Appeal and quality assessment for AI-generated images

Steve Göring, Rakesh Rao Ramachandra Rao, Rasmus Merten, Alexander Raake

Audiovisual Technology Group; Technische Universität Ilmenau, Germany
Email: [steve.goering, rakesh-rao.ramachandra-rao, rasmus-leo-lukas.merten, alexander.raake]@tu-ilmenau.de

Code & Data: `http://git.avt-imt.de/avt_ai_images`

Introduction

- increase of AI-generated images, e.g.:
  - DALL-E-2, Midjourney, Stable Diffusion [7], or Craiyon [1]
- text prompt → generated image (=text to image (T2I))
- example images, see Fig. 1
  - text prompt “Hyper-realistic photo of an abandoned industrial site during a storm” (p16)
- uncommon artificial-looking distortions, varying appeal visual quality
- published AVT-AI-Image-Dataset [3]:
  - appeal, realism, text prompt matching
  - 5 T2I generators
- related work: usually no comparison of several generators
- open: **image quality and appeal**

Overview of the AVT-AI-Image-Dataset

- AVT-AI-Image-Dataset: 27 text prompts, 16 from Drawbench [8]
- 11 real images included (p17 to p27); all images: resolution 512x512
- 146 images, full overview in [3], prompt selection see:

ID	Prompt	Origin
p11	A mechanical or electrical device for measuring time	Drawbench
p16	Hyper-realistic photo of an abandoned industrial site during a storm	Drawbench
p20	Purple flowers with yellow and a small bug	own
p23	A portrait of a mule	own
p27	A box with tools for home office	own

Subjective Test Design and Evaluation

- similar to [4, 6, 3]; AVRate Voyager [2] with two 1-5 sliders
- 25 participants (12 from clickworker.com, remaining from university)
- no training phase, ≈ 30 mins; partial runs excluded in results

Evaluation of Image Appeal

- SOS-analysis [5]; a value ≈ 0.33
- cross-test comparison: Pearson ≈ 0.91, Kendall ≈ 0.75, Spearman ≈ 0.9
- highest: Midjourney p16; lowest: Glide p11; compare Fig. 3

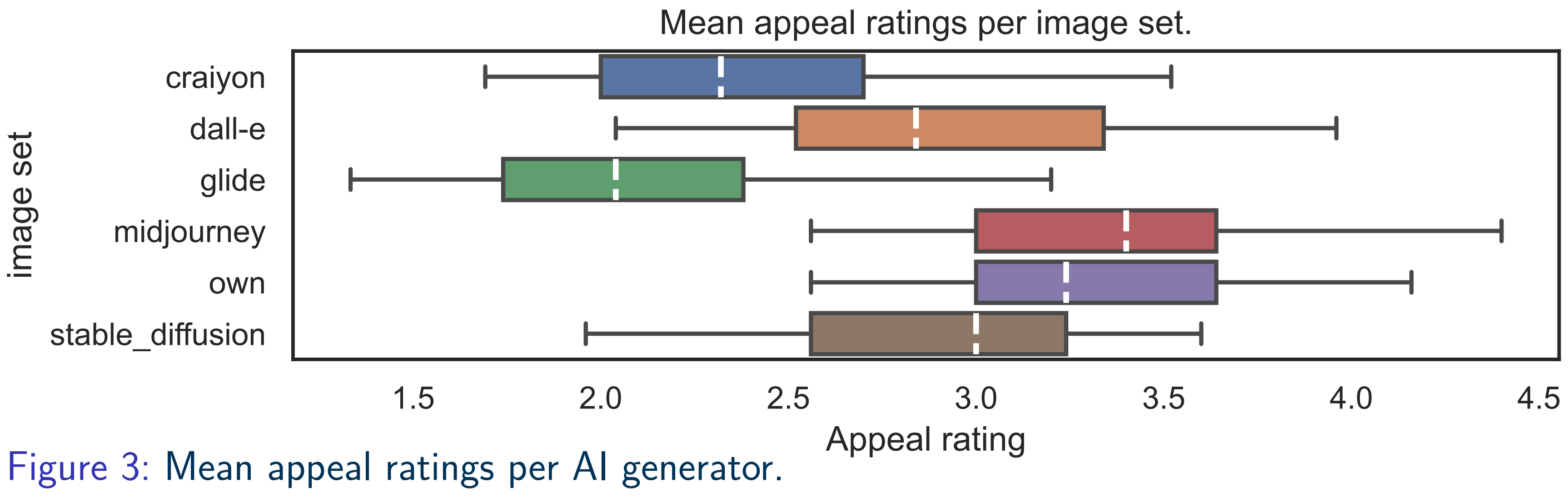

Figure 3: Mean appeal ratings per AI generator.

