Diffusion Image Analogies (Supplementary Material)

Adéla Šubrtová
CTU in Prague, FEE
Prague, Czech Republic
subrtade@fel.cvut.cz

Michal Lukáč
Adobe Research
San Jose, California, United States
lukac@adobe.com

Jan Čech
CTU in Prague, FEE
Prague, Czech Republic
cechj@fel.cvut.cz

David Futschik
CTU in Prague, FEE
Prague, Czech Republic
futscdav@fel.cvut.cz

Eli Shechtman
Adobe Research
Seattle, Washington, United States
elishe@adobe.com

Daniel Sýkora
CTU in Prague, FEE
Prague, Czech Republic
sykorad@fel.cvut.cz

ACM Reference Format:
Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. 2023. Diffusion Image Analogies (Supplementary Material). In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 6–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3588432.3591558

In this supplementary material, we present additional results showing examples of diffusion image analogies with gradually increasing analogy strength λ (see Figures 1–3 and also our supplementary video). This parameter is part of the user interface, enabling the user to find a proper balance between preserving the original content and pronouncing the notion of an edit derived from the given analogy. To extend the comparison of our approach with the synthetic baseline solution described in Section 4.1 of the main paper (Comparison), we also present additional examples in Figures 4–6. Those results demonstrate the shortcomings of using Stable Diffusion in the `img2img` mode [Rombach et al. 2022] together with BLIP [Li et al. 2022] to estimate text prompts. The first issue resides in the inability of BLIP to predict the caption with enough detail to capture the essence of the given analogy adequately (see Figure 5). The second problem is that while preserving the structure of the target image B, the desired properties defined by the analogy may not transfer fully (e.g., in Fig. 6, the head shape is altered, but the black markings are missing).

REFERENCES

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. 12888–12900.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0159-7/23/08.
https://doi.org/10.1145/3588432.3591558

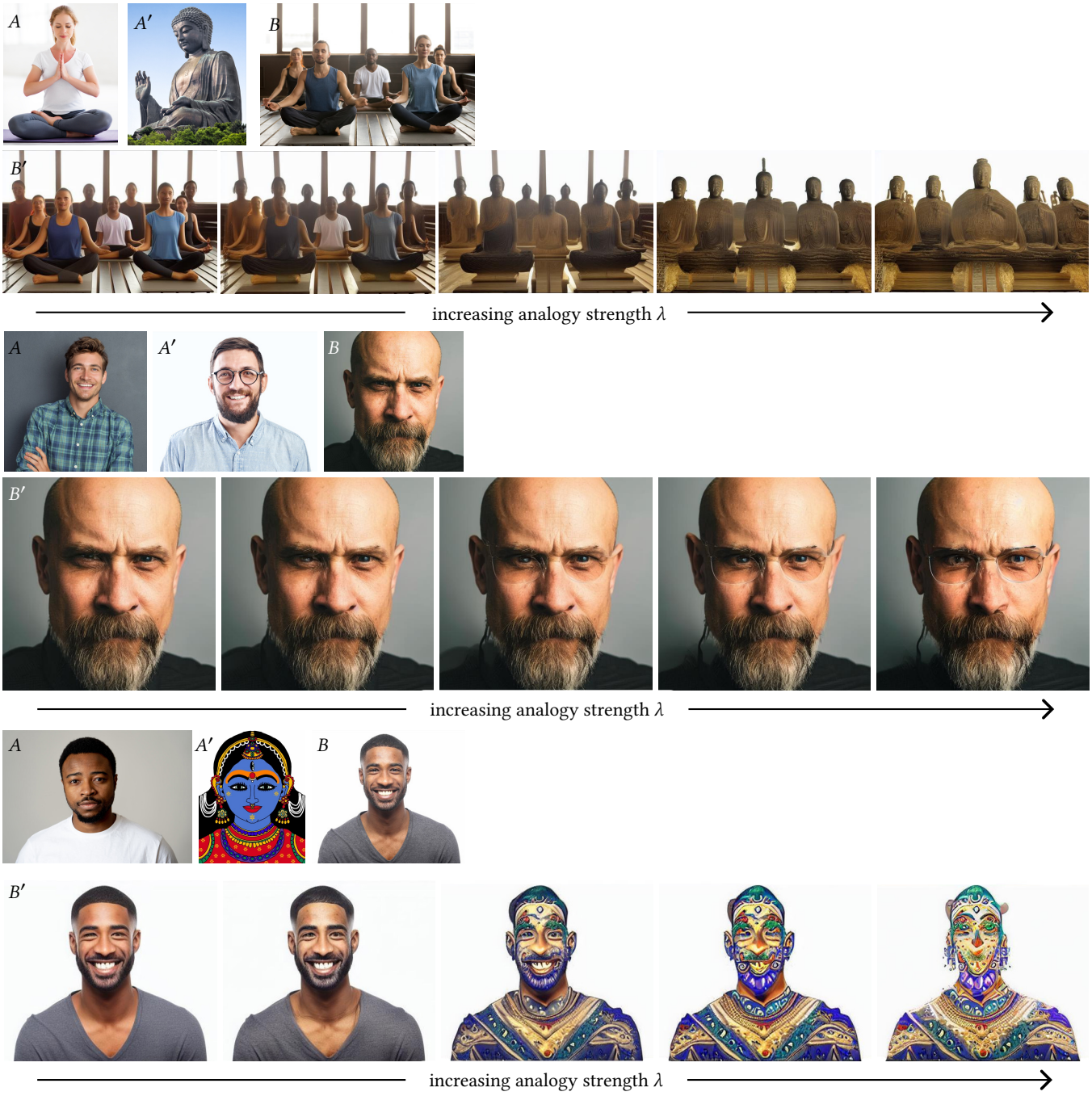

Figure 1: Examples of diffusion image analogies $A : A' :: B : B'$ produced using our approach with gradually increasing analogy strength λ. Note, how increasing λ makes the prescribed analogy more apparent. Source images: © Kevin Bidwell (Bald Man B), Adobe Stock the rest.

