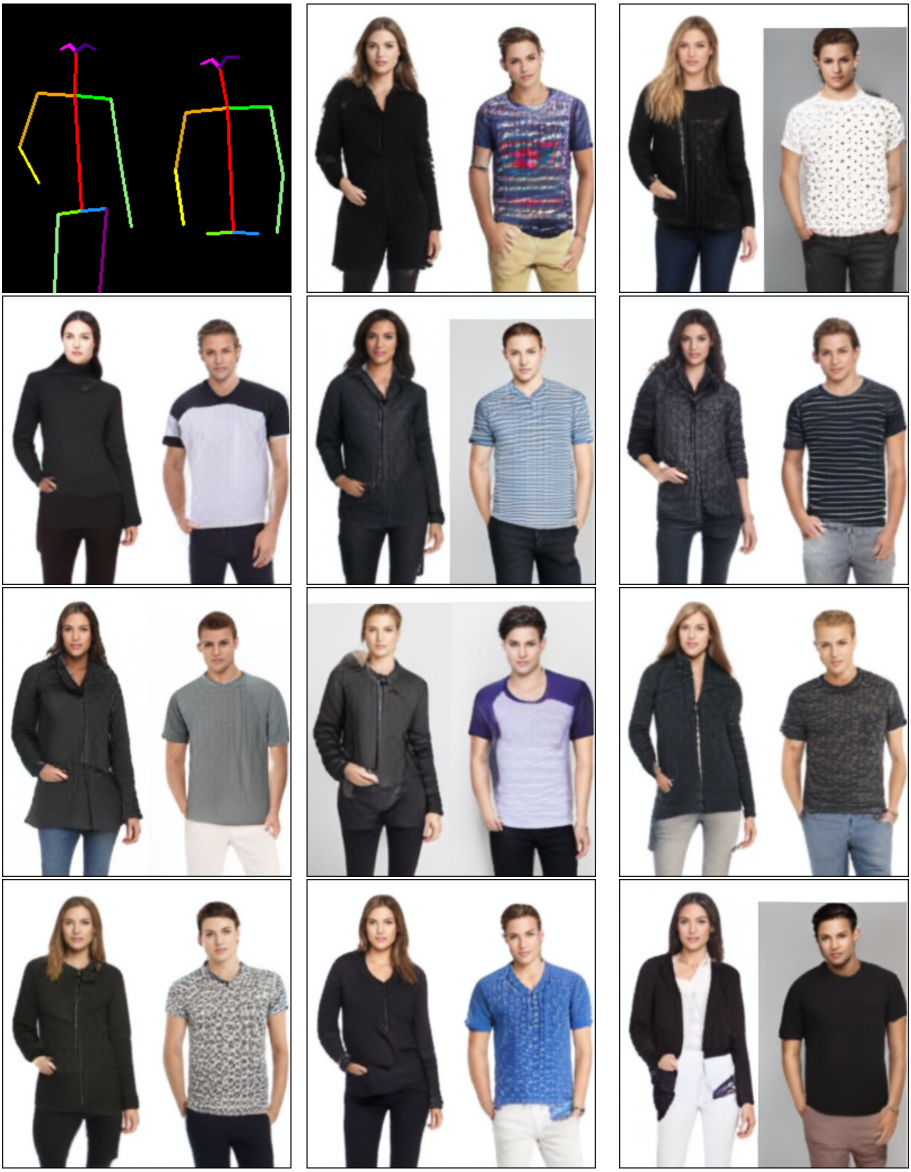

Fig. 9: ‘the lady is wearing a black long-sleeved parka.the man is wearing a multi-color short-sleeved tee.’