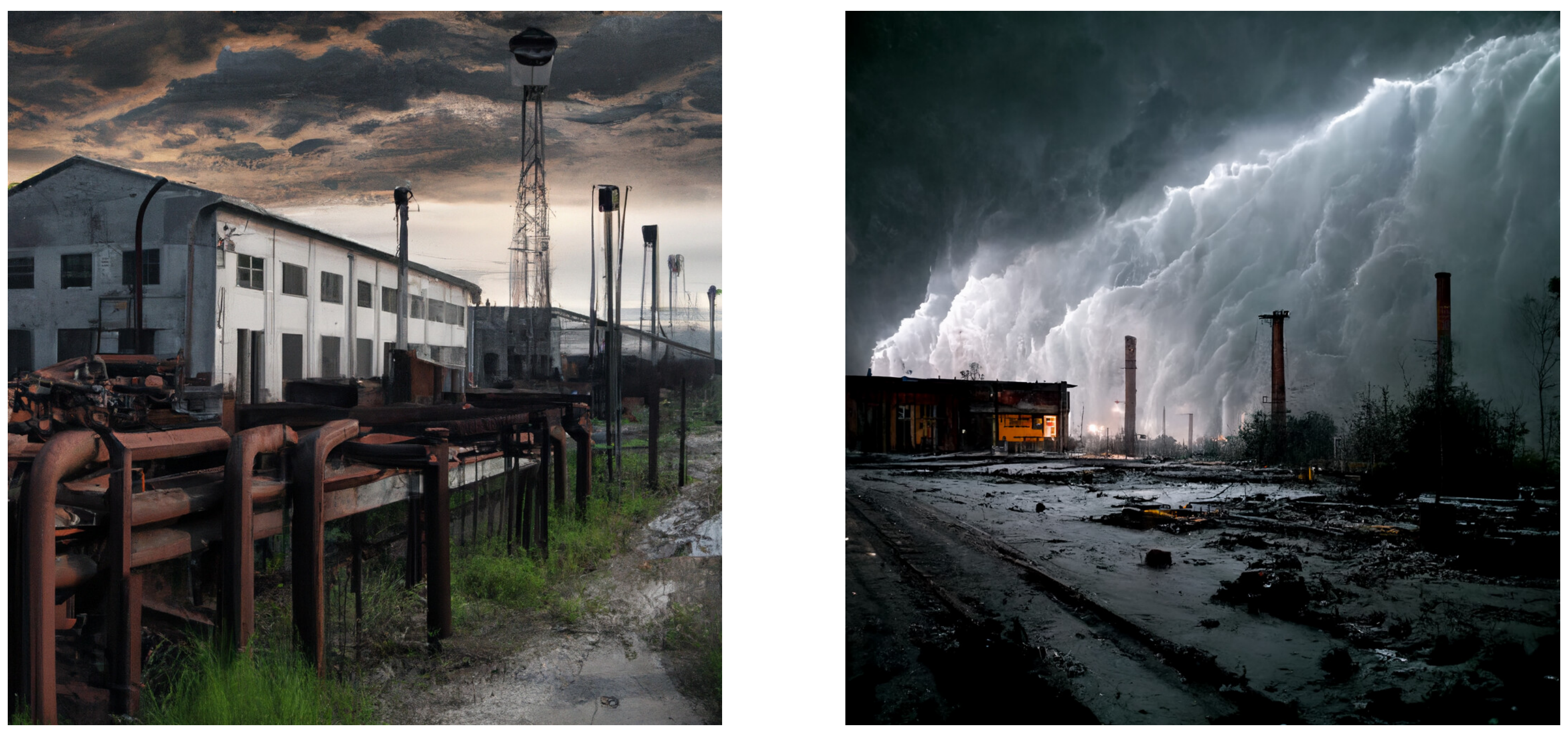

Figure 1: Generated images for p16: DALL-E-2 (left), Midjourney (right).