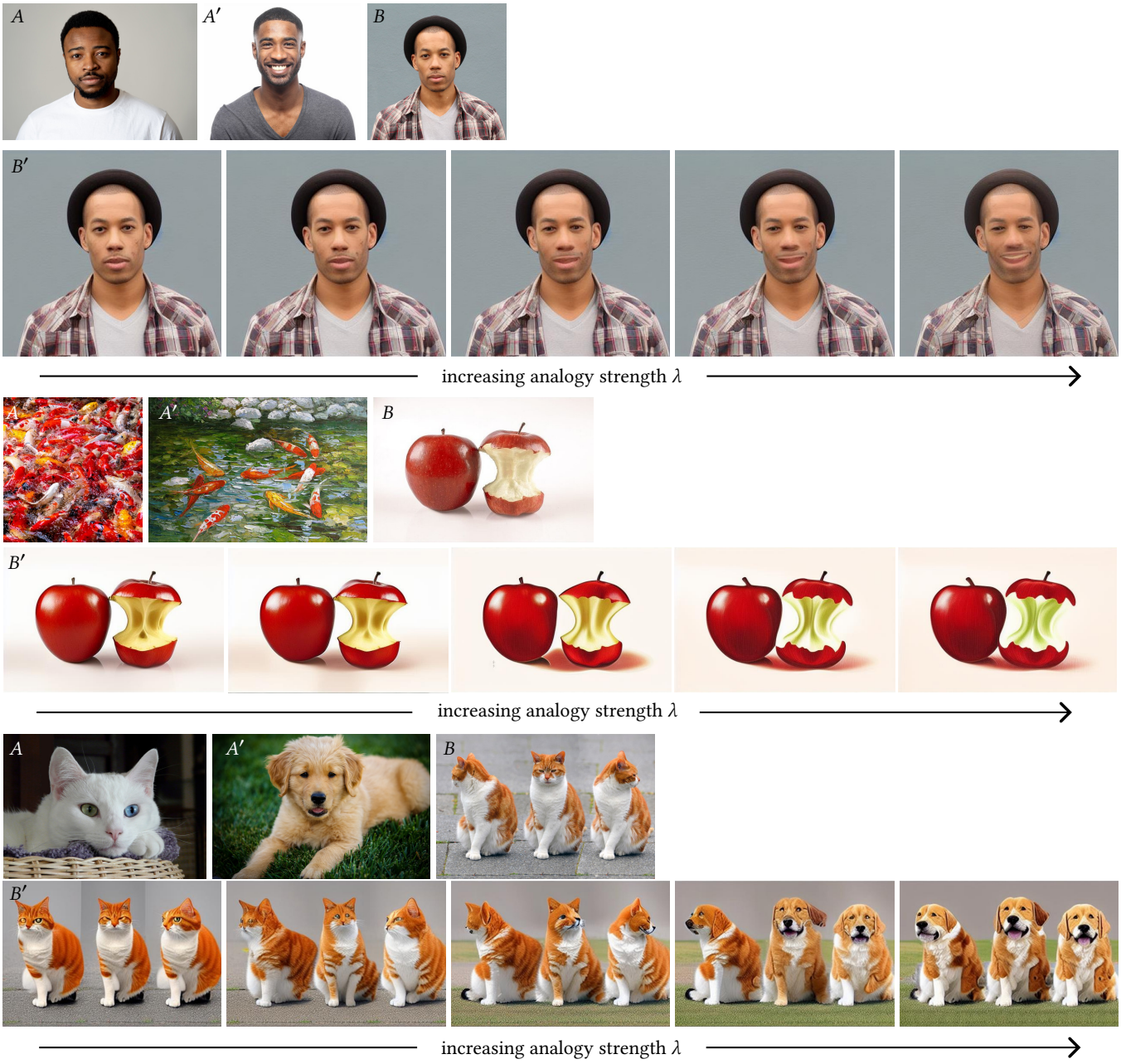

Figure 2: Examples of diffusion image analogies $A : A' :: B : B'$ produced using our approach with gradually increasing analogy strength λ (cont.). Source images: © Lucíola Correia (Apples B), © George Hodan (Three Cats B), Adobe Stock the rest.

