Learning Multi-dimensional Human Preference for Text-to-Image Generation

Supplementary Material

1. Text-to-Image Ranking

The score functions of text-to-image models serve to rank generated images by selecting generated images with higher scores to enhance the quality of text-to-image models. We extract some prompts from the validation set and generate 50 images for each prompt using the text-to-image models detailed in Tab. 2 of the main text. We compared the statistical method CLIP score [2], the PicScore [1] which learns human preferences, and our MPS method. Since PicScore only learns overall preferences for each image, we only employ our MPS conditioned on the overall score a fair comparison. We display the results of how these three models rank images generated from the same prompts. Fig. 1 shows the selected images with the highest scores according to each model. These qualitative comparisons indicate that our MPS is capable of selecting more preferable images than the baseline methods.

Figure 1. Comparing images selected by CLIP Score [2], PicScore [1], and our MPS.

2. Prompt category

We annotate the categories of the collected prompts based on the categories of Parti [3]. We merge some categories of Parti [3] and obtain 7 categories. These 7 categories, along with their corresponding original Parti categories, are as follows: Characters (People), Scenes (Indoor scenes, Outdoor scenes), Objects (Vehicles, World knowledge), Animals (Animals), Plants (Produce & plants), Arts (Artifacts, Arts, Illustrations), and Food (Food & beverage).

Additionally, as Fig. 2 in the main text shows, the initially collected prompts demonstrate a long-tail distribution in categories. Therefore, for categories with fewer prompts, we generate additional prompts leveraging Large Language Models (LLMs) to maintain a balanced distribution. Below are examples of generated prompts:

For the Animals category:

- A peacock spreading its magnificent tail by the lakeside, attracting the attention of passing tourists who stop to admire.
- A group of polar bears frolicking in the snow, rolling and playing around.
- Two little rabbits chasing each other and playing on a grassy field, enjoying the warm sunshine.
- A group of ants transporting food to their nest in an organized line, showing hard work and diligence.

For the Plants category:

- Sunlight filtering through the leaves, a bud ready to bloom sways in the gentle breeze.
- In the orchard, branches are laden with tempting sweet fruits.
- In autumn, the grapevines have turned yellow, with bunches of purple-red grapes glistening in the sunlight.
- A mountainside covered with red maple leaves, like a fiery beauty of autumn.

For the Arts category:

- A little girl holding hands with a giant bear as they walk through an enchanted forest, with magical animals passing by occasionally.
- Original illustration in a cartoon style, cute and suitable for a children's storybook cover.
- An ink wash painting employing the splashed ink technique to depict the ambiance of mountains and rivers, conveying a majestic and grand atmosphere.
- An impressionist oil painting depicting the leisurely ambiance of an afternoon, with sunlight cascading over a beautiful garden.

For the Food category:

- Delicious and crispy fried chicken.
- Green broccoli with a vegetable salad.
- Spicy and appetizing Korean kimchi.
- Sweet and sour, delicious strawberry milkshake.

3. More Visualizations

Fig. 2 shows more visualization results. We use Grad-CAM to visualize the words in the prompts and the regions in the images that MPS focuses on when scoring the generated images. With the help of the condition mask, when the Aesthetic condition is given, MPS tends to focus on words like 'handsome', 'high construct', and 'beautiful'. When alignment is the condition, MPS tends to focus on attributes (such as 'photo', '3D rendering'), quantities ('35mm lens'), and locations ('chemistry laboratory'). Meanwhile, under

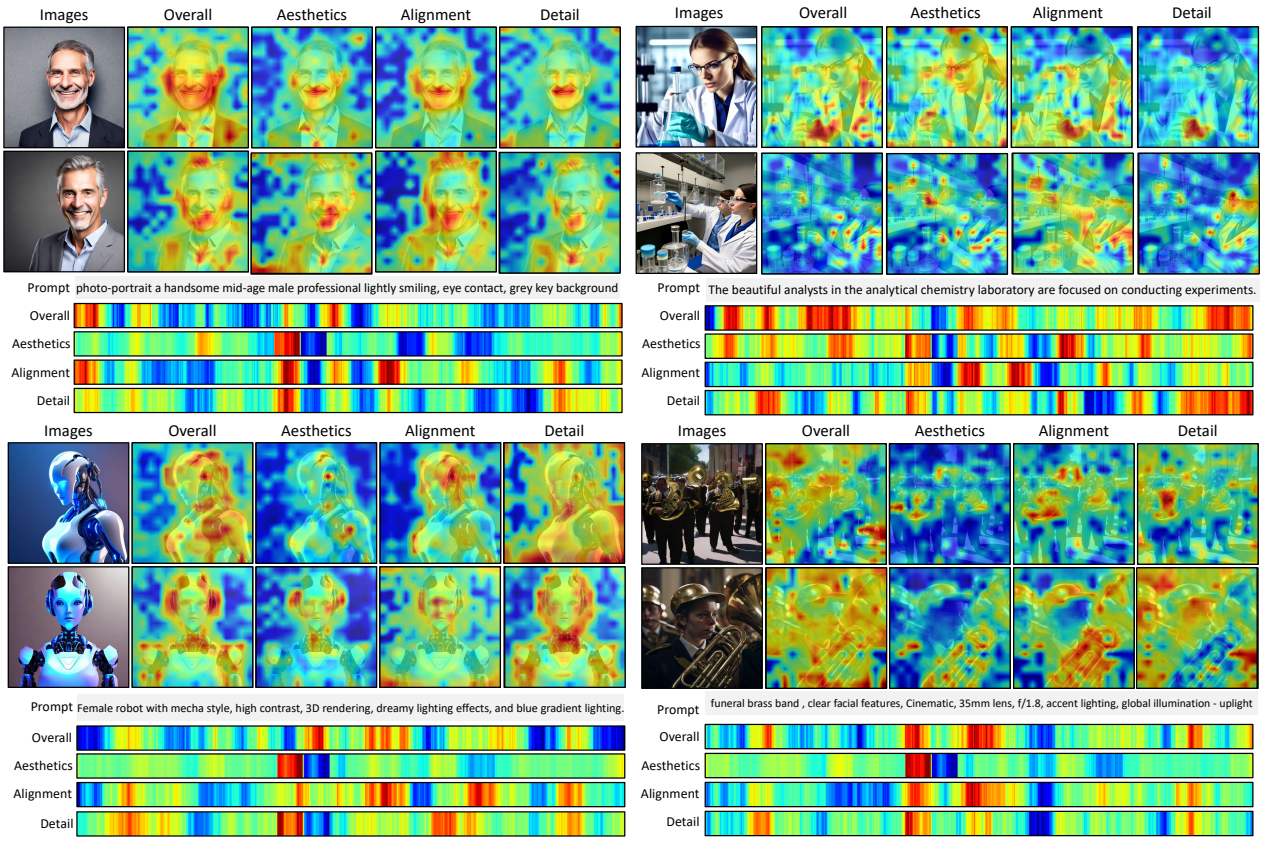

Figure 2. More visualizations. Areas with a redder color indicate regions that MPS focuses on more when scoring.

the detail condition, MPS tends to focus on specific regions of the image (such as 'face', 'hair', 'hands' and 'limbs'). The visual results demonstrate that MPS can focus on different regions of the image and prompt according to different conditions, thereby predicting varying human preferences under different conditions.

References

[1] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[3] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. 1