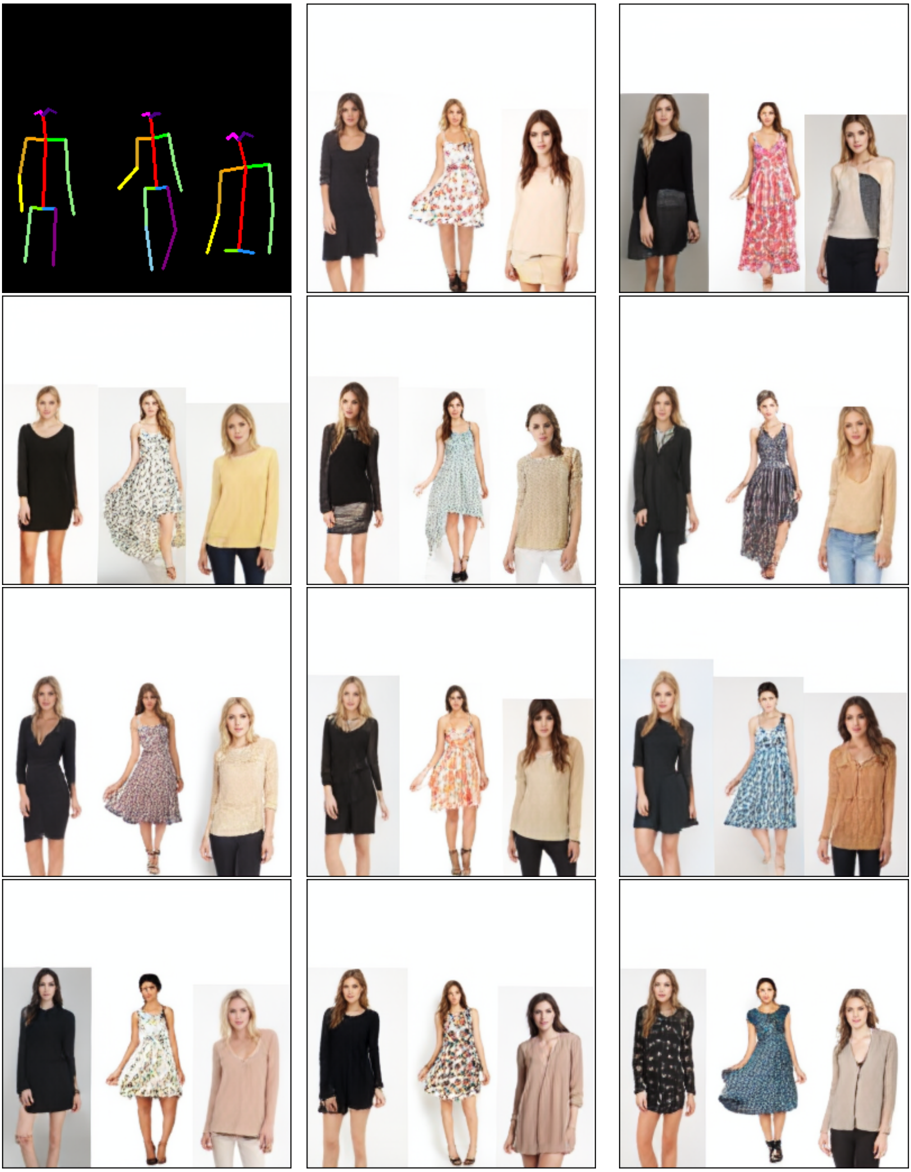

Fig. 10: ‘the lady wears a black long-sleeved romper.the lady wore a multicolor sleeveless dress.the lady is wearing a blouse with a long sleeved khaki.’

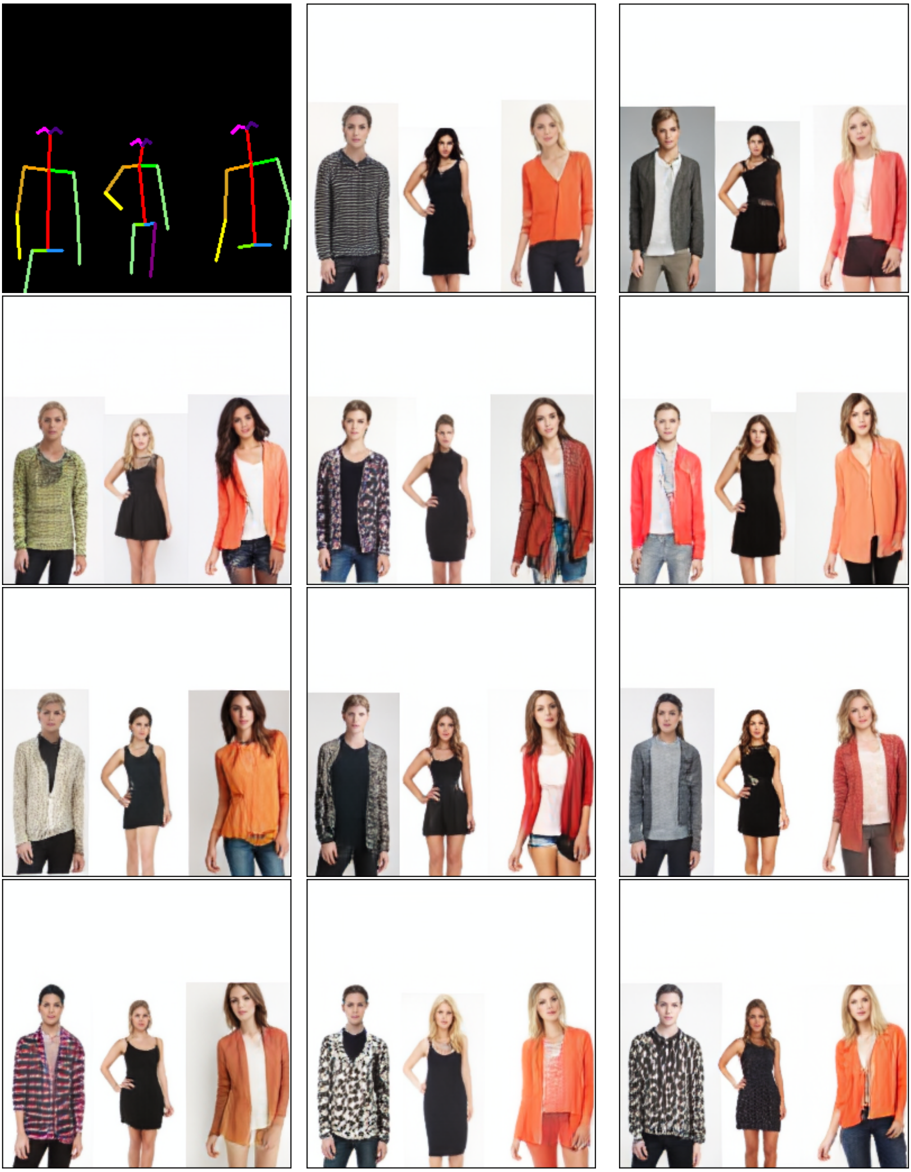

Fig. 11: ‘the man wore a cardigan with a multicolor long sleeve.the lady wore a black sleeveless dress.the lady wore a long orange blazer.’