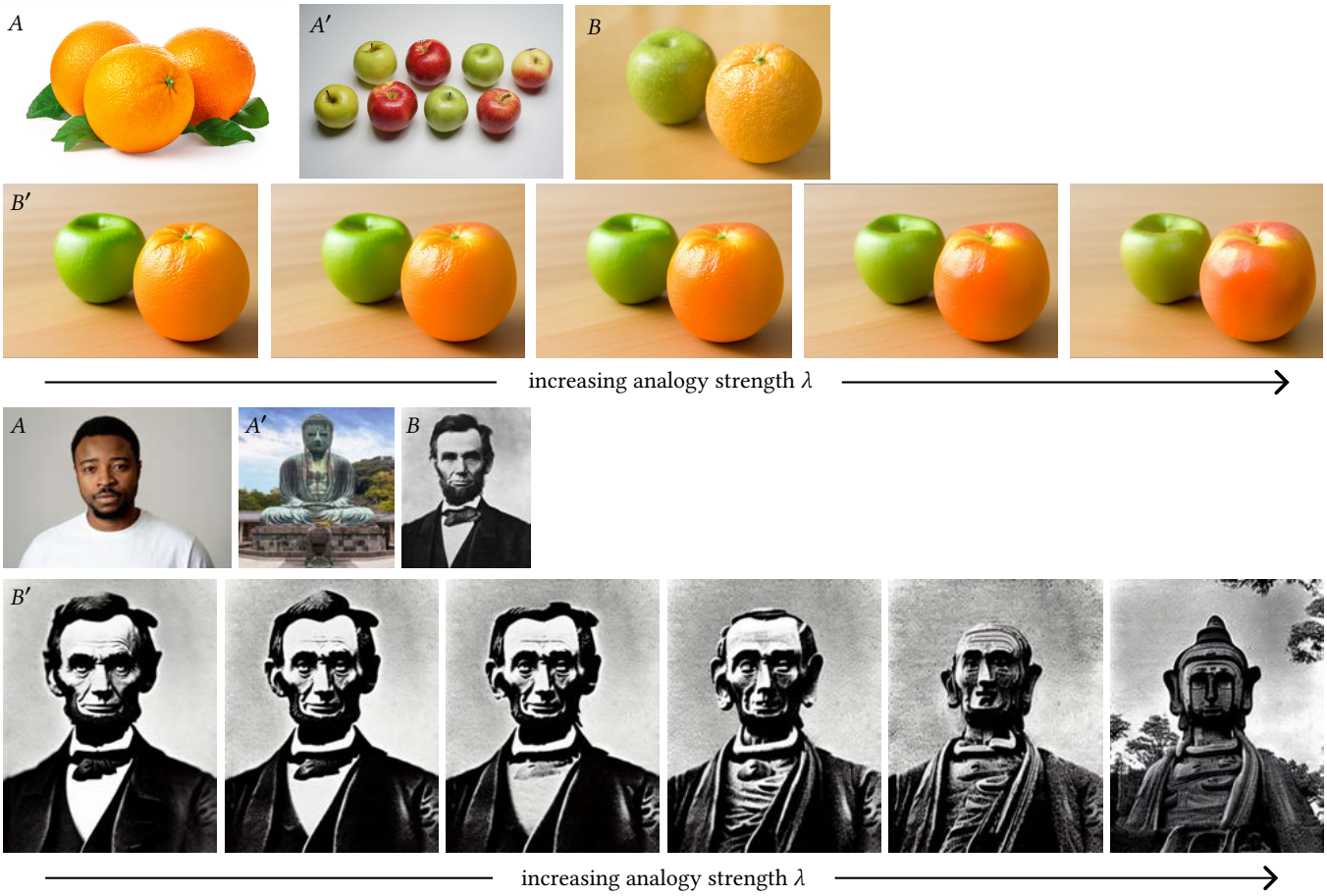

Figure 3: Examples of diffusion image analogies $A : A' :: B : B'$ produced using our approach with gradually increasing analogy strength λ (cont.). Source images: © Dllu (Apples A'), © The Busy Brain (Apple & Orange B), © U.S. Department of Agriculture (Lincoln B), Adobe Stock the rest.