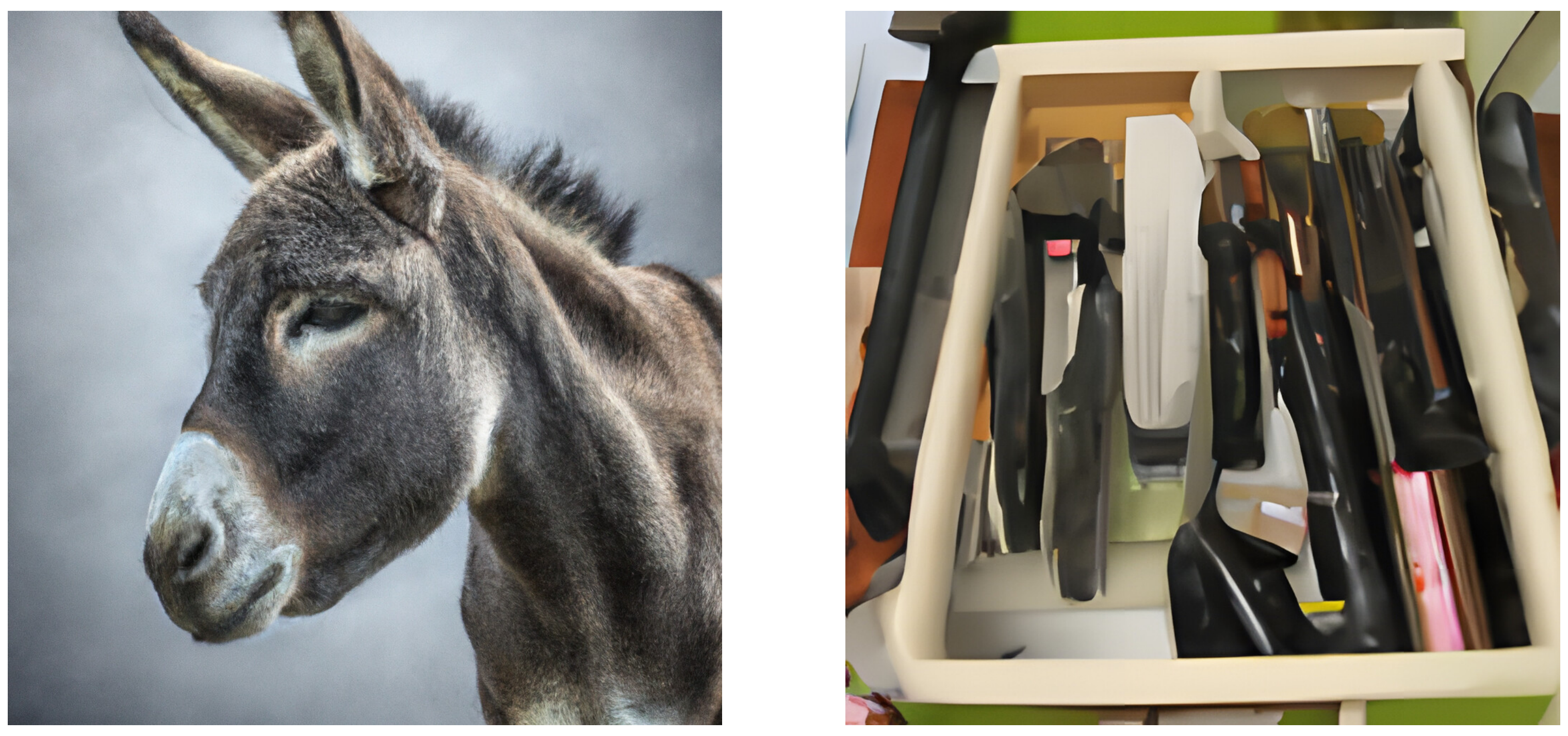

Figure 2: Best quality (left): DALL-E-2, p23 and worst (right): Glide, p27 .

Evaluation of Image Quality

- SOS-analysis [5]; a value ≈ 0.306
- Midjourney, DALL-E-2 best, see Fig. 4
- best: “own” p20 , DALL-E-2 p23; worst: Glide p27; see Fig. 2
- image quality models: best: MANIQA (0.44 PCC), BRISQUE (−0.39 PCC)
- appeal vs. quality:
  - overall: 0.80 PCC, higher appeal ↔ higher quality
  - glide: 0.57 PCC; “own”: 0.58 PCC
  - stable_diffusion: 0.62 PCC; dall-e: 0.63 PCC
  - midjourney: 0.74 PCC; craiyon: 0.77 PCC

