Towards fine-grained spatial control for soccer game image generation

Supplementary Material

A. More details on Stable Diffusion finetuning

We explored several methods to finetune Stable Diffusion for soccer image generation. We first tried *full-finetuning*, which is simply updating all the weight of the model to finetune. In order to keep the prior knowledge of Stable Diffusion we also explored finetuning by only updating some blocks of the network as suggested by [1]. Hence, we tested two approaches, one in which we only update the encoder blocks and the middle block of the UNet, and another one in which we only update the decoder blocks. We additionally tested the LoRA finetuning [23] on the Self/Cross-Attention layers of the UNet. Finally, we tested the Visual Prompt Tuning method of [26] to learn a set of tokens that will better ”describe” a soccer game image than the unique prompt we are using in this work. Full-finetuning is the most naive approach but was giving most qualitative samples. One should note that this finetuning stage is done once for all. The finetuned Stable Diffusion can be subsequently used for any type of control.

B. Ablation on the heatmap for calibration control

To assess the effectiveness of our field heat map calibration control, we replace it with the binary mask of the pitch. Figure 7 shows the results for this experiment.

We can see on Figure 7 that both control signals allows to capture the overall orientation of the field. However, using the binary mask doesn’t capture the calibration. This is particularly true when the shape doesn’t give any clue on the area where we are on the pitch. In fact we can notice on the second row of the figure that the model is not generating the field markings at the same positions. Therefore, it is not consistent with the calibration. This is expected as the binary mask only contains the shape information. This experiment shows that our two dimensional heap map that serves as a kind of positional embedding is an effective way to control calibration.

C. Color control comparison

To further validate the design of our control model, we compare our model with the baselines on color-conditioned human image generation. We use the CIHP dataset [15], which is a human attribute segmentation dataset. It is one of the largest multi-person human parsing dataset with $38,280$ diverse human images. CIHP is annotated with rich information of person items. The images in this dataset are labeled with pixel-wise annotations on 20 categories and instance-level identification. We use the segmentation maps to extract the color of each segment as its average pixel color. We then train all methods on this dataset for 100K training steps with batch size 8 and learning rate 1×10^{-5} on a single NVIDIA A100 80G GPU. Figure 8 shows the results. We can clearly see that ControlNet and Uni-ControlNet mostly respect the shapes and do not follow very well the color instructions. In contrast, our approach of projecting conditioning images using the pretrained VAE of Stable Diffusion effectively follows color instructions.

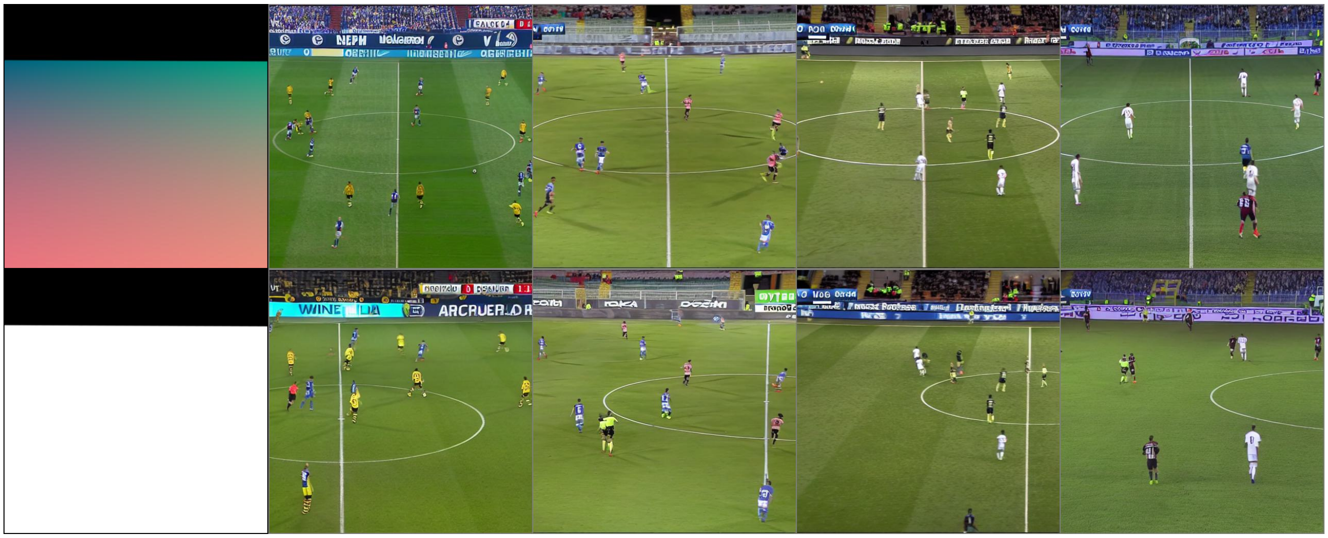

Figure 7. Comparison between calibration control with the proposed heat map and binary mask. In this case, the binary mask does not allow for precise field calibration.

