Supplementary Material: Towards a Simultaneous and Granular Identity-Expression Control in Personalized Face Generation

1. Study on Improved Midpoint Sampling

In training, we need $\mathbf{x}_0^*$ to impose the identity and expression constraints. The more accurate estimation of $\mathbf{x}_0^*$ is, the more accurate identity and expression losses are. We conduct an experiment on the dataset constructed by randomly selecting 500 image pairs from FFHQ, to evaluate three sampling methods: 1) one-step sampling using Eq. 1, 2) midpoint sampling used in [2], 3) our proposed improved midpoint sampling. MSE is used to measure the error between sampling results and ground truth, thus showing the image reconstruction performance.

As shown in Fig. 1, all sampling methods can decrease the reconstruction errors along with the training steps increasing. Our sampling method can achieve lower MSE than others in all periods. Fig. 2 shows the reconstruction results by denoising $\mathbf{x}_t$ using different sampling methods. Our results are not only more faithful to ground truth $\mathbf{x}_0$, but also more realistic and clear in the regions of eyes, mouths and even the reflection of sunglasses.

Fig. 3 shows the qualitative comparison of three sampling methods in generating final faces. In terms of identity preserving, the faces produced by our sampling method contain more facial details, e.g. wrinkles and whiskers, thus being more similar to source B. In terms of expression preservation, our result in the 3rd row exhibits less angry expression, thus being more consistent with the expression in source A. In the 4th row, our result displays the expression of slightly opening mouth, looking better than others.

Tab. 1 shows the quantitative comparison of three sampling methods. The improved midpoint sampling gets the best score in ID. and the second best score in Exp., which indicates it can impose more effective constraints on the training of diffusion model. Besides, the effectiveness of identity and expression losses can be validated by removing ID&Exp losses.

2. Challenging Cases

In Fig. 4, we show the face reenactment results under some challenging conditions, such as the significant differences in poses and lighting between source and target. Our method performs well under these challenging conditions.

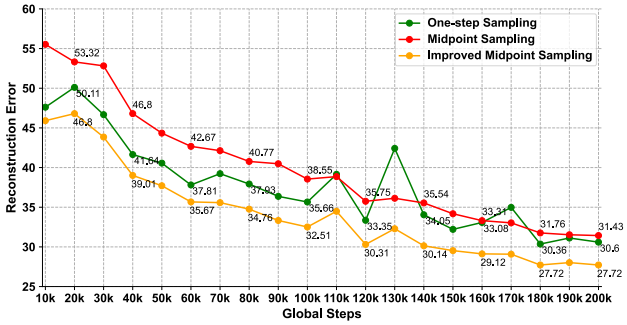


Figure 1. Image reconstruction performance of three sampling methods.

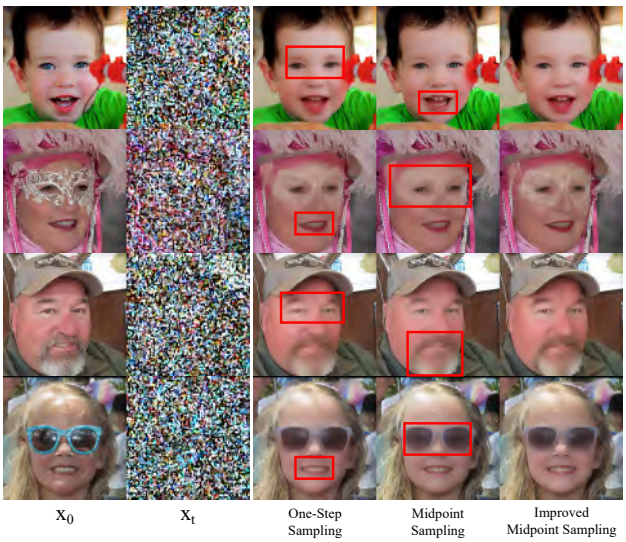


Figure 2. Image reconstruction results by using three sampling methods.

3. Fine-grained Expression Controlling Results

Fig. 6-39 shows all of the fine-grained expression synthesis results with 135 labels of expression text [1]. Readers can zoom in for more details. Due to the size limitations of the submitted files, these results are highly compressed so that some regions are distorted. The uncompressed results are available in the project homepage: https://diffsfsr.github.io/.

Methods	ID.↑	Exp.↓
w./o. ID&Exp Losses	67.5	0.71
One-step Sampling	83.1	**0.62**
Midpoint Sampling [2]	74.0	0.70
Improved Midpoint Sampling (ours)	**87.0**	0.63

Table 1. Quantitative results of using three sampling methods. All values are scaled up by a factor of 100 for simplicity.

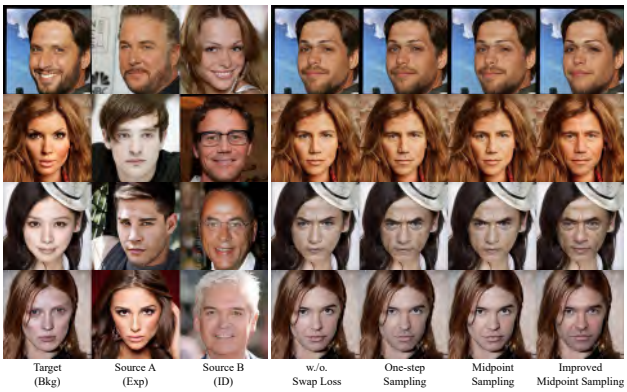


Figure 3. SFSR results by using three sampling methods.

Figure 4. Face reenactment under challenging conditions. The source faces are totally different from the target faces in lighting (1st row) and poses (2nd and 3rd rows).

4. Limitations

Although the facial expressions between reference and synthesized images are close to each other in our framework, it can be found that their facial expressions could not fully reflect the semantic information by the text label. For example, the text label "agitation" is not consistent with the reference image in the 4th row of Fig. 6. This can be attributed to the flaws from the dataset [1], which cannot guarantee that all images can fully display their corresponding expression labels. Additionally, these expression labels have more or less ambiguity among them, leading to the overlapping of their semantics. What's more, as shown in Fig. 5, if the face is wrongly detected, our method will still process it. If no face is detected, our method will be terminated. Our method also inherits the limitations of Stable Diffusion, such as artifacts in teeth and limbs.

Figure 5. Some typical failure cases.

5. Ethical Statement

Our method offers certain contributions and inspiration in both academia and industry. Everyone can rely on our method to quickly customize their photos. However, we recognize the ethical issues associated with the ability to generate human images with high fidelity. The spread of this technology could lead to malicious tampering with images and the dissemination of false information. We therefore emphasize the importance of developing and following ethical guidelines and using this technology responsibly. We hope that our contribution will foster further discussion and research on the safe and ethical use of computer vision.

References

[1] Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding. Semantic-rich facial emotional expression recognition. *IEEE Trans. Affect. Comput.*, 13(4):1906–1916, 2022. 1, 2

[2] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proc. of CVPR*, pages 8568–8577, 2023. 1, 2

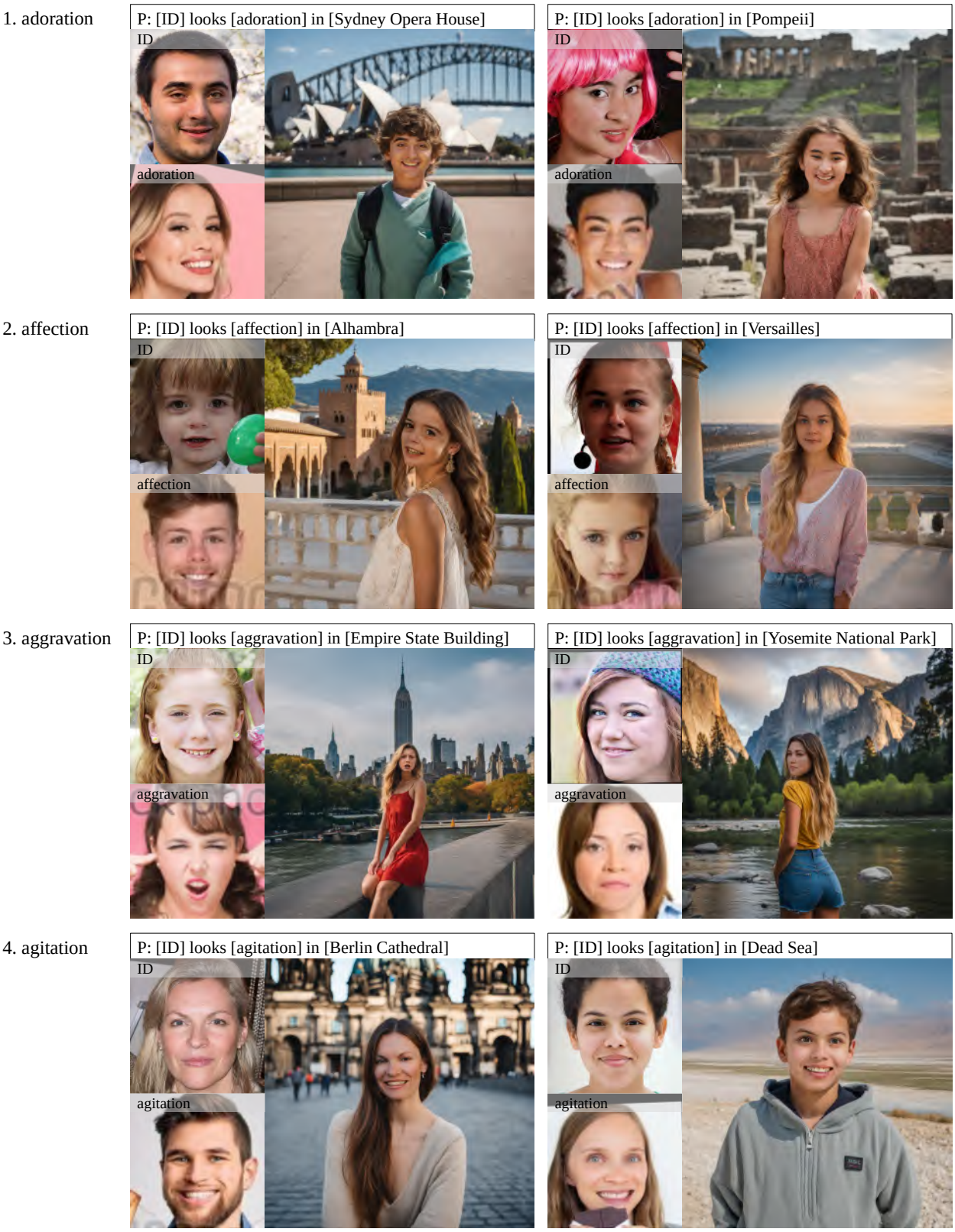


Figure 6. Resulting samples of the full set of 135 expression labels. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

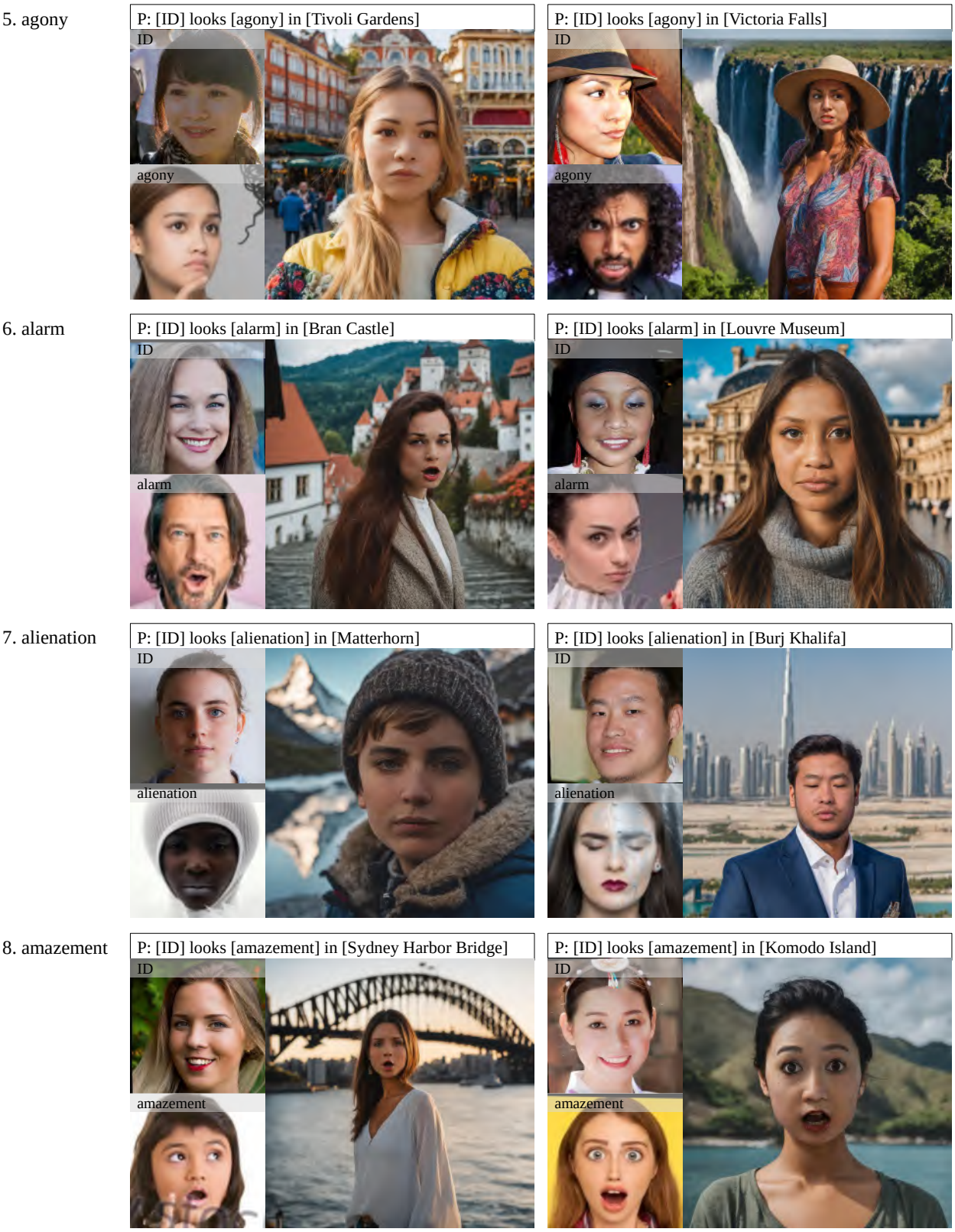


Figure 7. Continues from Figures 6. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

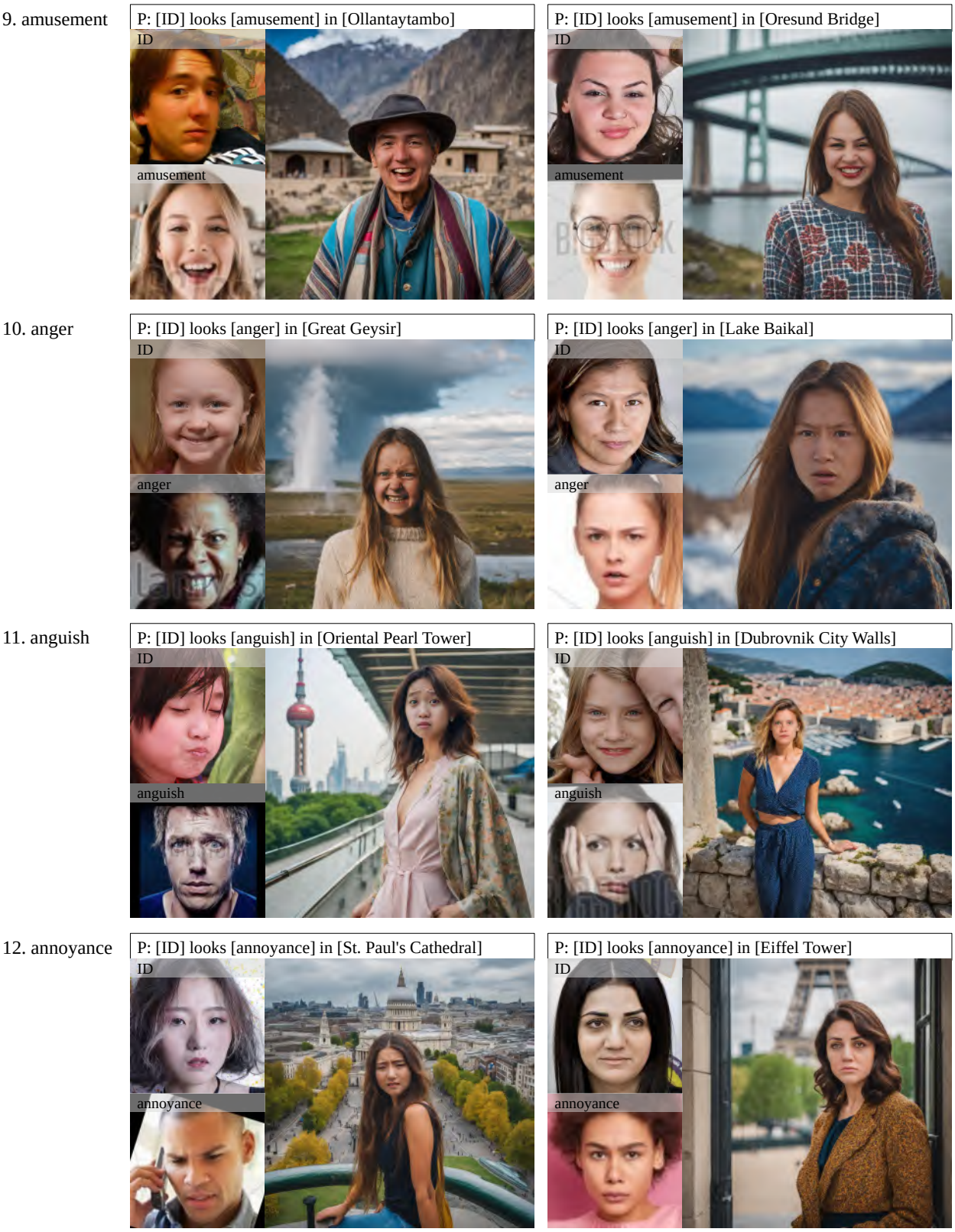
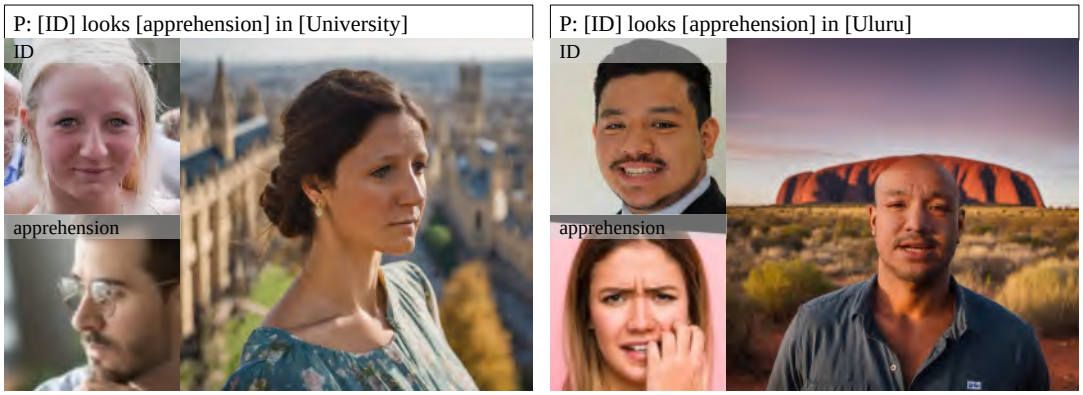


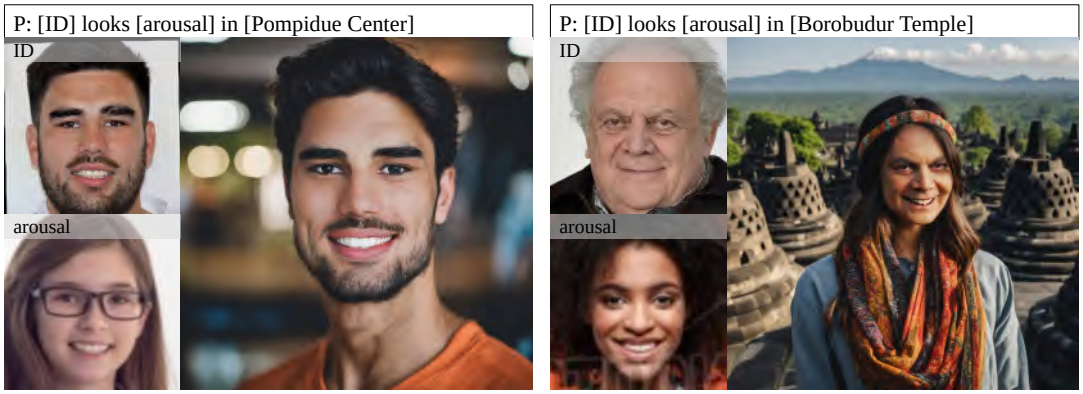
Figure 8. Continues from Figures 6-7. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

13. anxiety

14. apprehension



15. arousal



16. astonishment

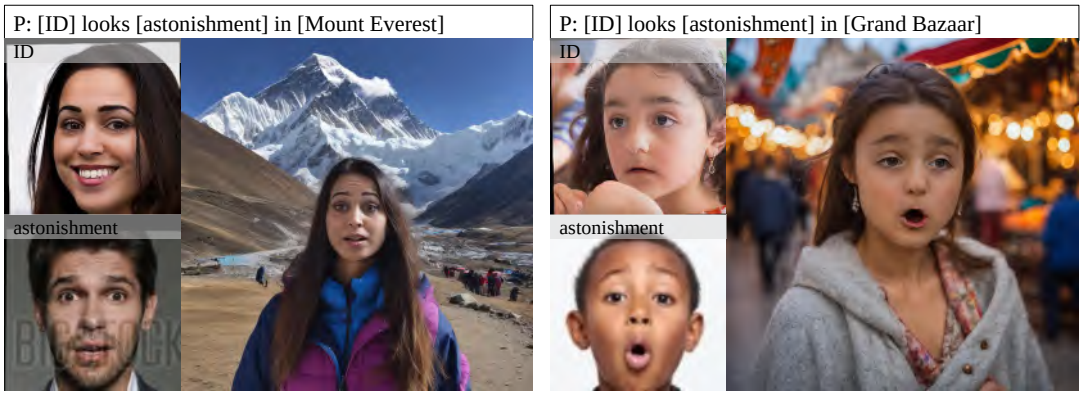


Figure 9. Continues from Figures 6-8. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

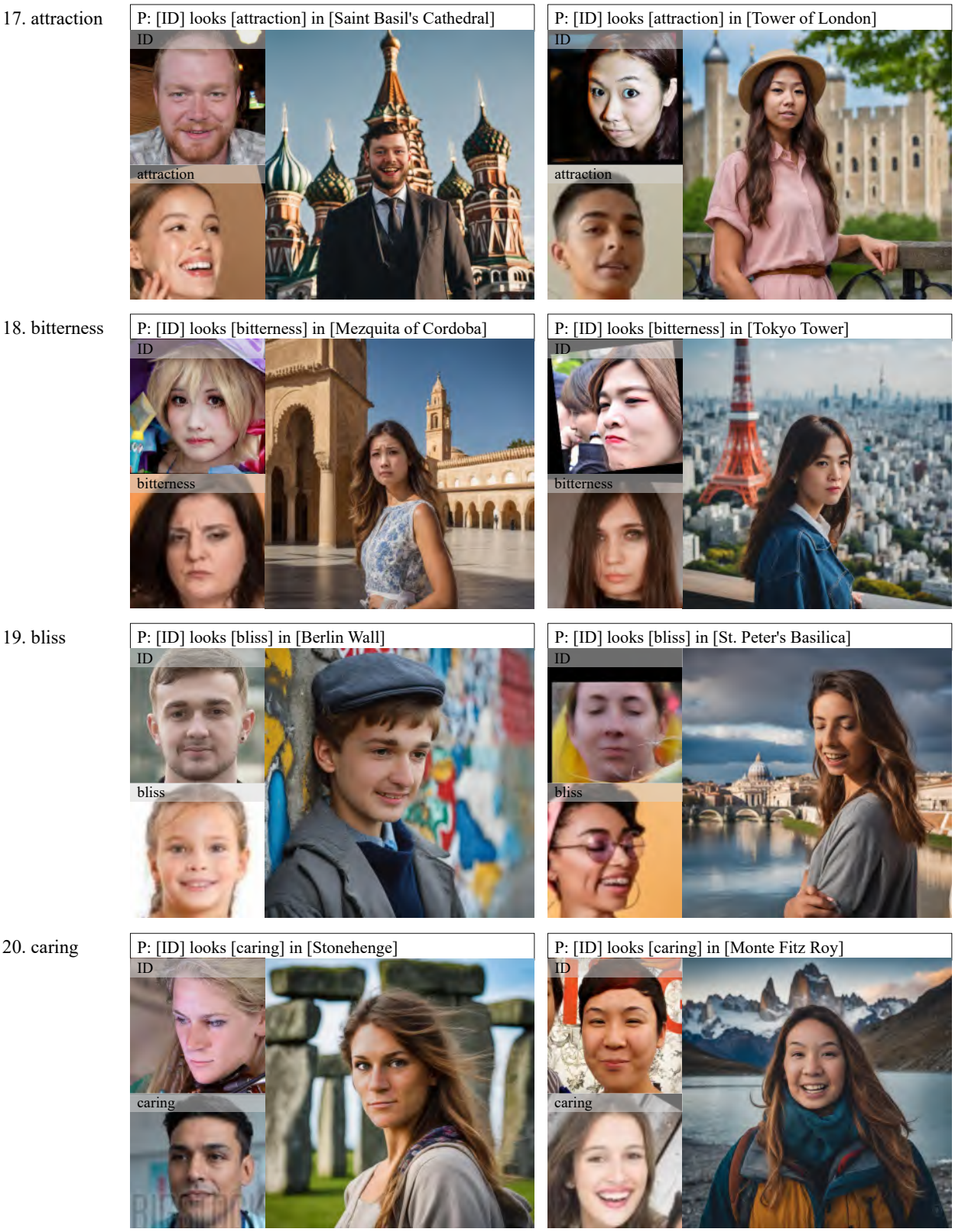


Figure 10. Continues from Figures 6-9. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

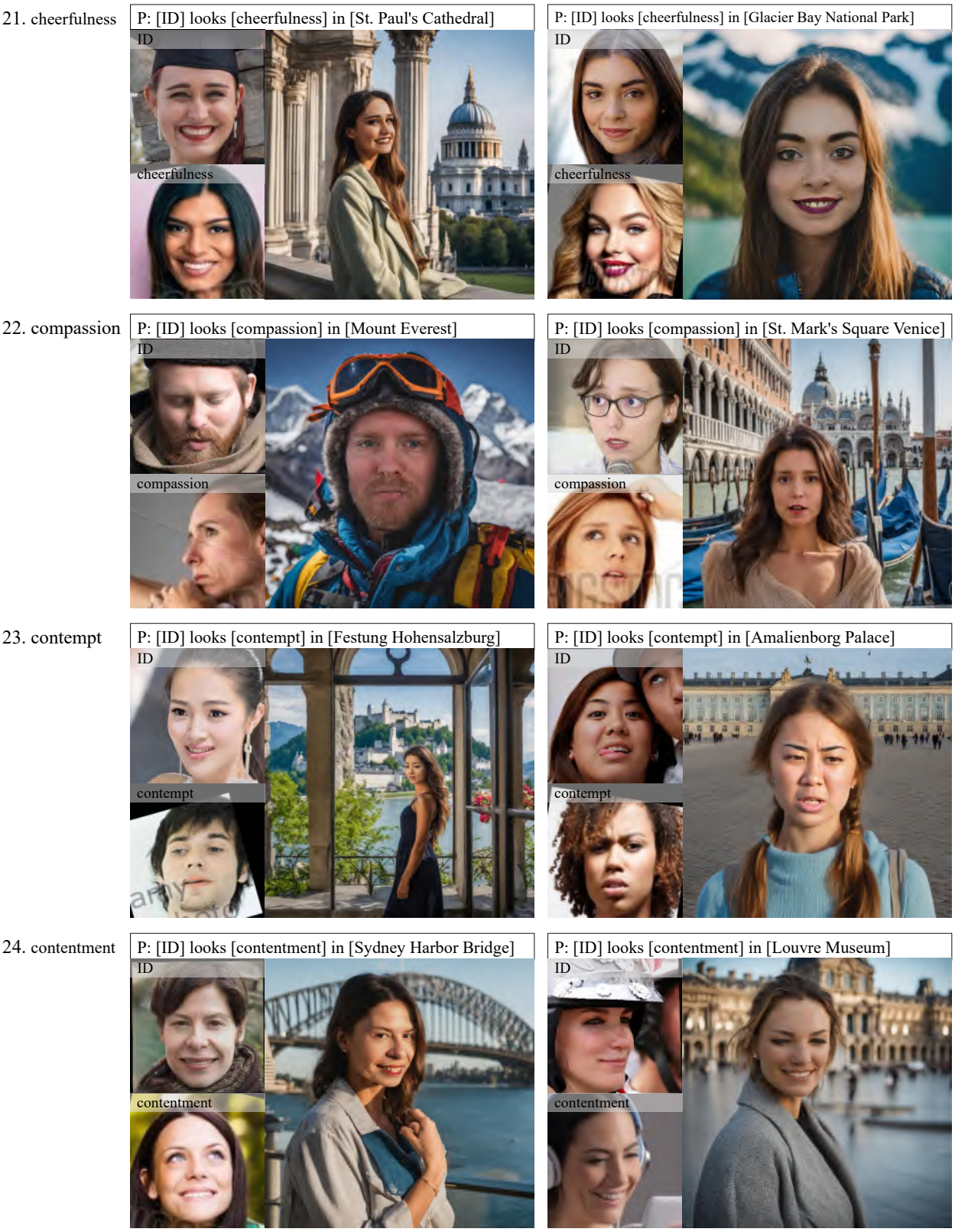
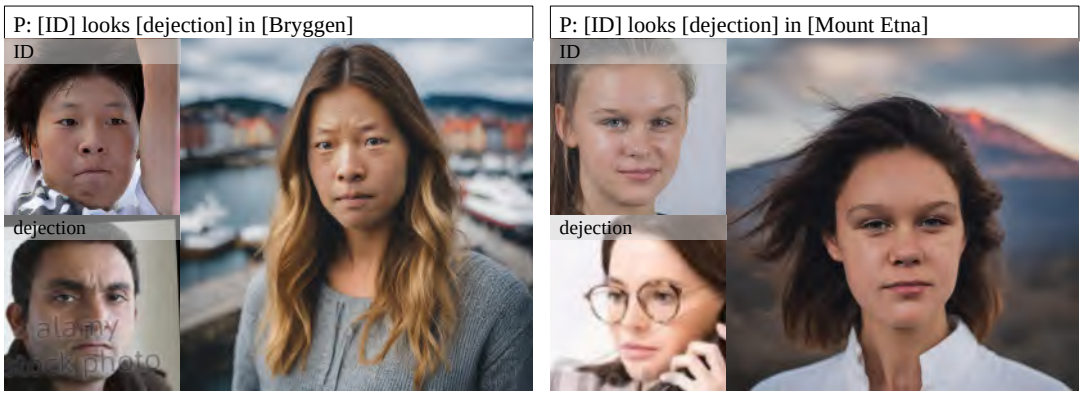


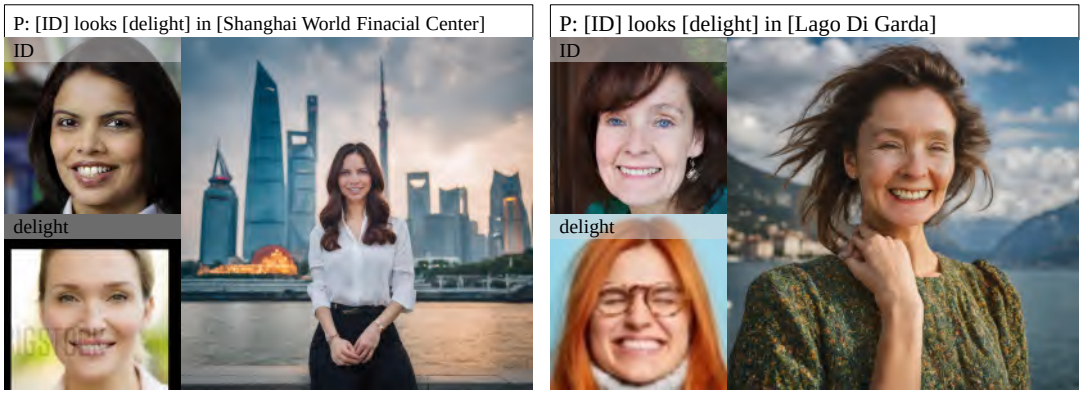
Figure 11. Continues from Figures 6-10. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

25. defeat

26. dejection



27. delight



28. depression

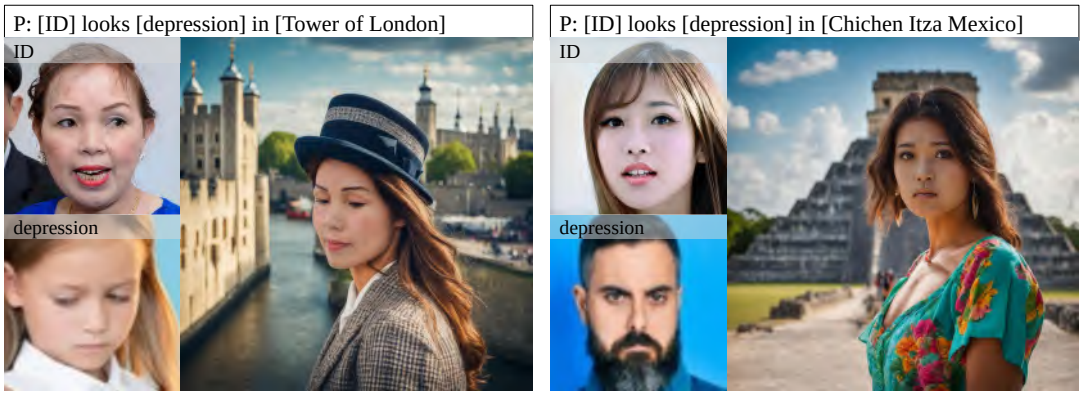
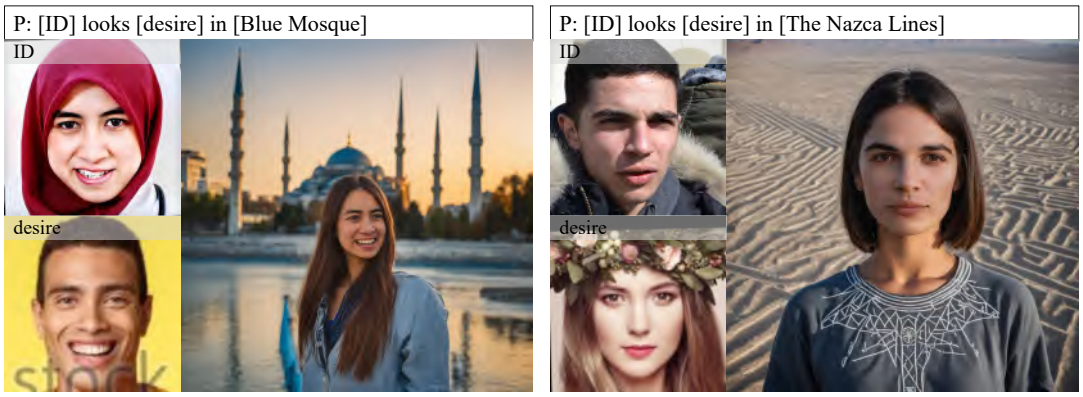


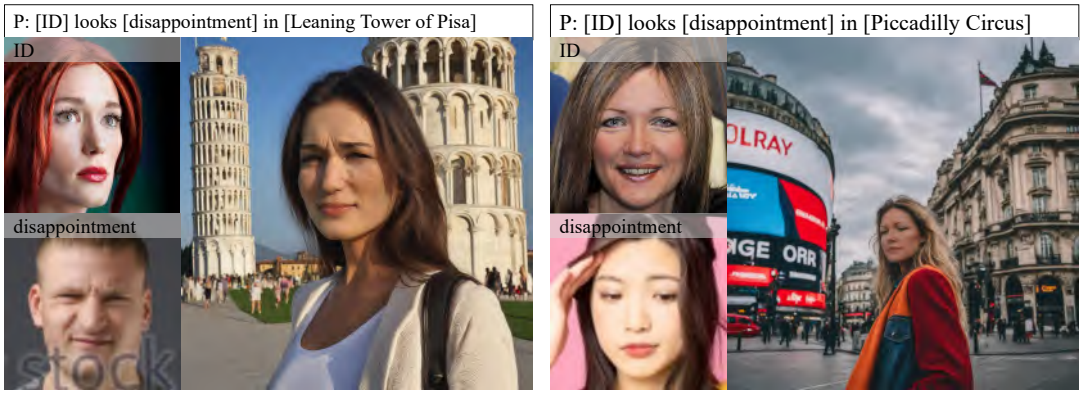
Figure 12. Continues from Figures 6-11. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

29. desire



30. despair

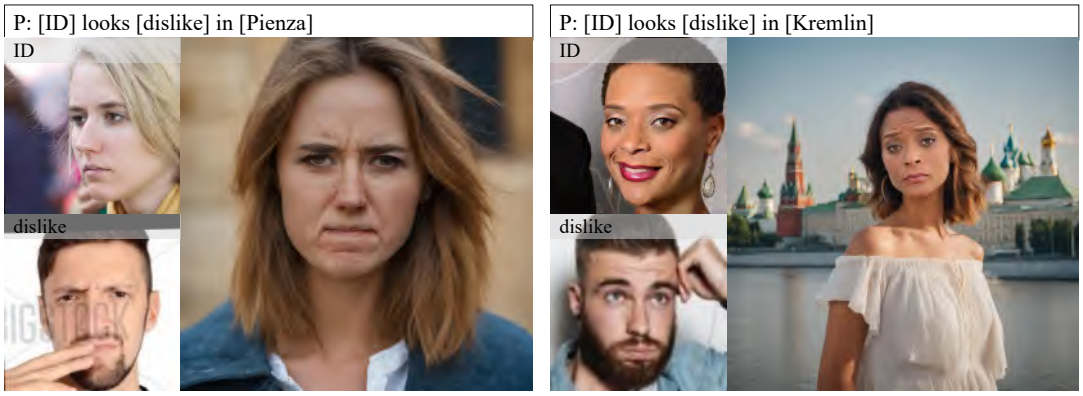
31. disappointment



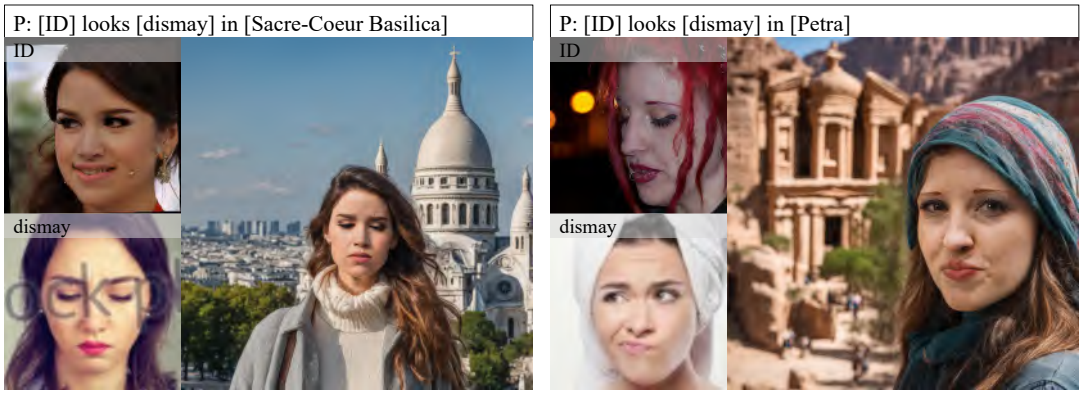
32. disgust

Figure 13. Continues from Figures 6-12. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

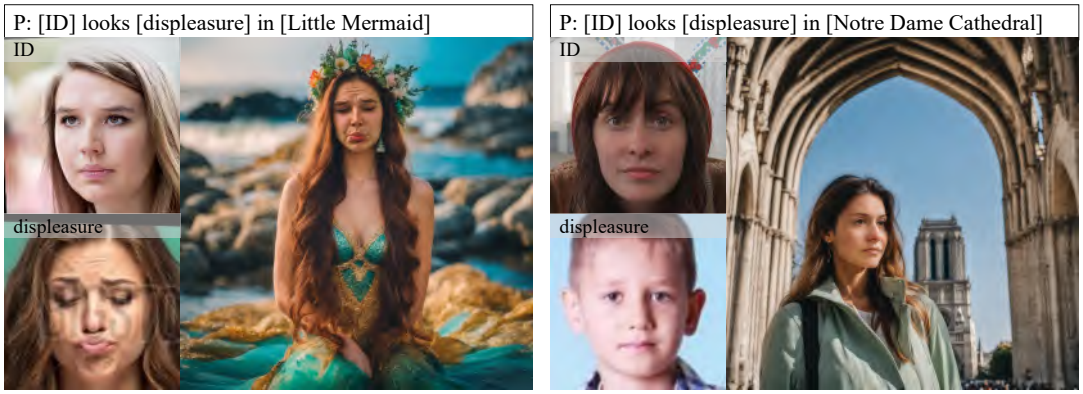
33. dislike



34. dismay



35. displeasure



36. distress

Figure 14. Continues from Figures 6-13. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

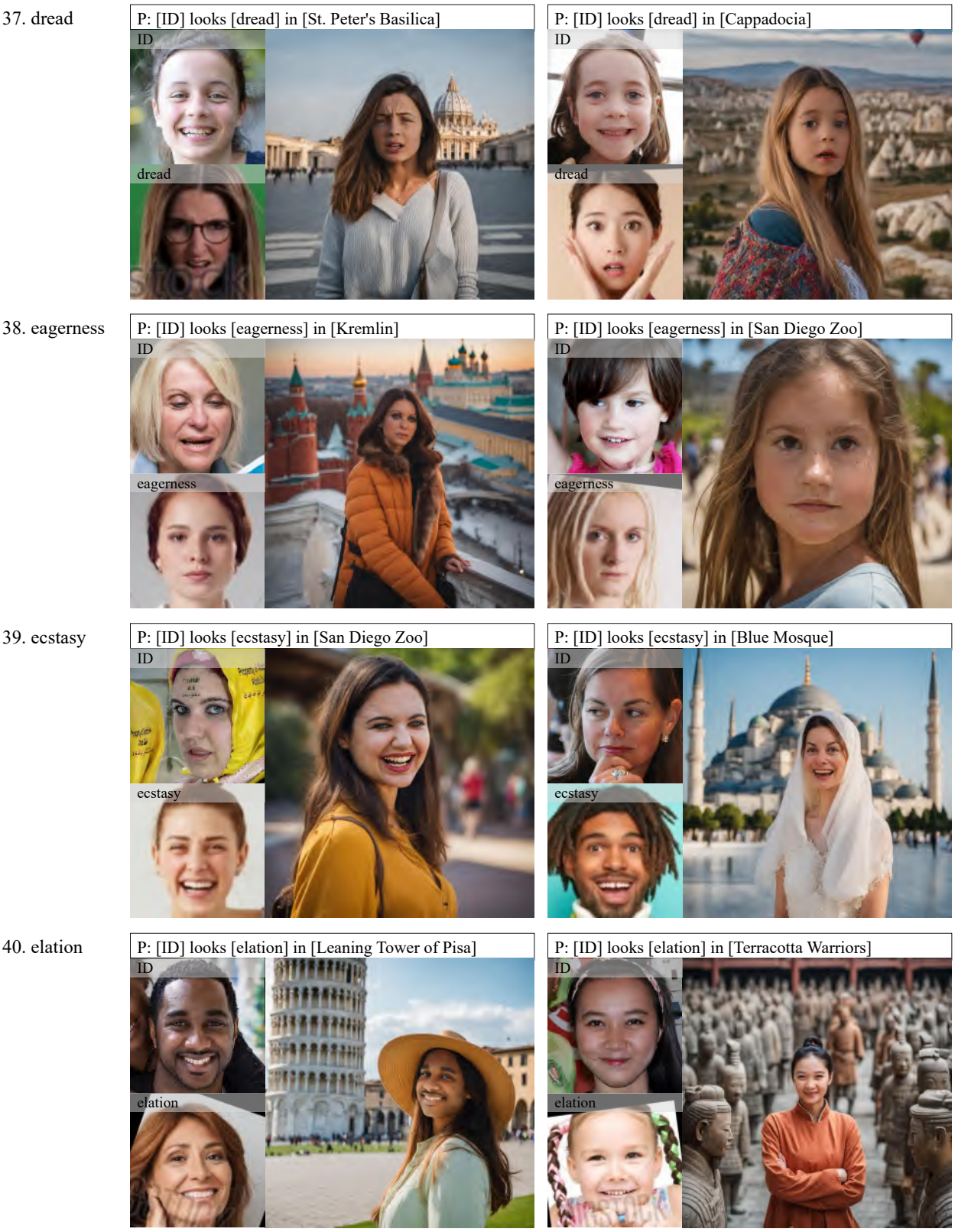
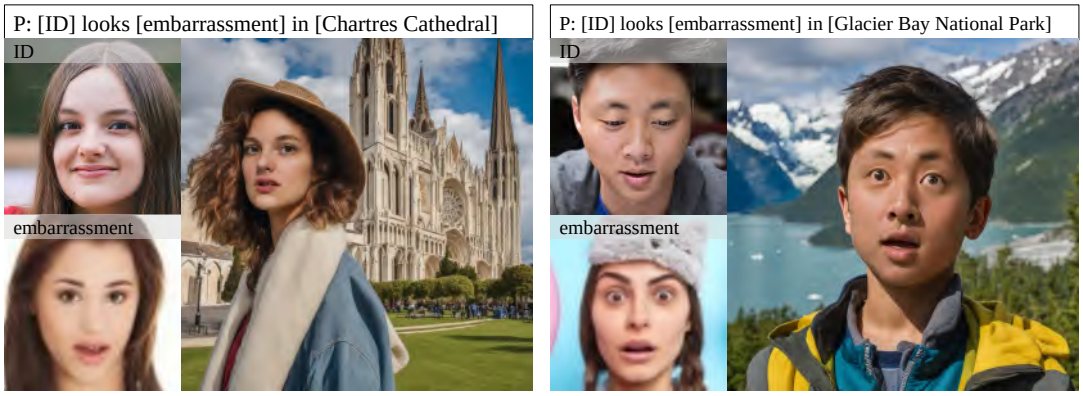
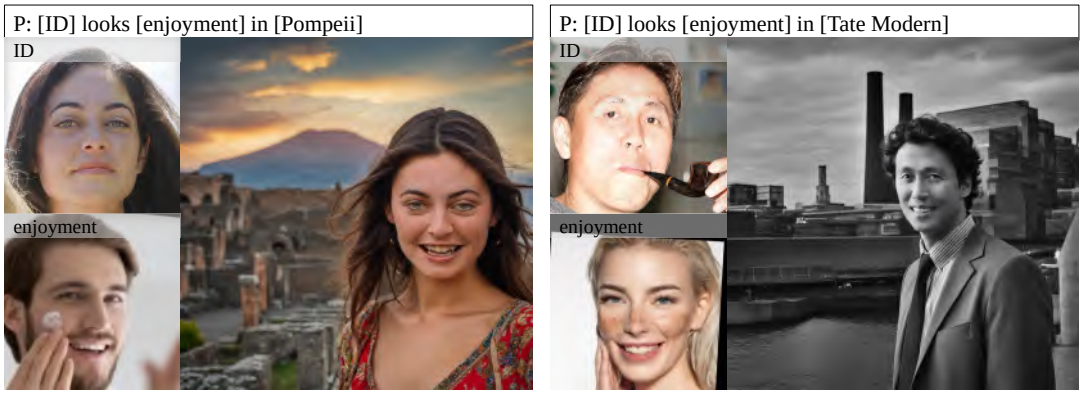


Figure 15. Continues from Figures 6-14. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

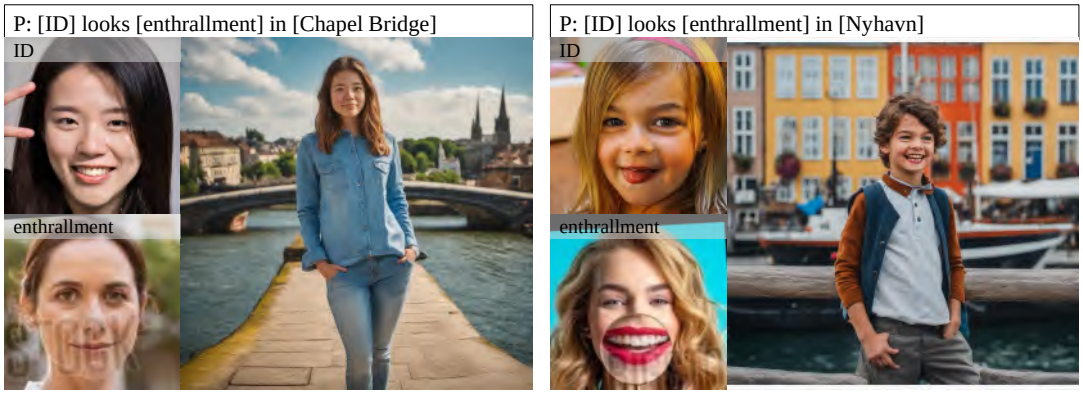
41. embarrassment



42. enjoyment



43. enthrallment



44. enthusiasm

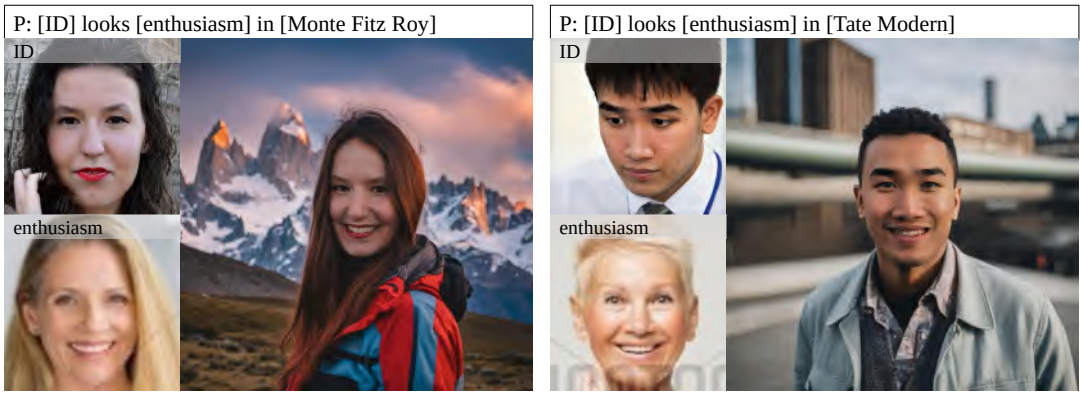


Figure 16. Continues from Figures 6-15. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