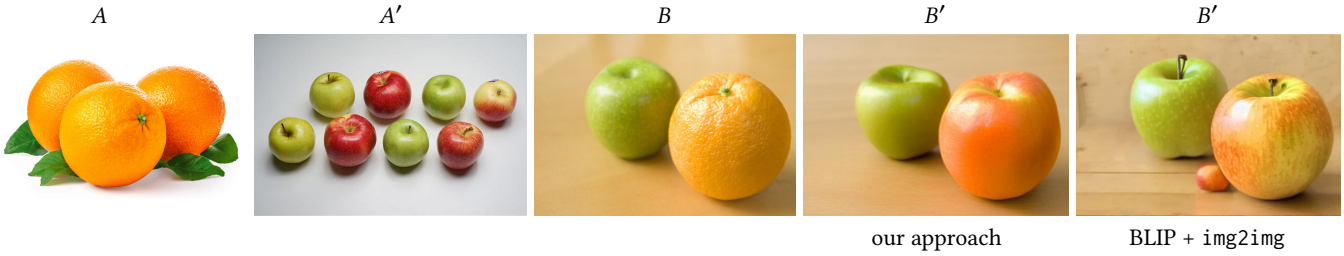

Figure 4: Additional comparison with Stable Diffusion running in the `img2img` mode [Rombach et al. 2022] where BLIP [Li et al. 2022] was used to estimate text prompts of the images from which the analogy of CLIP features is computed (c.f. the main paper for detailed explanation). In this example A =*"three oranges with leaves on a white background"*, A' =*"a group of apples sitting on top of a white table"*, and B =*"two oranges and an apple on a table"*. The erroneous notion of three pieces in the estimated description of B slightly biases the result towards a spurious small fruit visible in the output image B'. Source images: Adobe Stock A, © Dllu A', © The Busy Brain B.

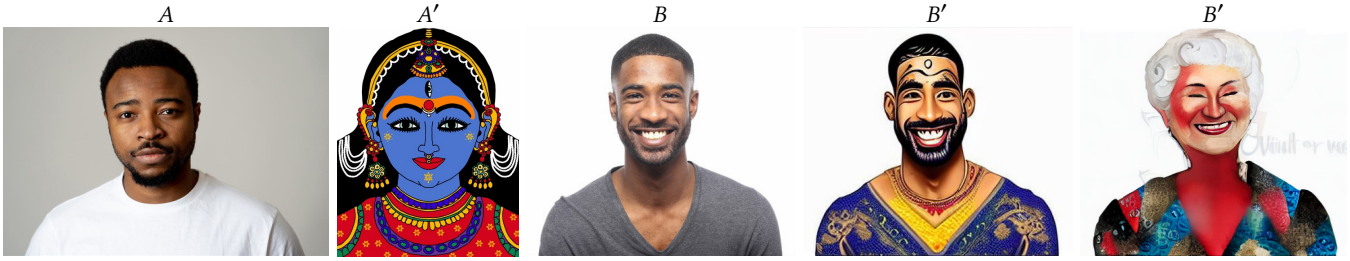

Figure 5: Comparison with Stable Diffusion (cont.): A =***"a man in a white T-shirt looks at the camera"***, A' =***"a painting of a woman in a red dress"***, **and** B=***"a man with a smile on his face"***. **The estimated description of** A' **does not express the style with sufficient accuracy. Source images: Adobe Stock.**

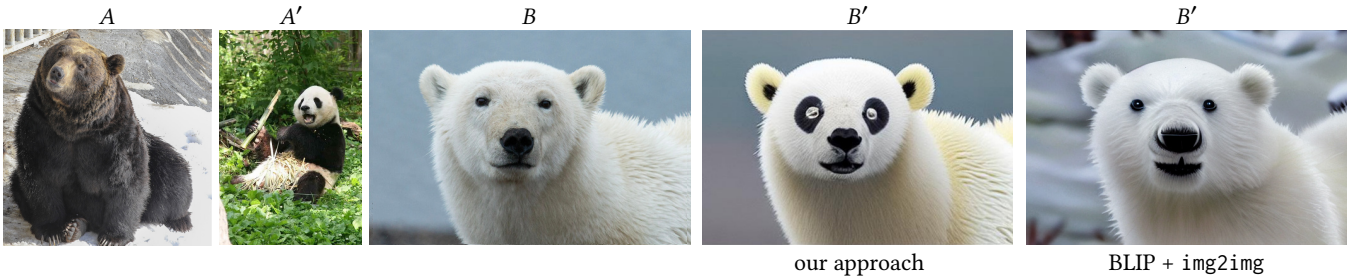

Figure 6: Comparison with Stable Diffusion (cont.): A =***"a large brown bear sitting in the snow"***, A' =***"a panda bear sitting in the grass eating bamboo"***, **and** B:***"a close up of a polar bear looking at the camera"***. **Due to relatively cluttered description of** A' **the analogy is not sufficiently strong to reproduce panda's face while at the same time preserving the prescribed structure. Source images: © Artanisen** A**, © Cliff** A'**, Adobe Stock** B**.**