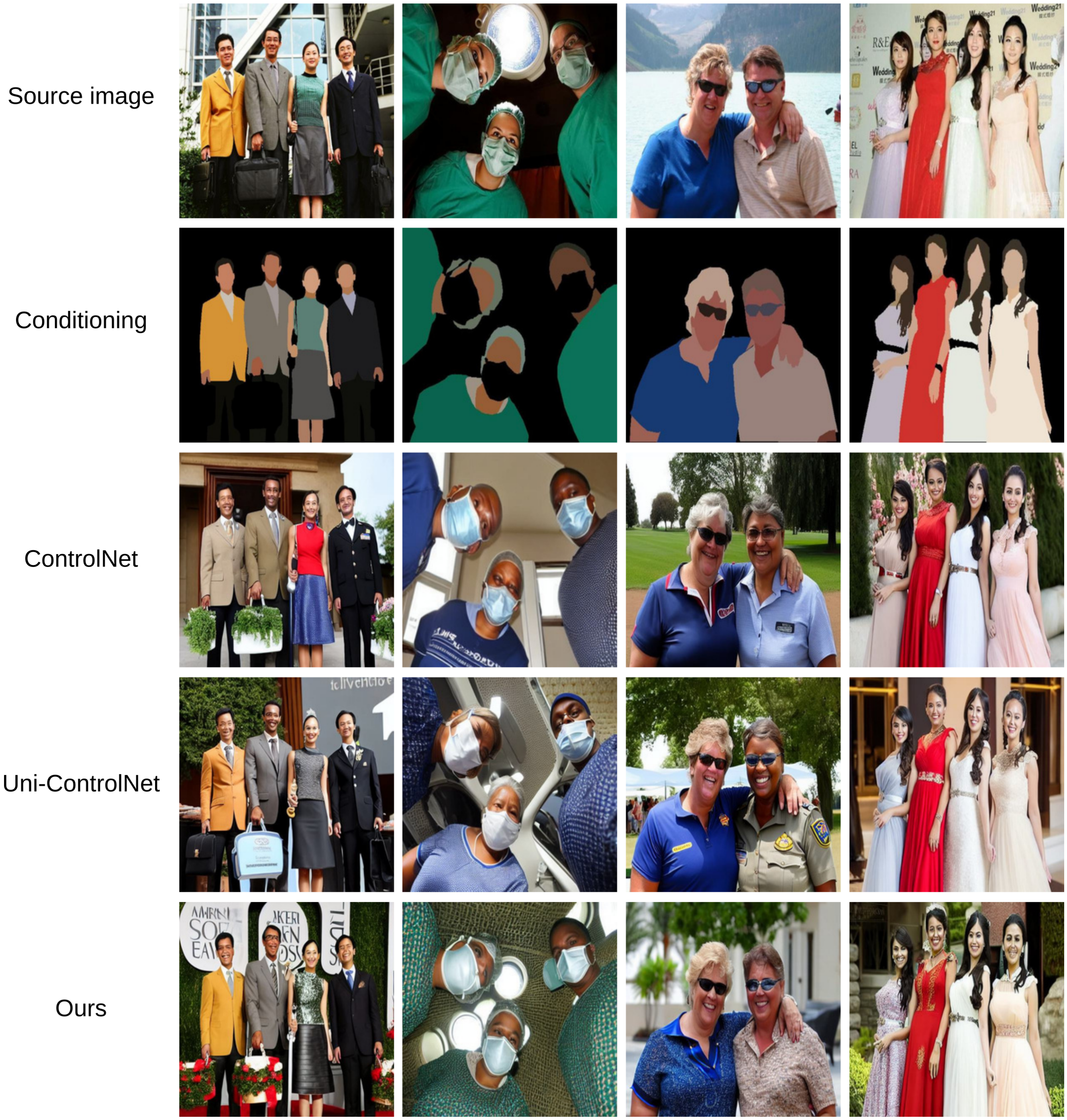

Figure 8. Comparison between the different methods of clothes color conditioning on human image generation. For each method and each conditioning, we generated 5 images and selected the best one. All samples have been obtained with classifier-guidance scale $\beta = 9.0$ and 50 denoising steps.