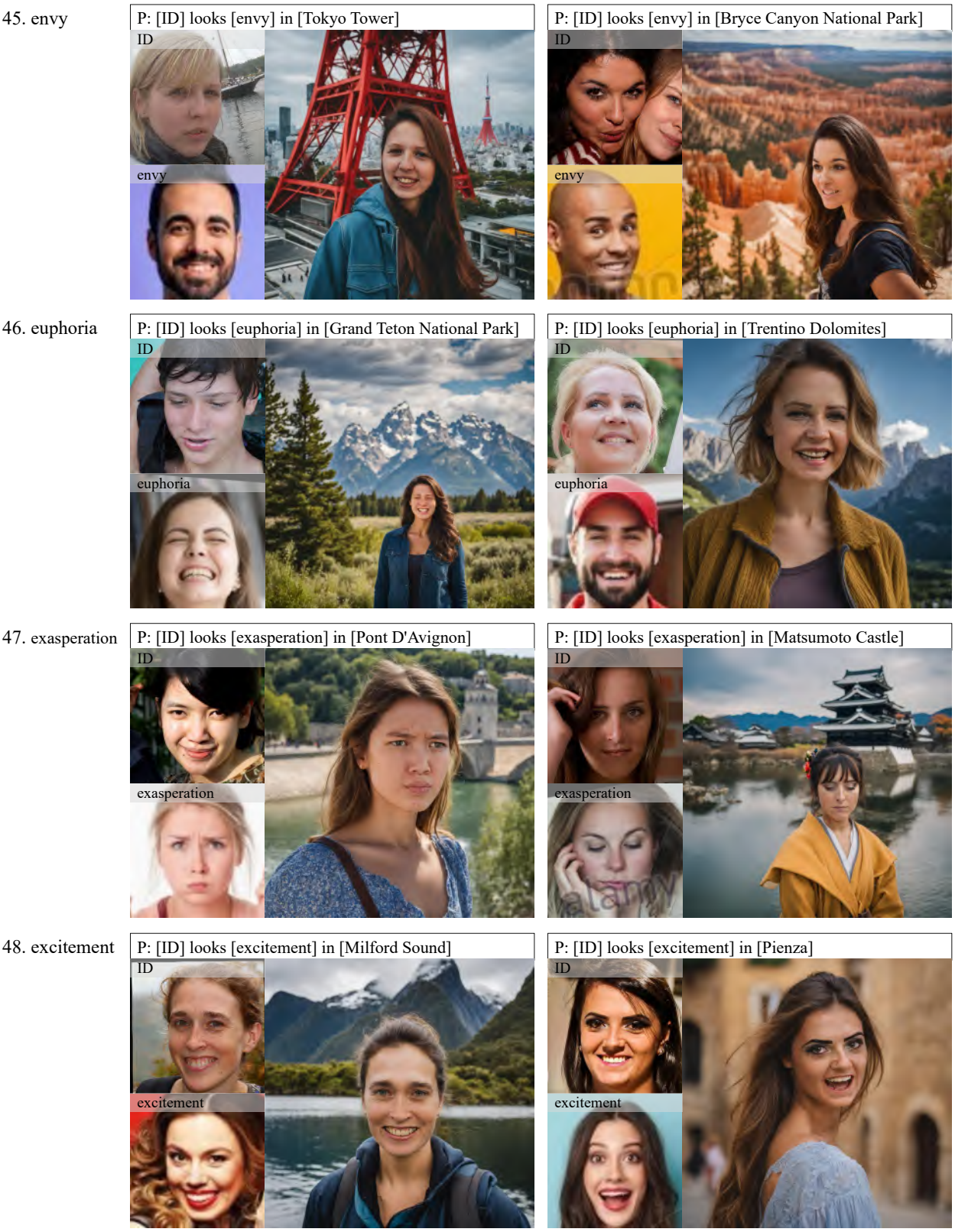


Figure 17. Continues from Figures 6-16. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

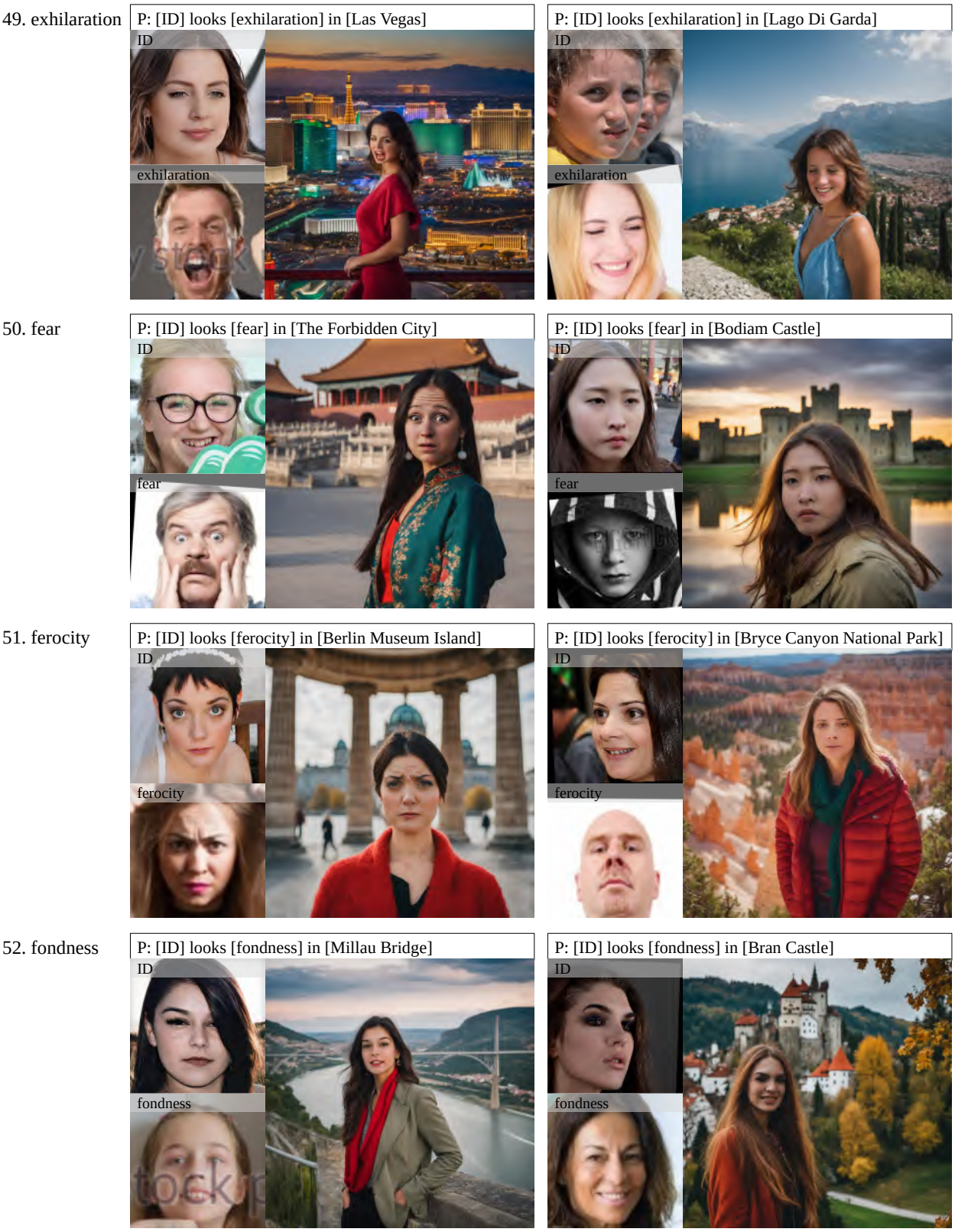
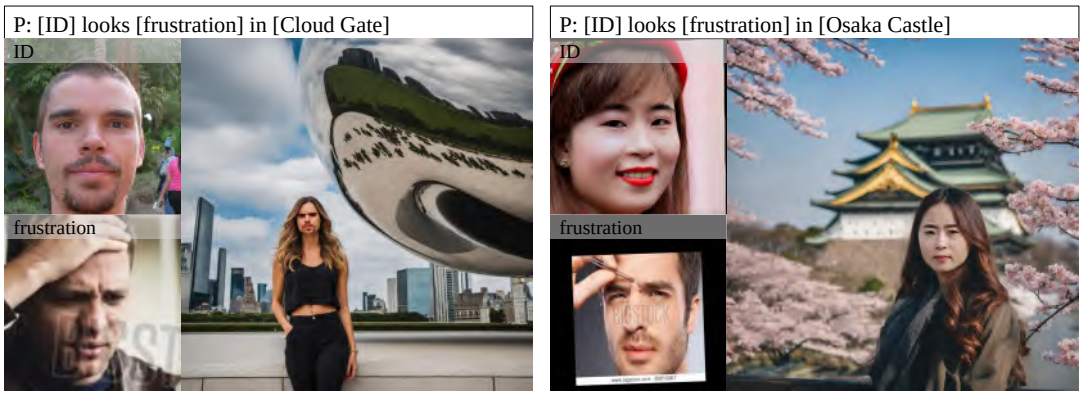


Figure 18. Continues from Figures 6-17. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

53. fright

54. frustration



55. fury

56. gaiety

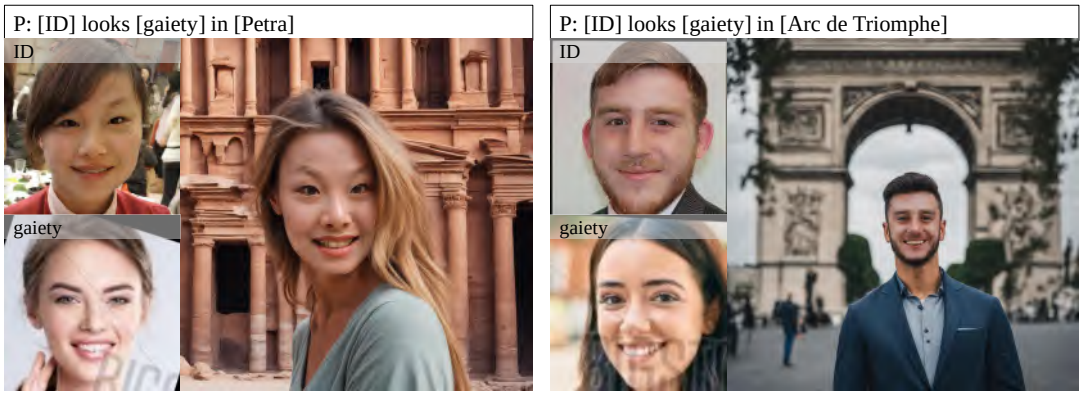
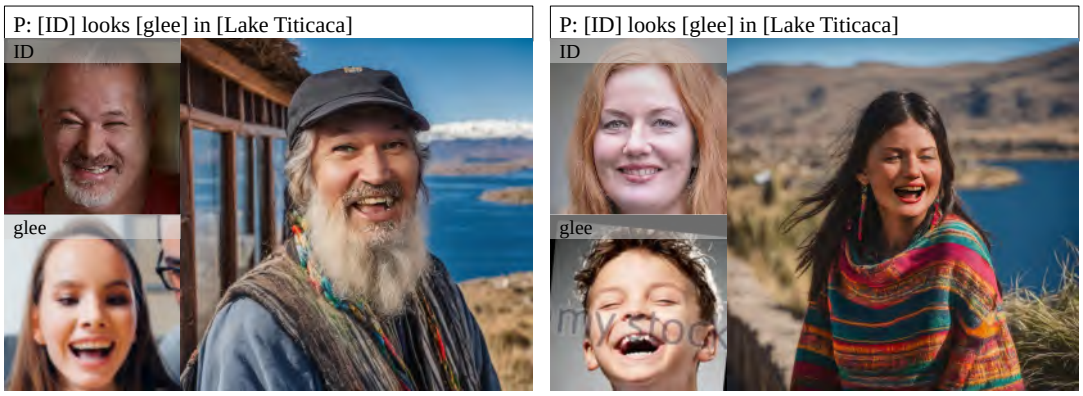


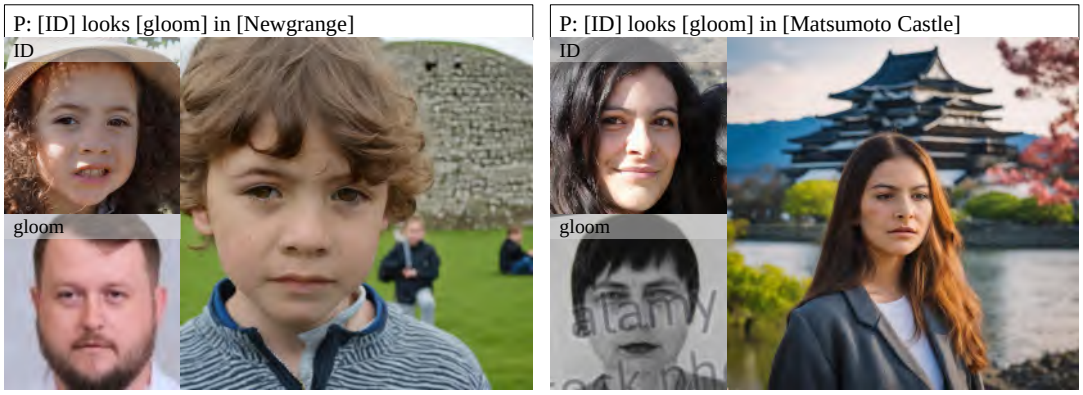
Figure 19. Continues from Figures 6-18. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

57. gladness

58. glee



59. gloom



60. glumness

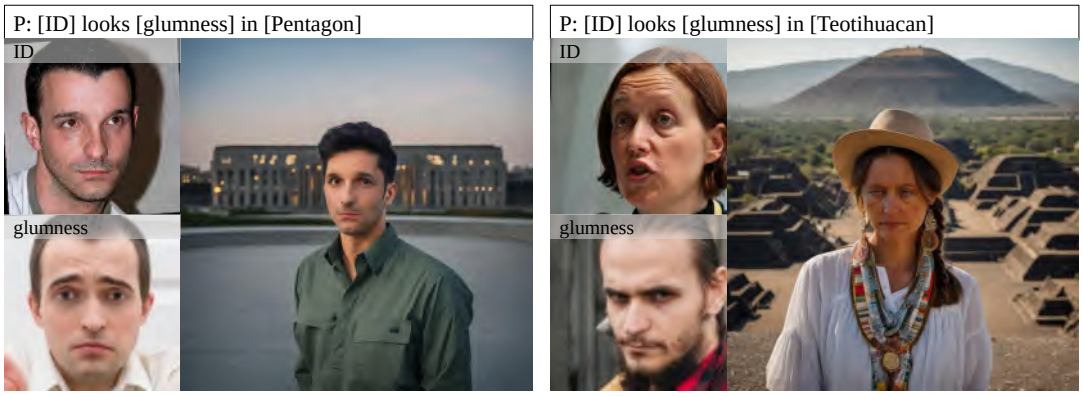
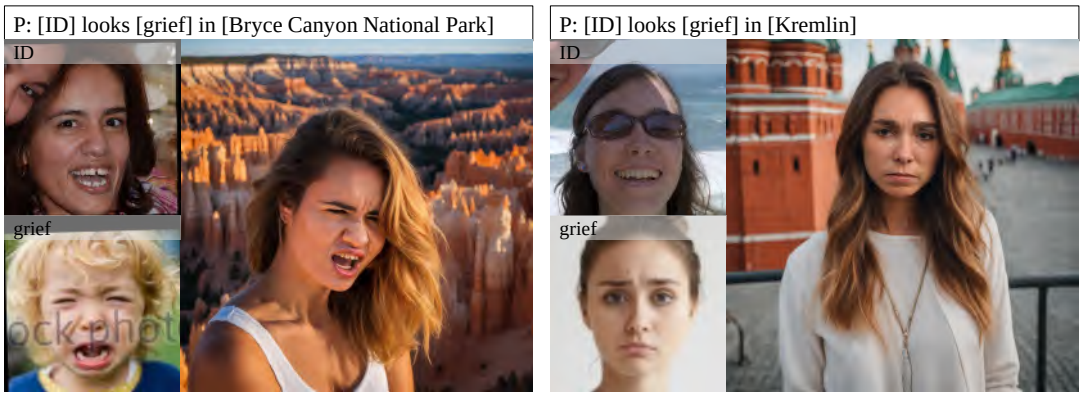


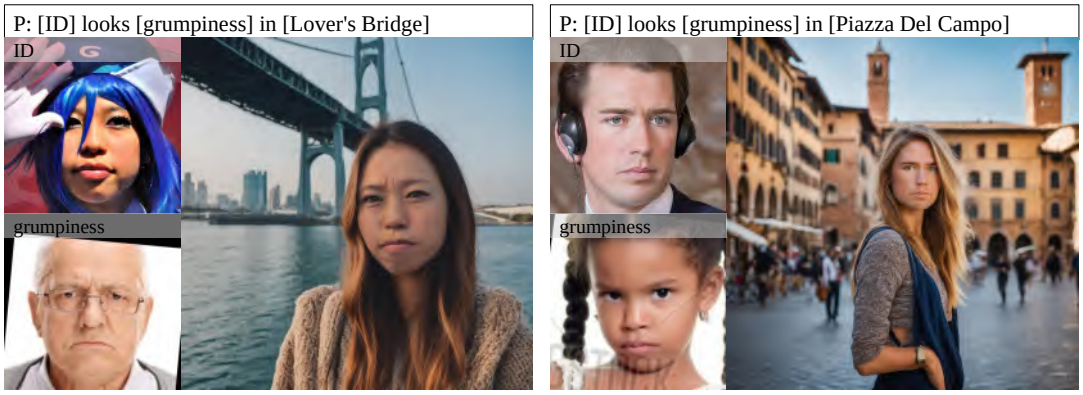
Figure 20. Continues from Figures 6-19. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

61. grief



62. grouchiness

63. grumpiness



64. guilt

Figure 21. Continues from Figures 6-20. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

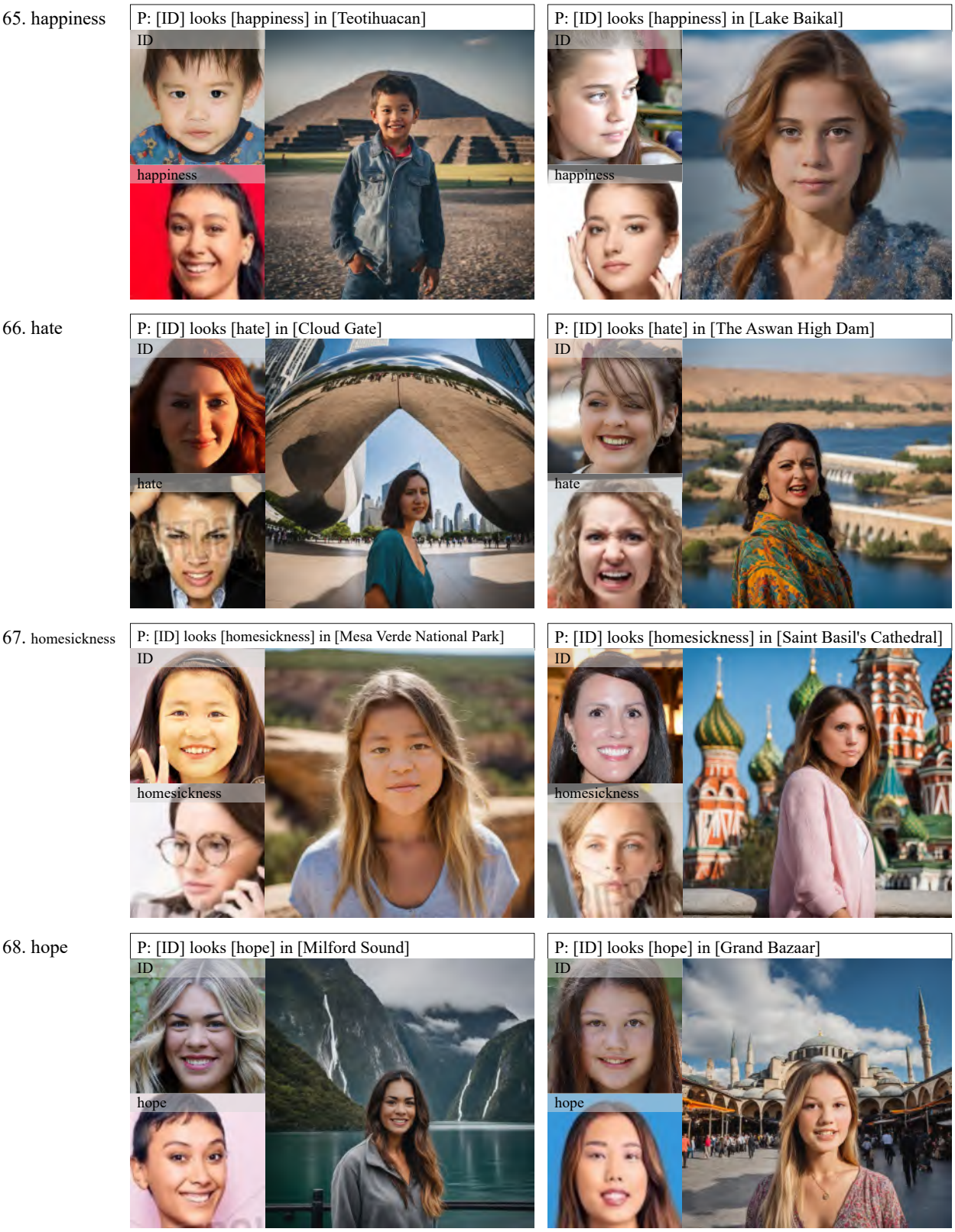


Figure 22. Continues from Figures 6-21. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

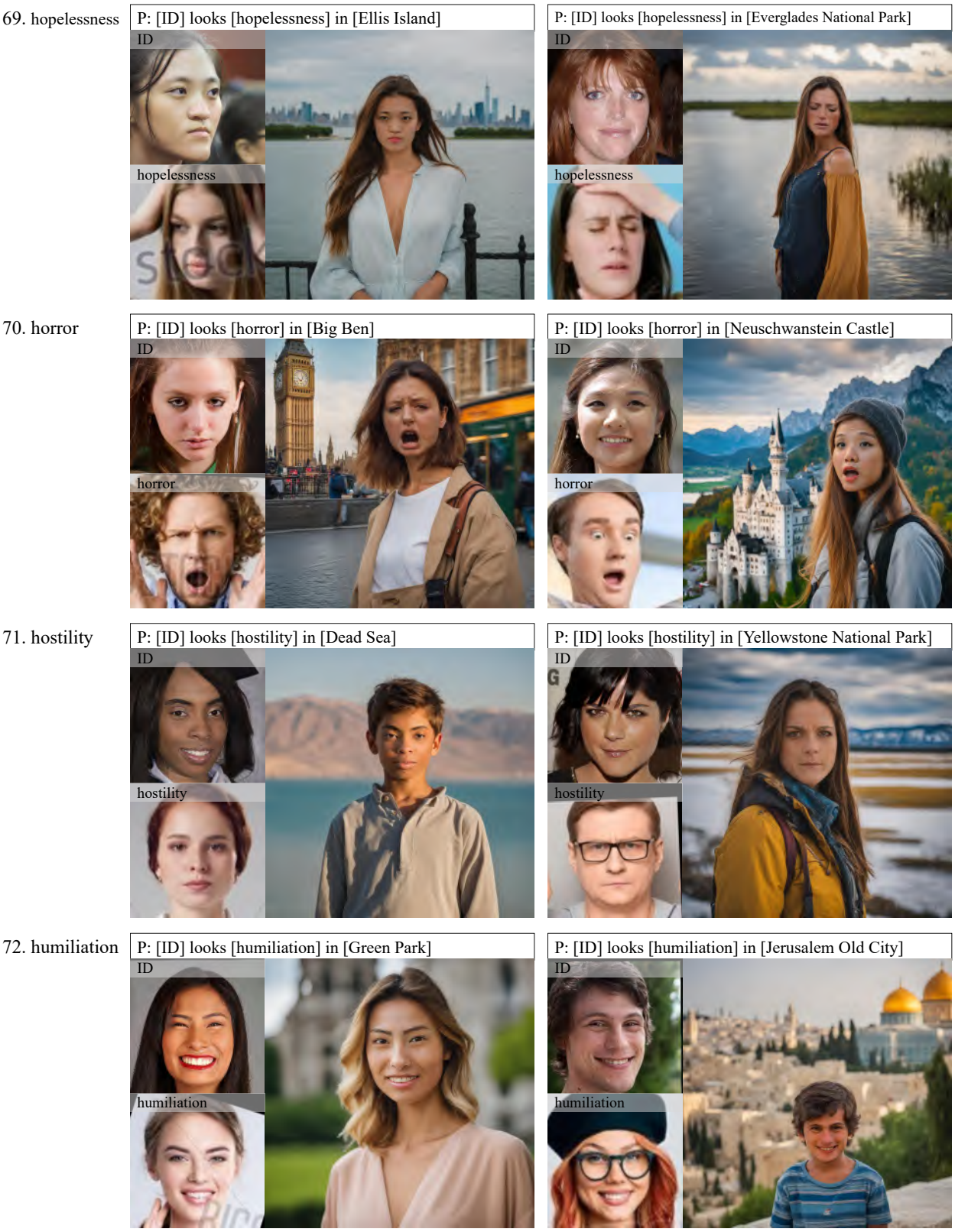
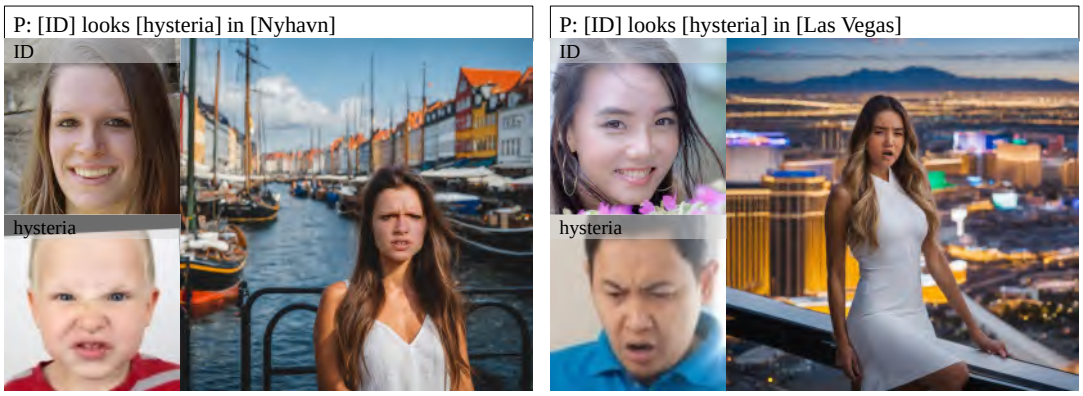


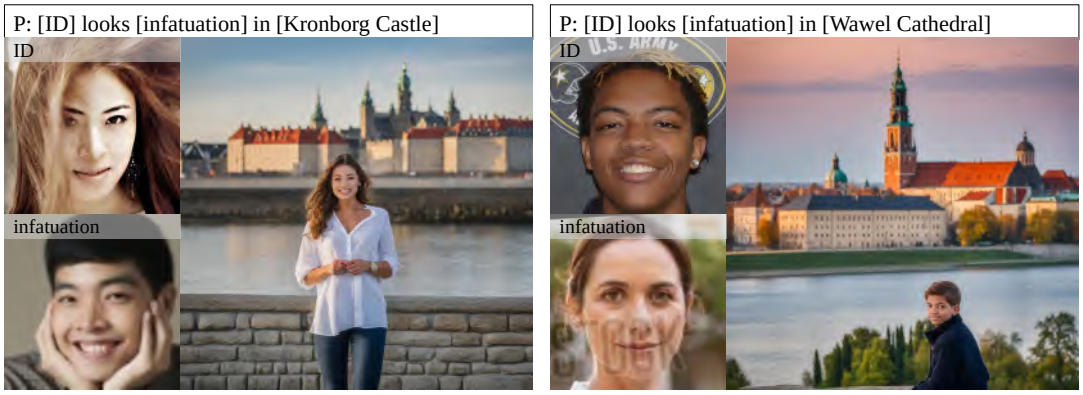
Figure 23. Continues from Figures 6-22. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

73. hurt

74. hysteria



75. infatuation



76. insecurity

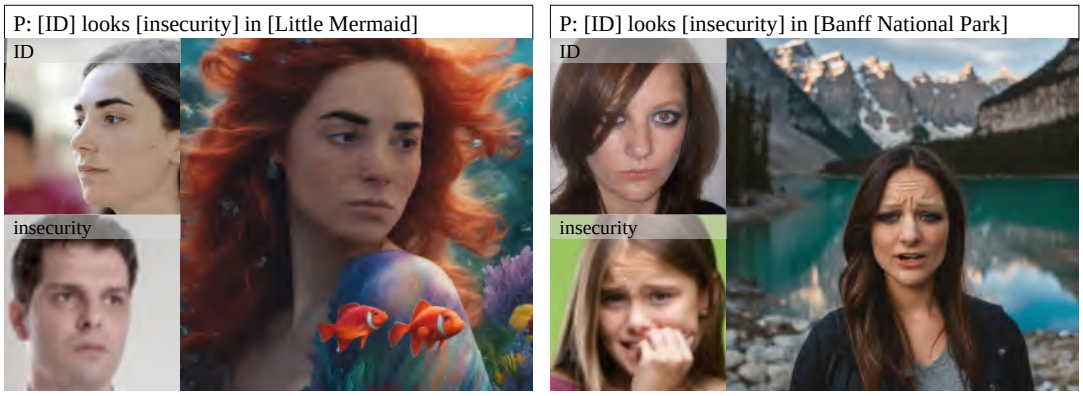
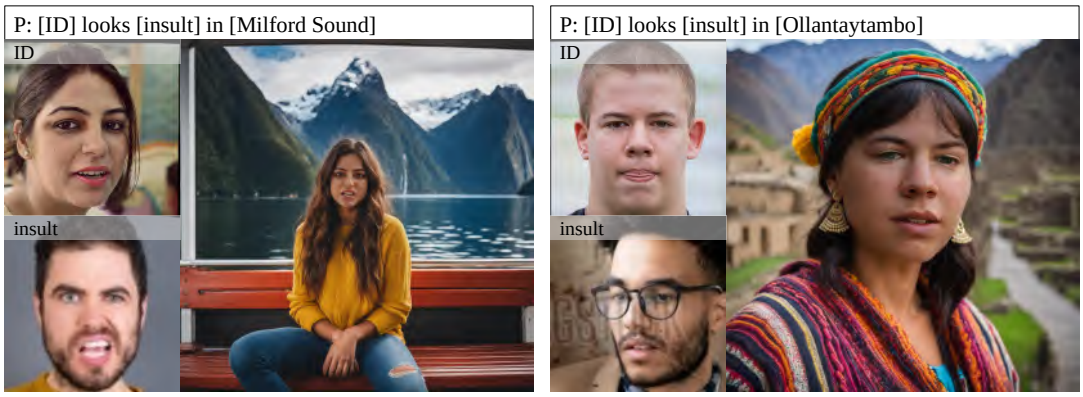
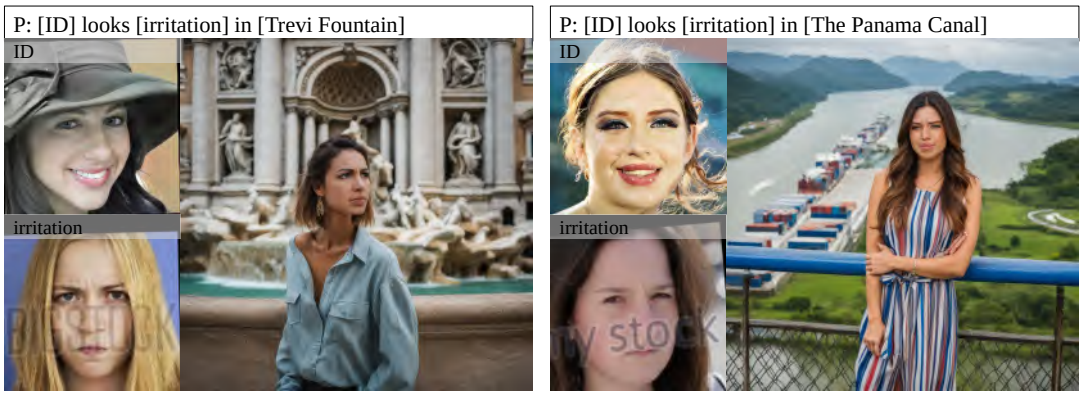


Figure 24. Continues from Figures 6-23. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

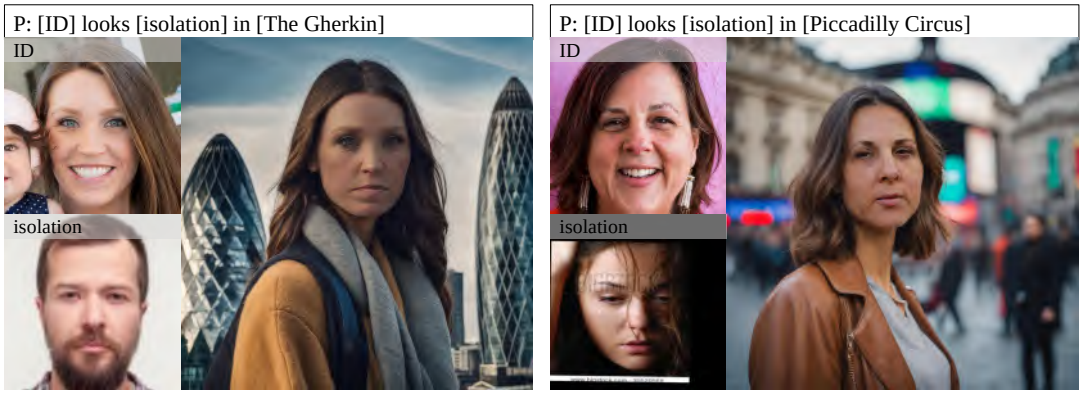
77. insult



78. irritation



79. isolation



80. jealousy

Figure 25. Continues from Figures 6-24. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

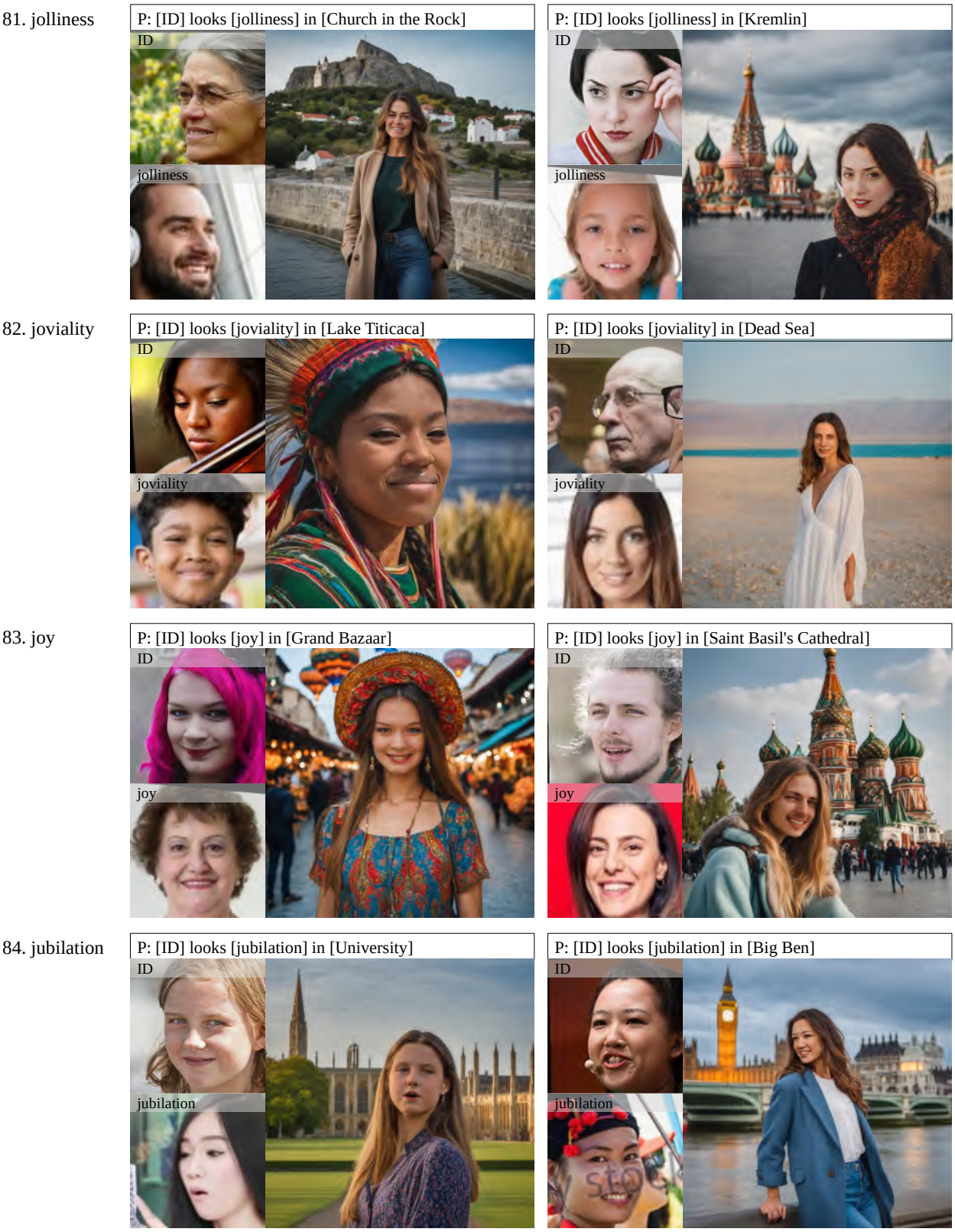
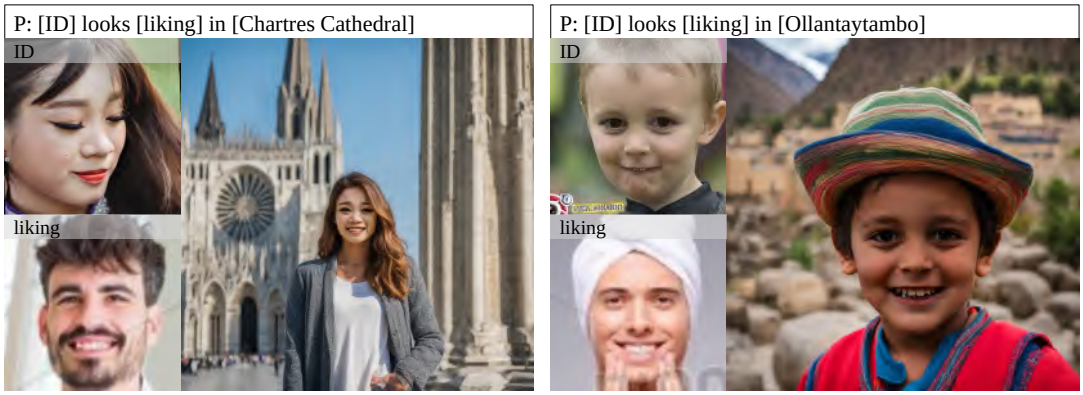
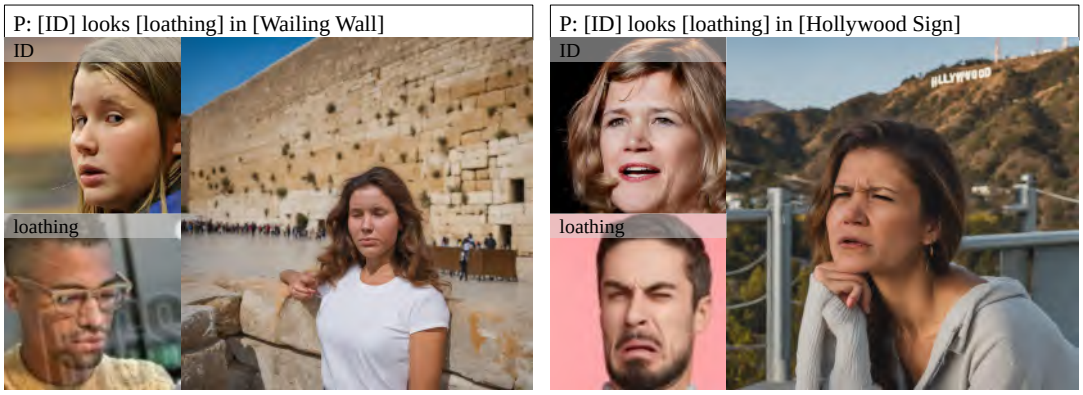


Figure 26. Continues from Figures 6-25. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

85. liking



86. loathing



87. loneliness

88. longing

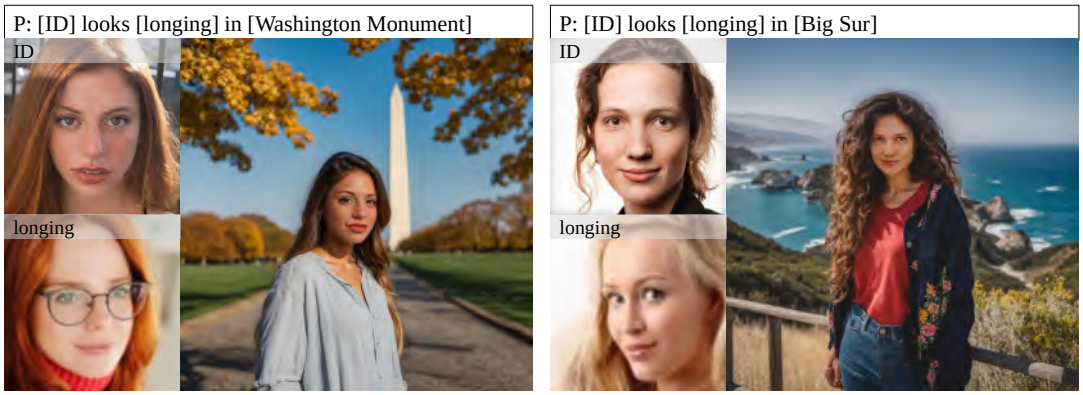
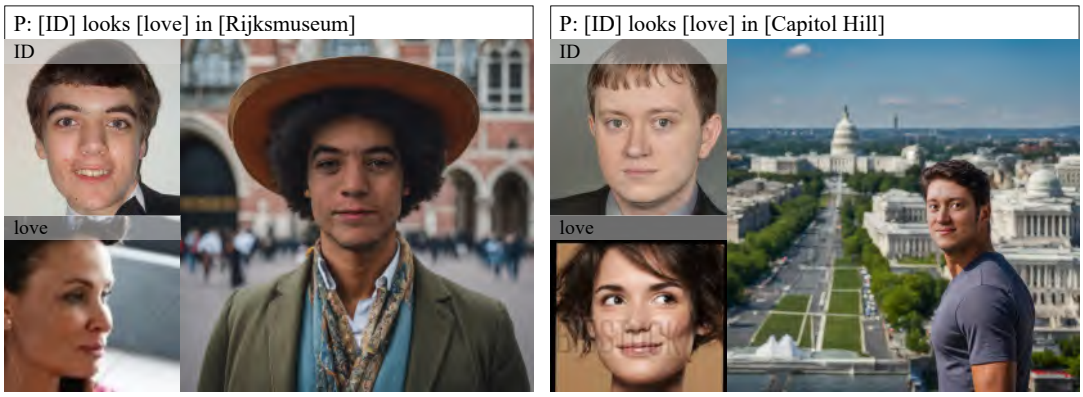


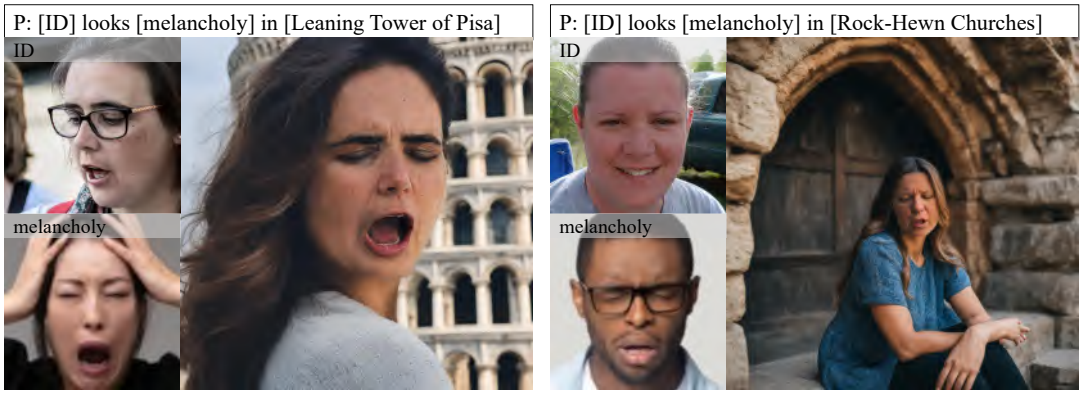
Figure 27. Continues from Figures 6-26. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

89. love



90. lust

91. melancholy



92. misery

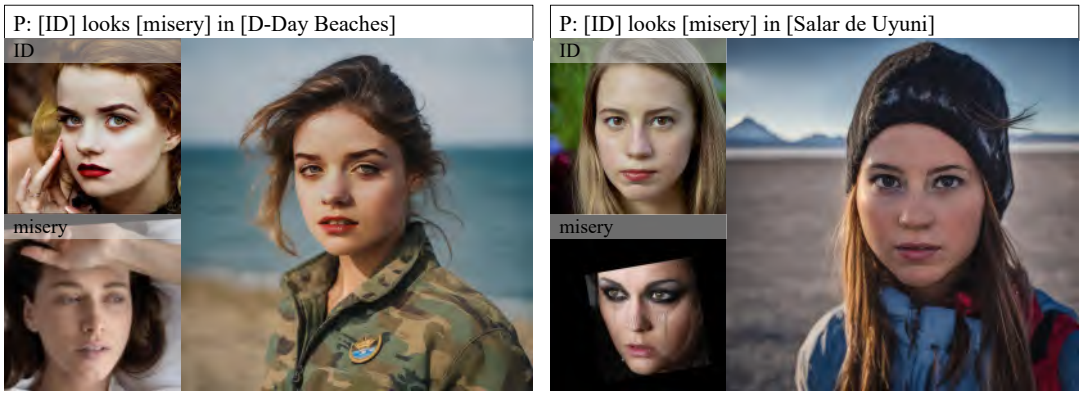


Figure 28. Continues from Figures 6-27. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

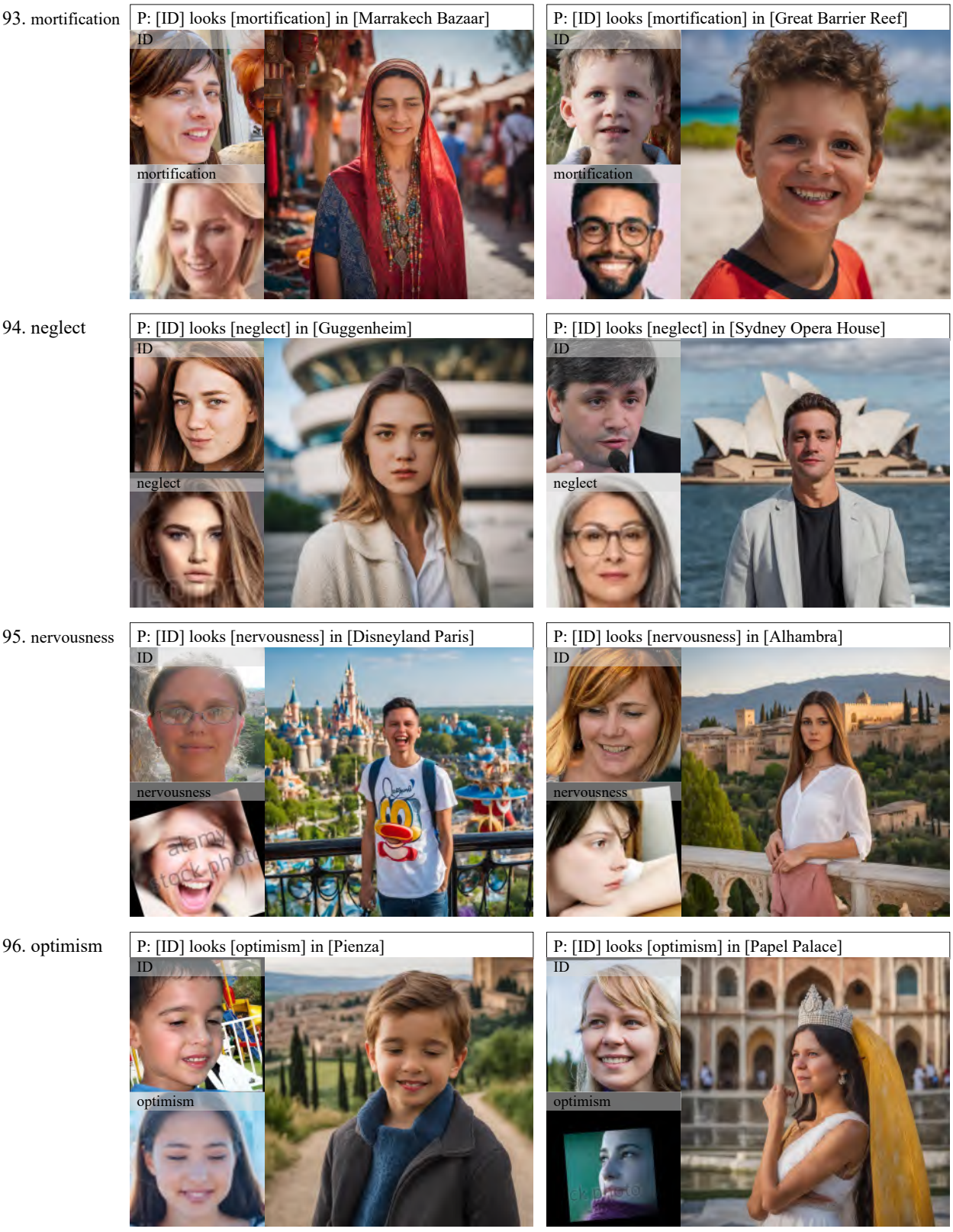


Figure 29. Continues from Figures 6-28. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

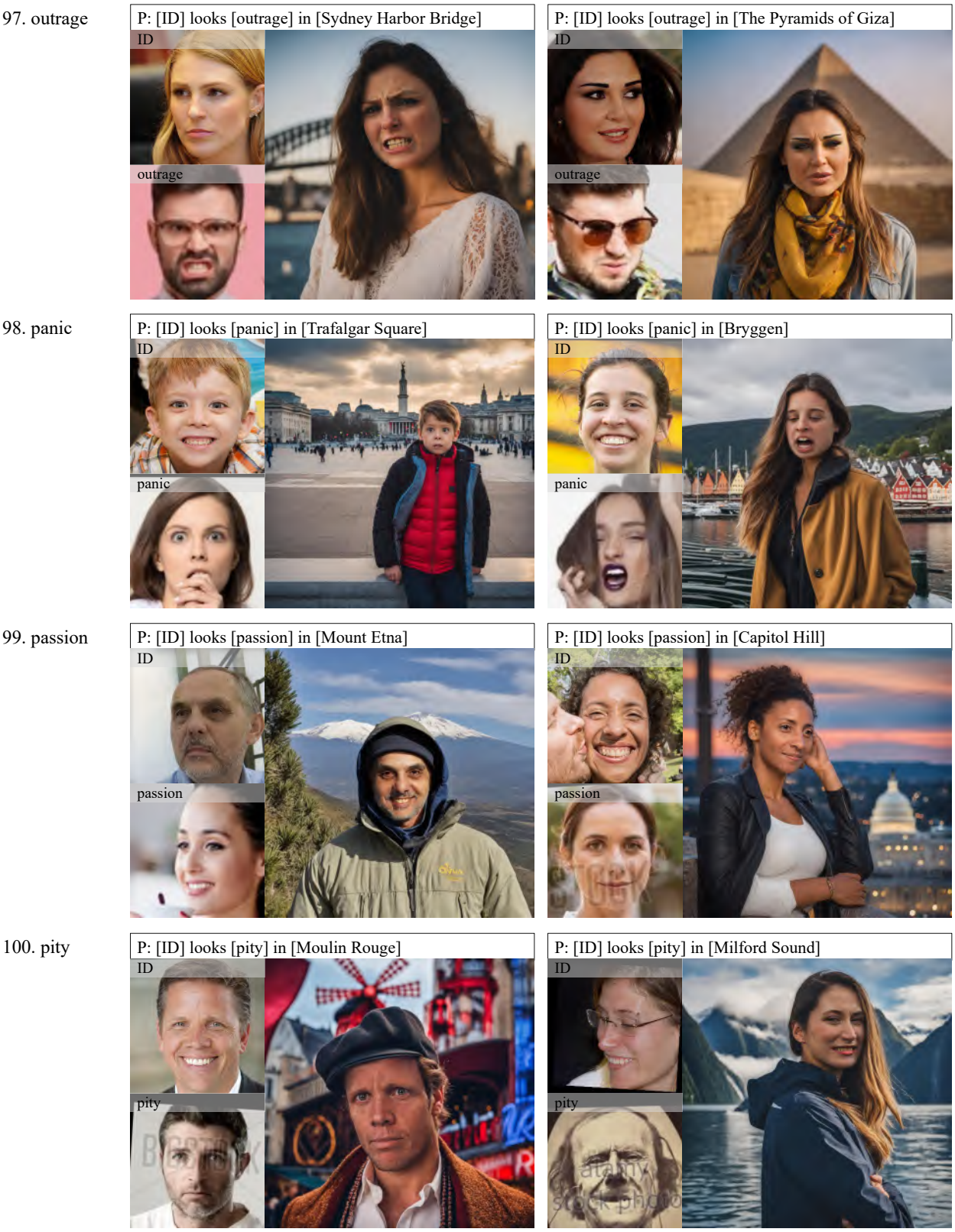


Figure 30. Continues from Figures 6-16. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

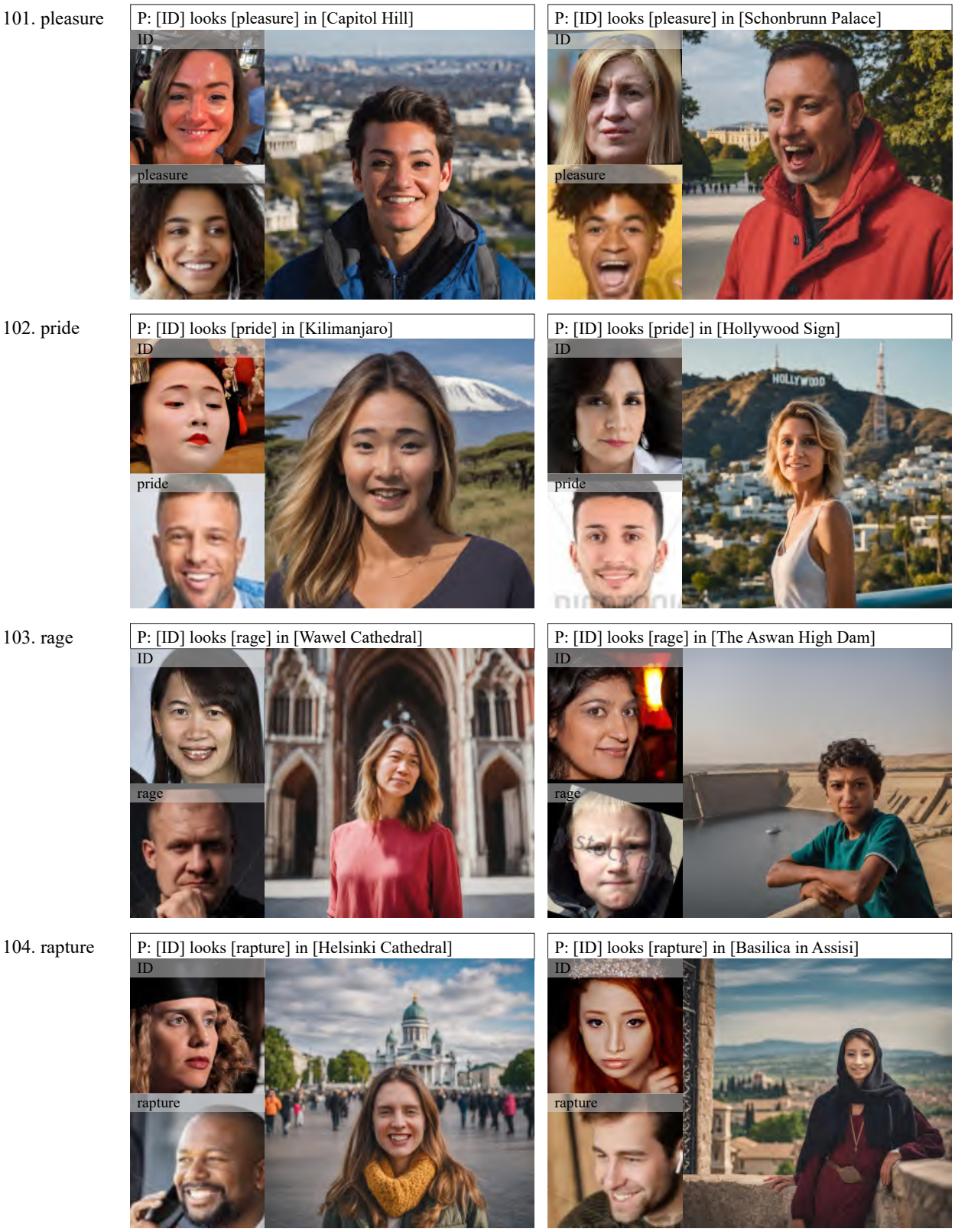
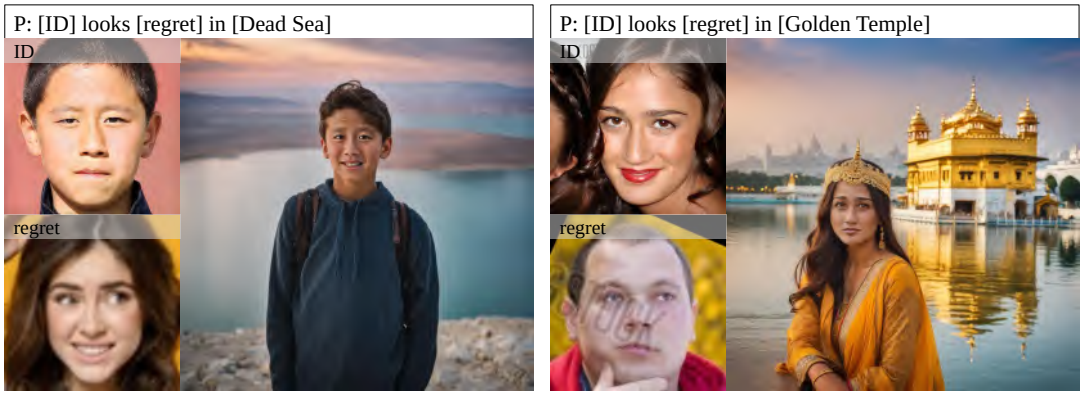
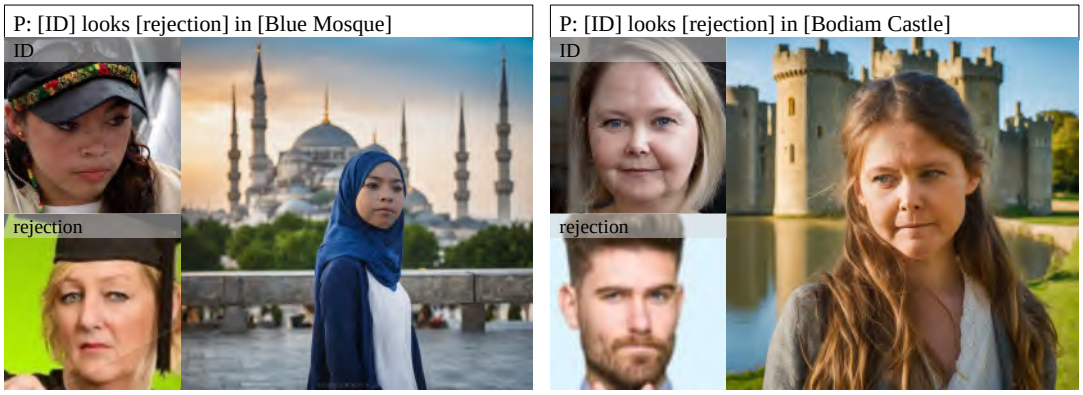


Figure 31. Continues from Figures 6-30. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

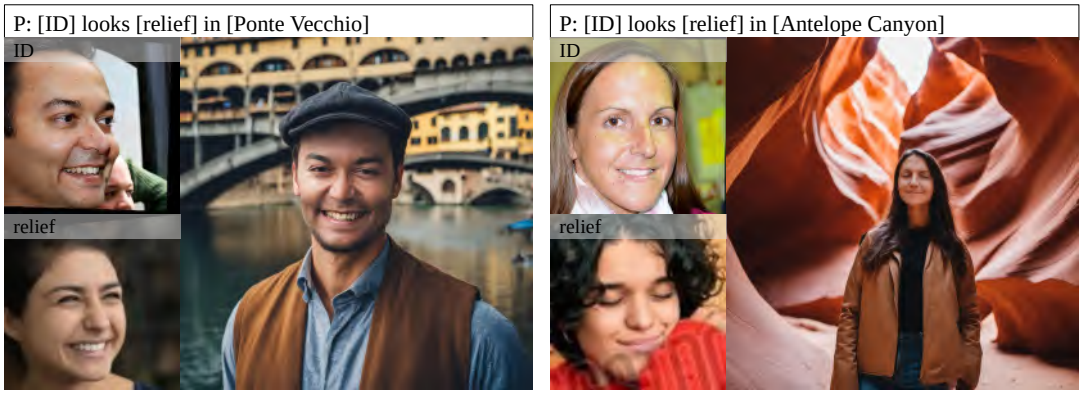
105. regret



106. rejection



107. relief



108. remorse

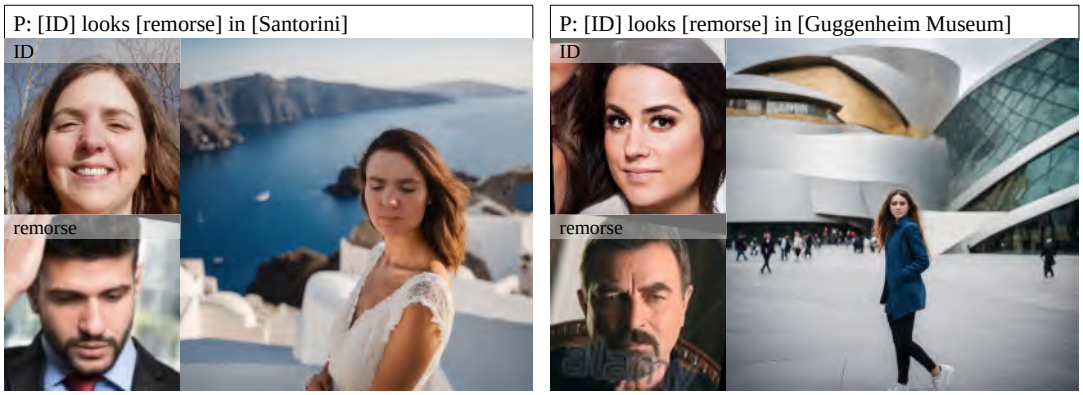


Figure 32. Continues from Figures 6-31. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

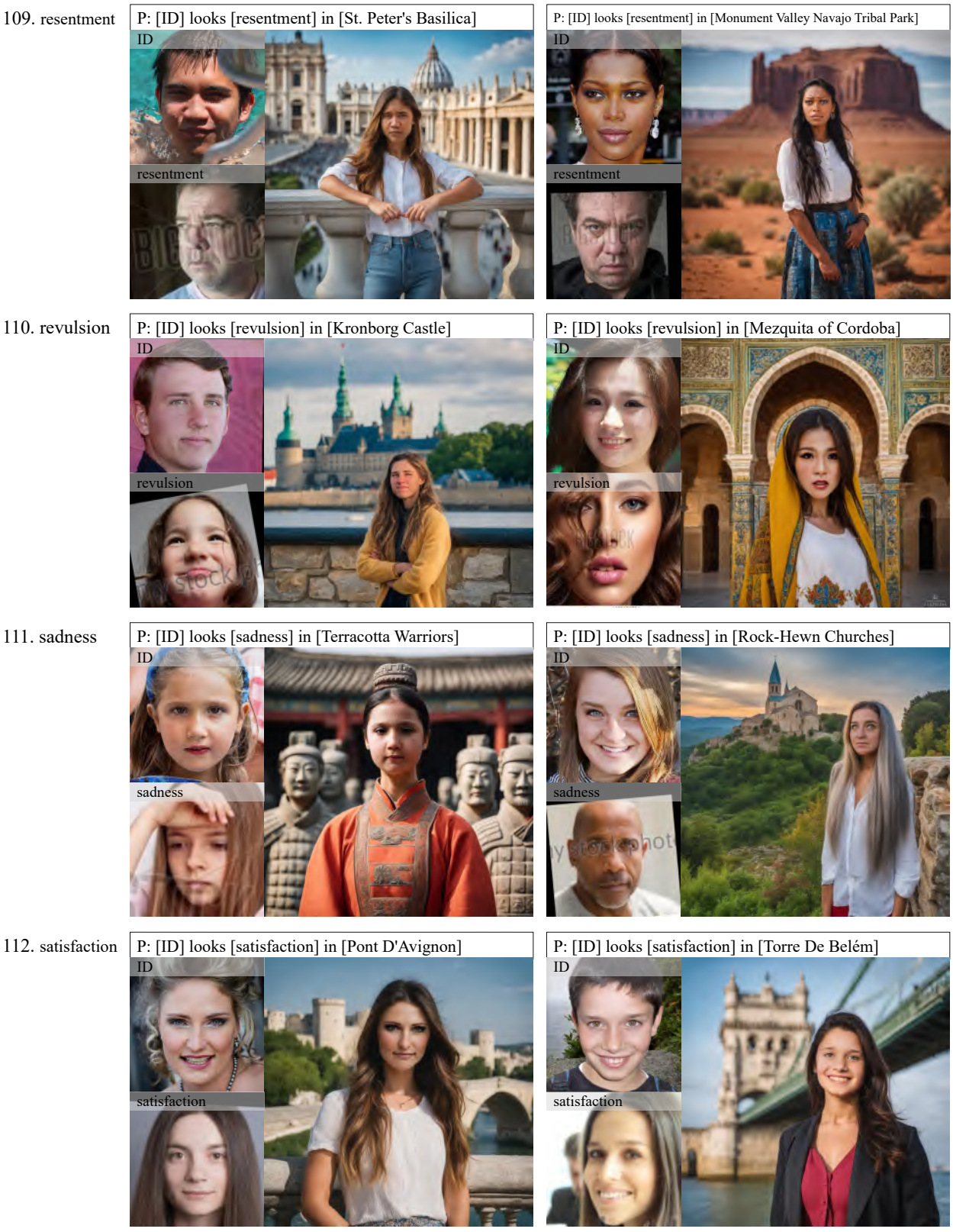


Figure 33. Continues from Figures 6-32. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

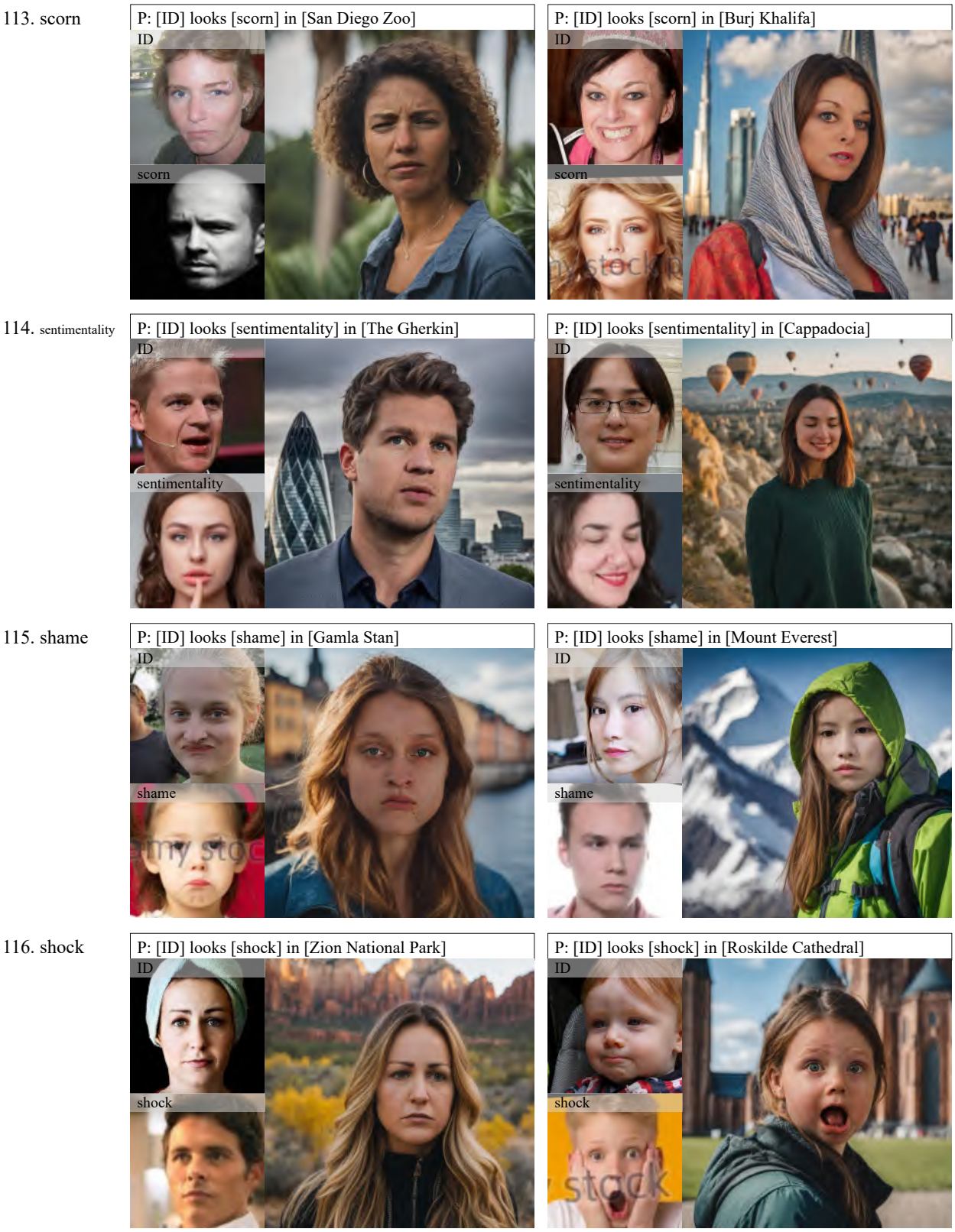
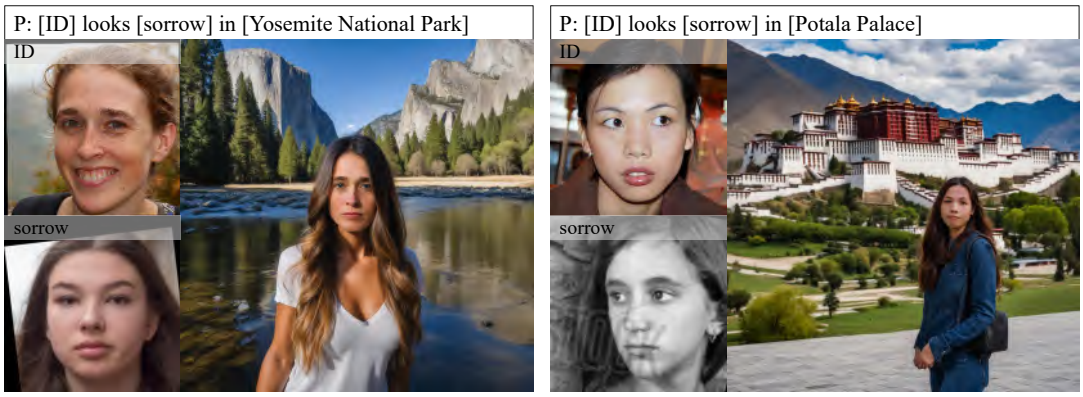
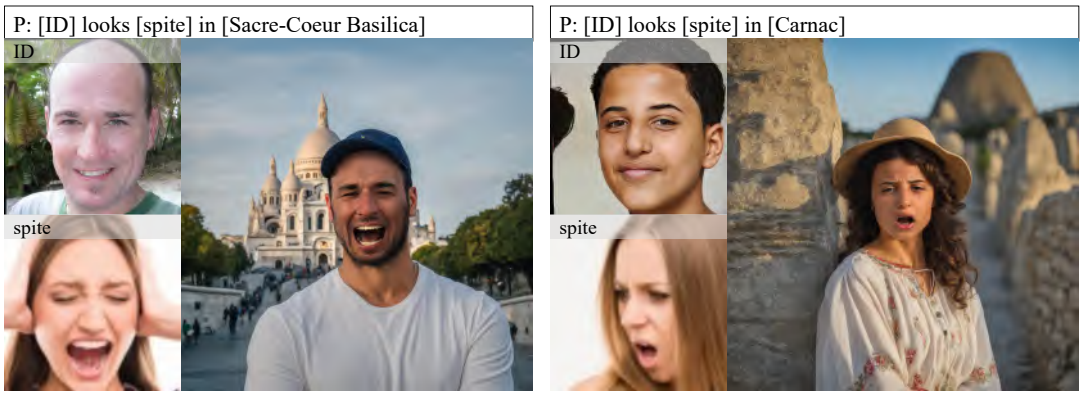


Figure 34. Continues from Figures 6-33. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

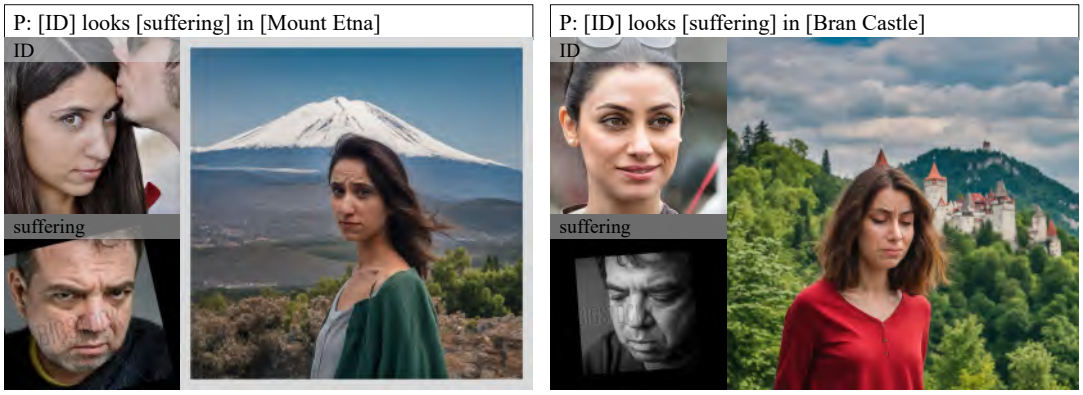
117. sorrow



118. spite



119. suffering



120. surprise

Figure 35. Continues from Figures 6-34. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

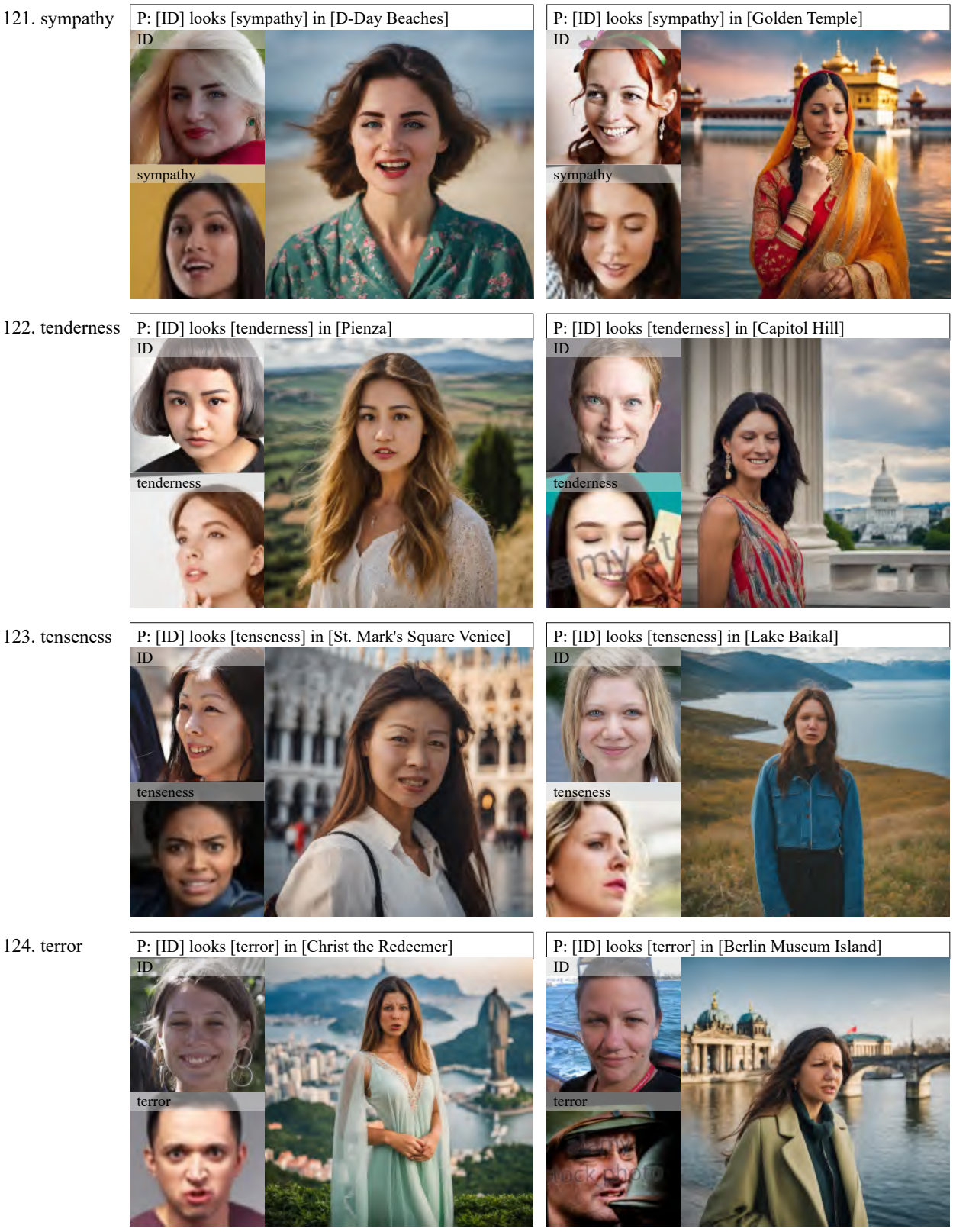


Figure 36. Continues from Figures 6-35. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

Figure 37. Continues from Figures 6-36. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

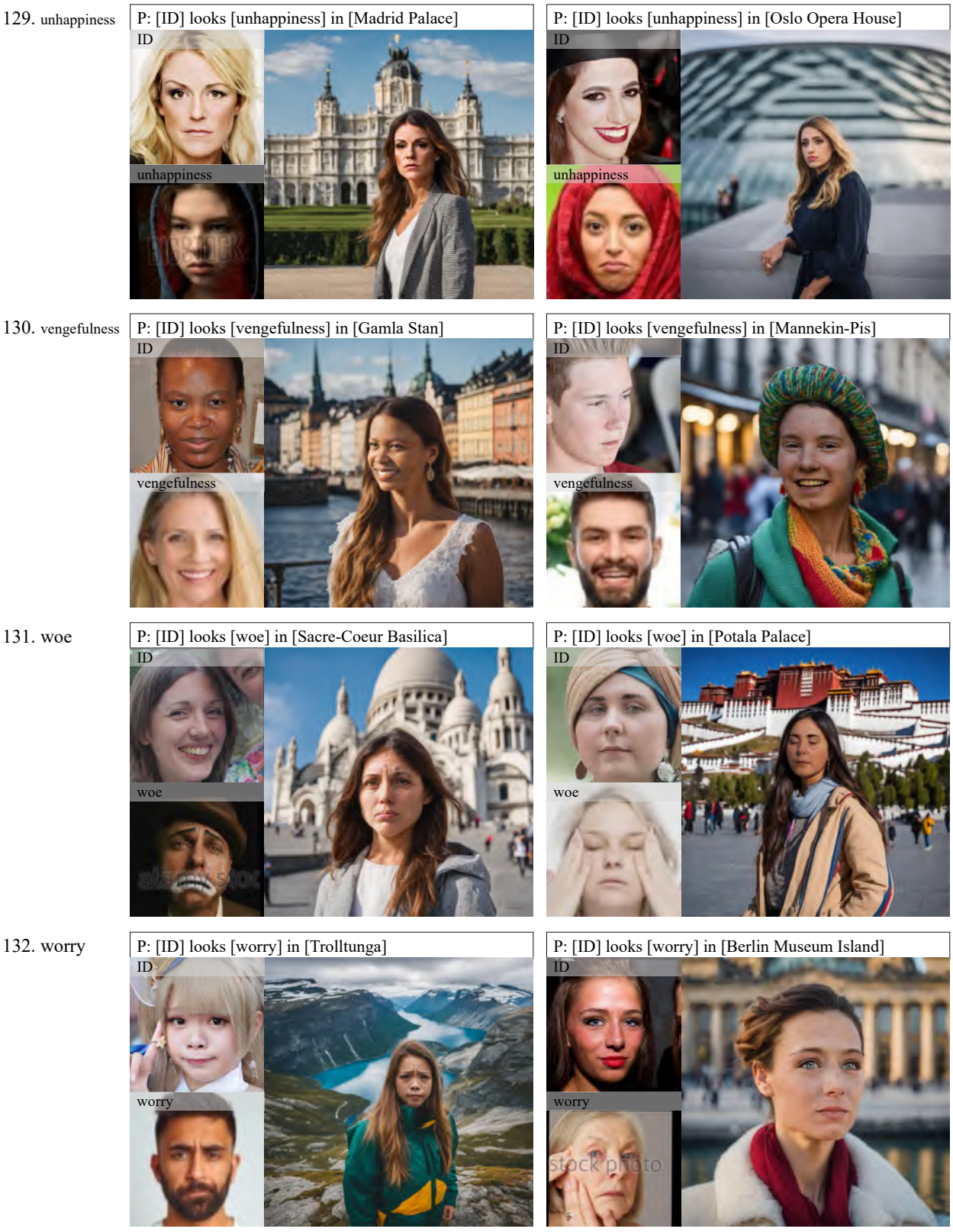
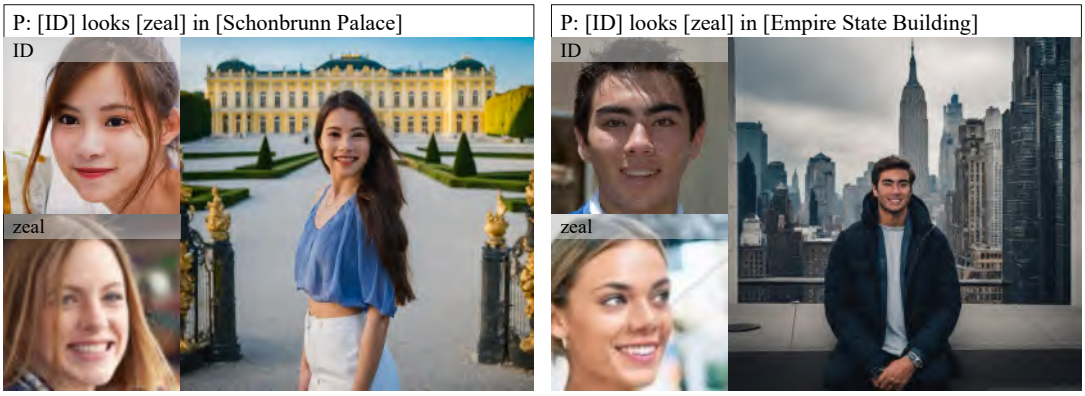


Figure 38. Continues from Figures 6-37. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.

133. wrath

134. zeal



135. zest

Figure 39. Continues from Figures 6-38. The input text prompt is shown at the top. The image in the top right corner refers to the ID image and the image in the bottom right corner refers to the expression reference image. The image on the right showcases the resulting image according to the inputs of the text prompt and ID image. Please zoom in for more details.