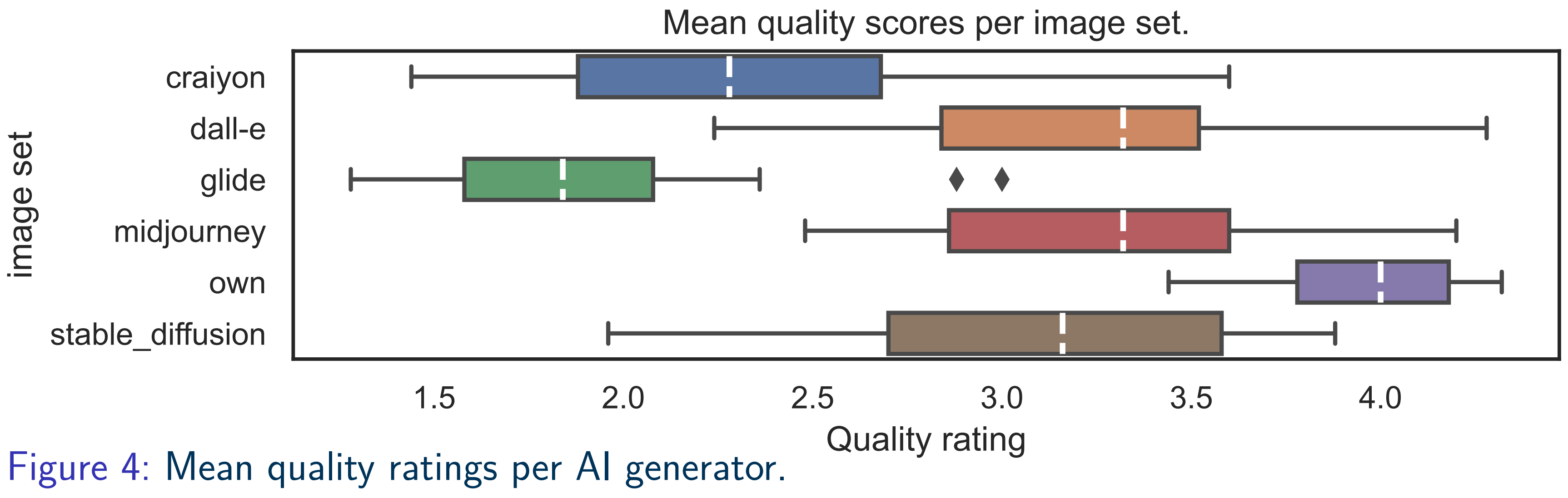

Figure 4: Mean quality ratings per AI generator.

Conclusion

- limited subjective evaluation for AI-generated images for different generators
- evaluation: AVT-AI-Image-Dataset appeal/ quality; crowdsourcing
- Glide and Craiyon: overall low appeal and quality
- DALL-E-2 and Midjourney: similar high appeal/ quality to real photos

Future Work

- objective quality models: low performance for AI-generated images
- prediction models and features for AI-generated images
- larger datasets
- newer AI generators

References

[1] B. Dayma et al. *DALL · E Mini.* July 2021.

[2] S. Göring, R. Rao Ramachandra Rao, S. Fremerey, and A. Raake. “AVRate Voyager: an open source online testing platform”. In: *23st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2021, pp. 1–6.

[3] S. Göring, R. Rao Ramachandra Rao, R. Merten, and A. Raake. “Analysis of Appeal for realistic AI-generated Photos”. In: vol. 11. IEEE, 2023, pp. 38999–39012.

[4] S. Göring, R. Rao Ramachandra Rao, and A. Raake. “Quality Assessment of Higher Resolution Images and Videos with Remote Testing”. In: *Quality and User Experience (QUEX)* 8 (2023).

[5] T. Hoßfeld, R. Schatz, and S. Egger. “SOS: The MOS is not enough!” In: *3rd int. workshop on quality of multimedia experience*. IEEE. 2011, pp. 131–136.

[6] R. Rao Ramachandra Rao, S. Göring, and A. Raake. “Towards High Resolution Video Quality Assessment in the Crowd”. In: *13th Int. Conf. on Quality of Multimedia Experience (QoMEX)*. 2021.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

[8] C. Saharia et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *arXiv preprint arXiv:2205.11487* (2022).

Acknowledgment

The authors would like to thank the participants for taking part in this crowd test. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – DFG-437543412. Furthermore, we want to thank the “AG Wissenschaftliches Rechnen” of the TU Ilmenau for computing resources.

TECHNISCHE UNIVERSITÄT ILMENAU

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation