ID-Patch: Robust ID Association for Group Photo Personalization

Supplementary Material

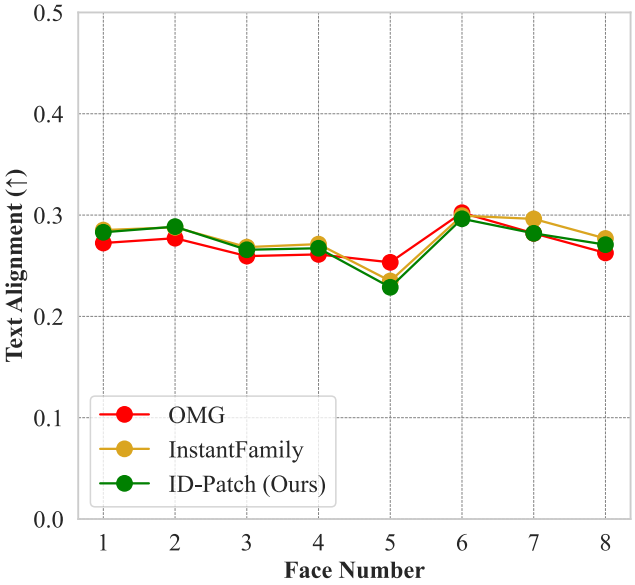

Figure A1. Performance evaluation of different model generations from the perspective of the text alignment score. There is no significant trend or difference between different number of faces. All methods achieve similar scores.

1. Additional Experiment Setups

1.1. Training Dataset

Our training set comprises a robust collection of images, totaling 17 million single-person and 1.95 million multi-person instances. These images are sourced from both purchased data and publicly available sources. Each image has been carefully cropped and resized according to predefined specifications and subsequently organized into distinct data buckets based on shape categories. BLIP-2 [21] is employed to obtain image captions.

Face localization in images is performed using the MTCNN framework [52]. For generating the pose conditions, we employ the HRNet-DEKR model [11] for pose estimation. To enhance the accuracy of this estimation, we compute the average distance between facial keypoints detected by MTCNN (considered as ground truth) and those estimated by HRNet-DEKR. Poses with a large keypoint distance are filtered out to ensure precision.

2. Additional Quantitive Results

Details regarding the text alignment scores with varying numbers of faces are illustrated in **Fig. A1**. It is evident that all three methods exhibit comparable performance in terms of text alignment. Beside, we do not observe a significant change of the text alignment score across different number of faces.

3. Additional Visualizations

Additional pose-free generated images of our proposed ID-Patch can be found in **Fig. A2**, where the generation is conditioned on individual face locations while corresponding head sizes and body poses are inferred implicitly. Our method can accommodate a large number of people with diverse ethnic backgrounds, yet generating visually appealing group photo results.

Additional comparisons of pose-conditioned generated images are available in **Fig. A3** and **Fig. A4**. In these visualizations, our proposed method, ID-Patch, consistently achieves robust ID association without ID leakage. In contrast, the other two methods experience significant ID leakage as the total number of faces in the generated images increases.

Figure A2. Additional visualizations of ID-Patch pose-free generation. Here the condition image is only ID patches rendered onto a black canvas without any pose condition.

