Fig. 1. More cases to clarity the "average collapse". Left: we train our model with our proposed contrastive loss. Right: we optimize the model just using CLIP similarity. The model optimized in a contrastive way could handle different text inputs while the other one fails to achieve that.

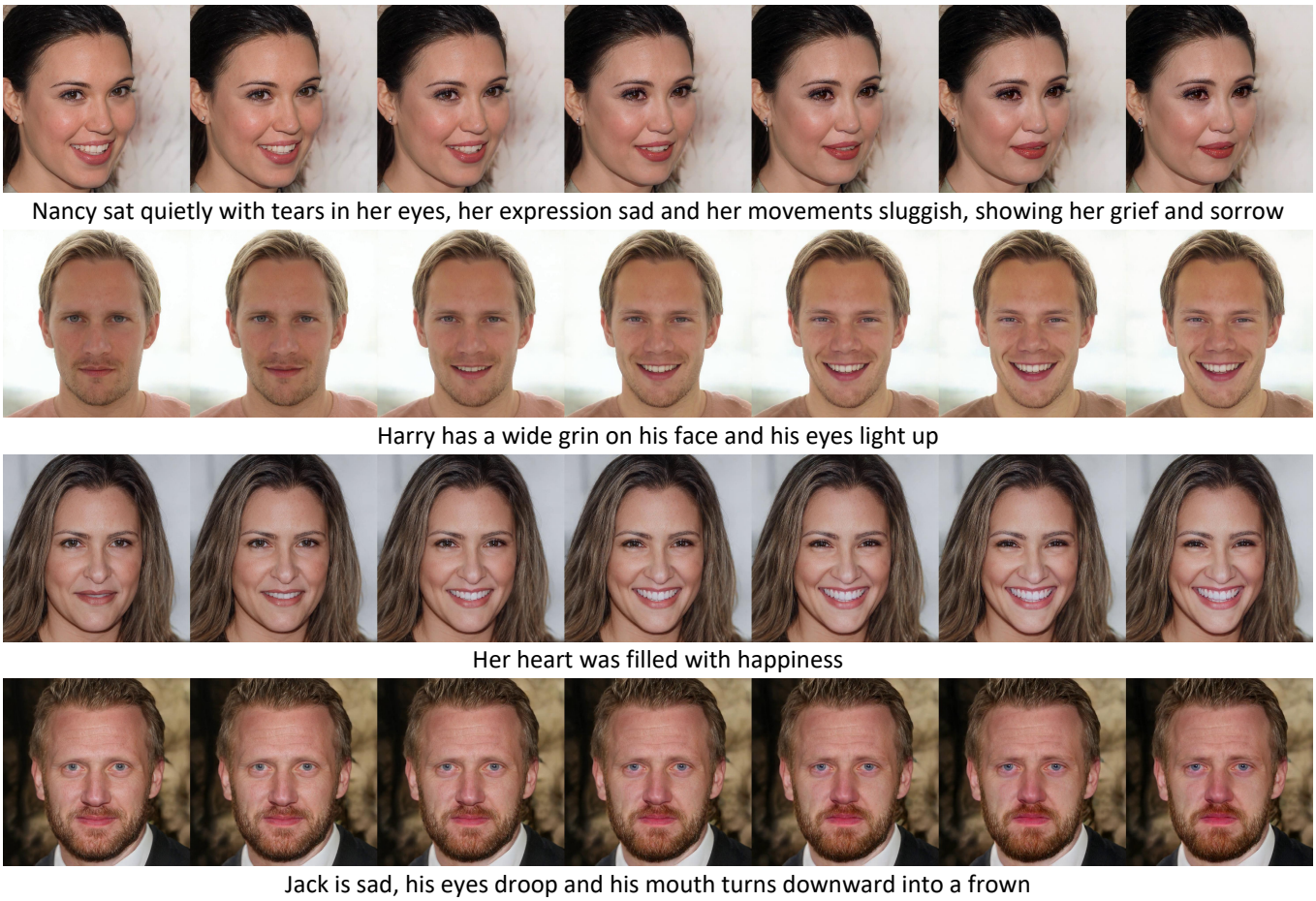

Fig. 2. More cases to show that our model can handle complex text prompts beyond "The person is".

Fig. 3. More cases to demonstrate our model's ablity to handle different attributes and facial expressions. All results are from one trained model.