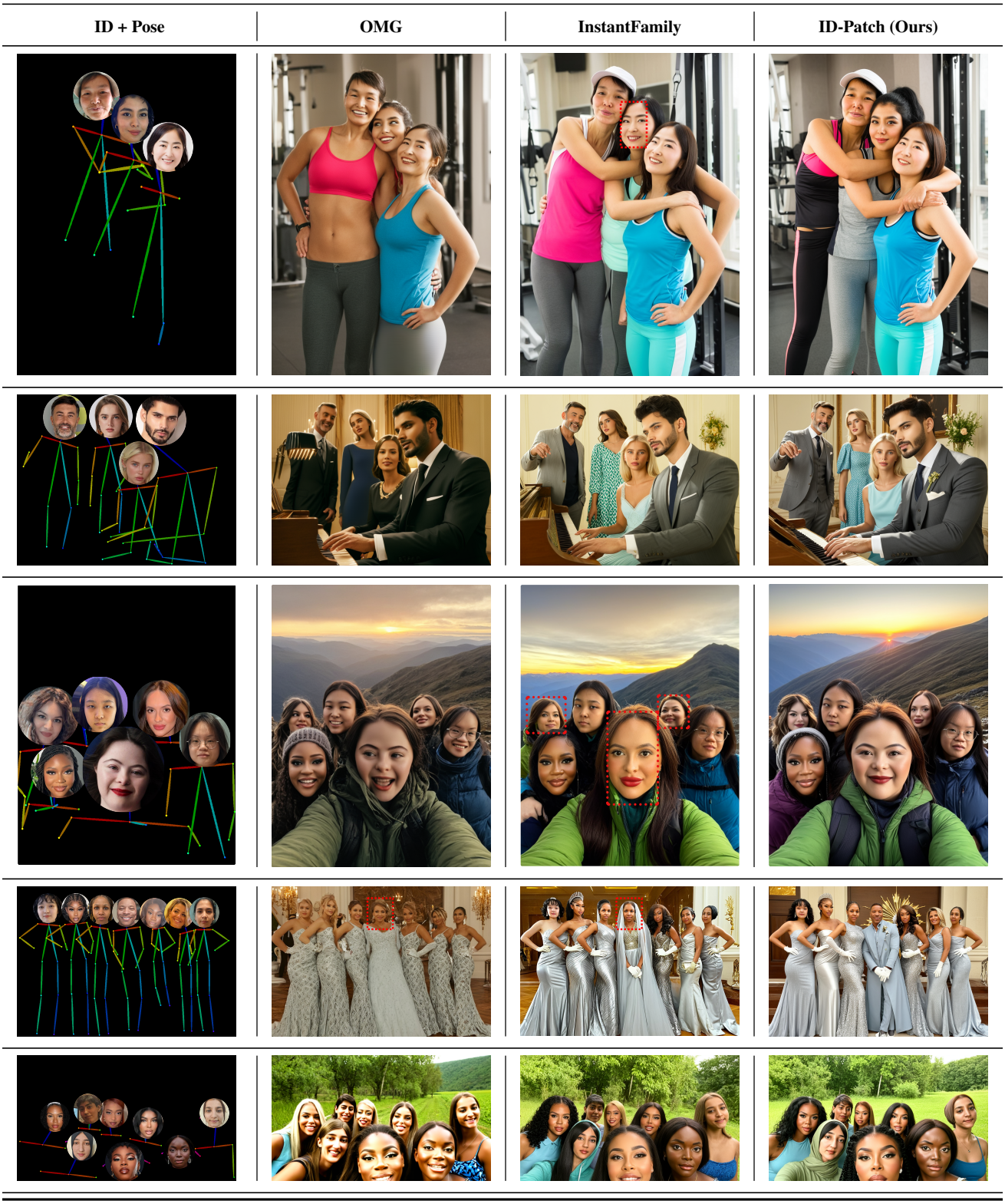

Figure A3. Part I: Additional comparison with baselines on pose-conditioned generation, where red dashed boxes highlight instances with low identity resemblance.

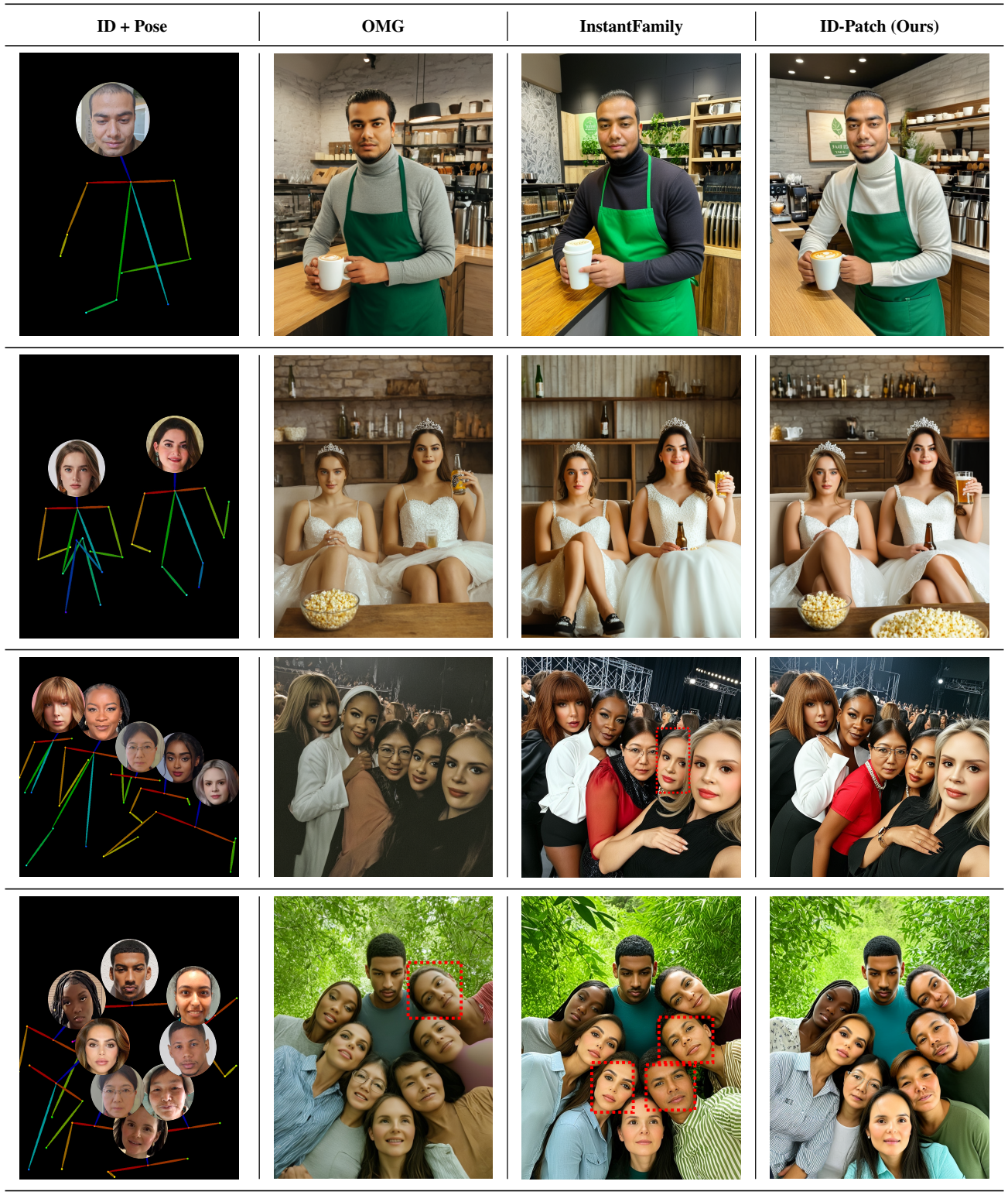

Figure A4. Part II: Additional comparison with baselines on pose-conditioned generation, where red dashed boxes highlight instances with low identity resemblance.