Improving Subject-Driven Image Synthesis with Subject-Agnostic Guidance –Supplementary Material–

1. Experimental Settings

To train ELITE-SAG, we use a subset of the the WebLI [2] dataset for training. We randomly select 70M data from the master dataset. We further extract 1M data containing *dogs* and *cats* with their face size greater than 128×128 as our domain-specific dataset. The remaining data is used as our general-domain dataset. The dataset mixing ratio is 0.1. In this work, the weak condition $\mathbf{c_0}$ is obtained simply by replacing the special token with the class of the subject (*e.g.*, "dog" or "cat"). We train our models with 8 TPUv4 chips for 300,000 iterations. The learning rate is set to 10^{-4}. The method is implemented in JAX [1].

2. Additional Results for Textual Inversion

Additional results are shown in Fig. 1. With our SAG, Textual Inversion is able to produce results that are better align with the text descriptions.

3. Additional Results for SuTI

We provide additional results for applying SAG on SuTI in Fig. 2, without SAG, while identity is successfully preserved, the outputs often fail to capture the details specified by the text prompts. In contrast, text alignment is much better with SAG, without sacrificing identity.

4. Additional Results for DreamSuTI

As depicted in Fig. 3, given a SuTI fine-tuned on a given style, our SAG leads to better style alignment while being able to preserve subject identity.

5. Additional Comparison Using ELITE-SAG

We provide additional comparison with DreamBooth [4], Textual Inversion [3], and ELITE [5]. As shown in Fig. 6 to Fig. 4, while existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.

References

[1] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 1

[2] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Chechik Gal Bermano, Amit H., and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

[5] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 1

Figure 1. **More Results on Textual Inversion.** With SAG, Textual Inversion is able to produce results that are better align with the text descriptions.

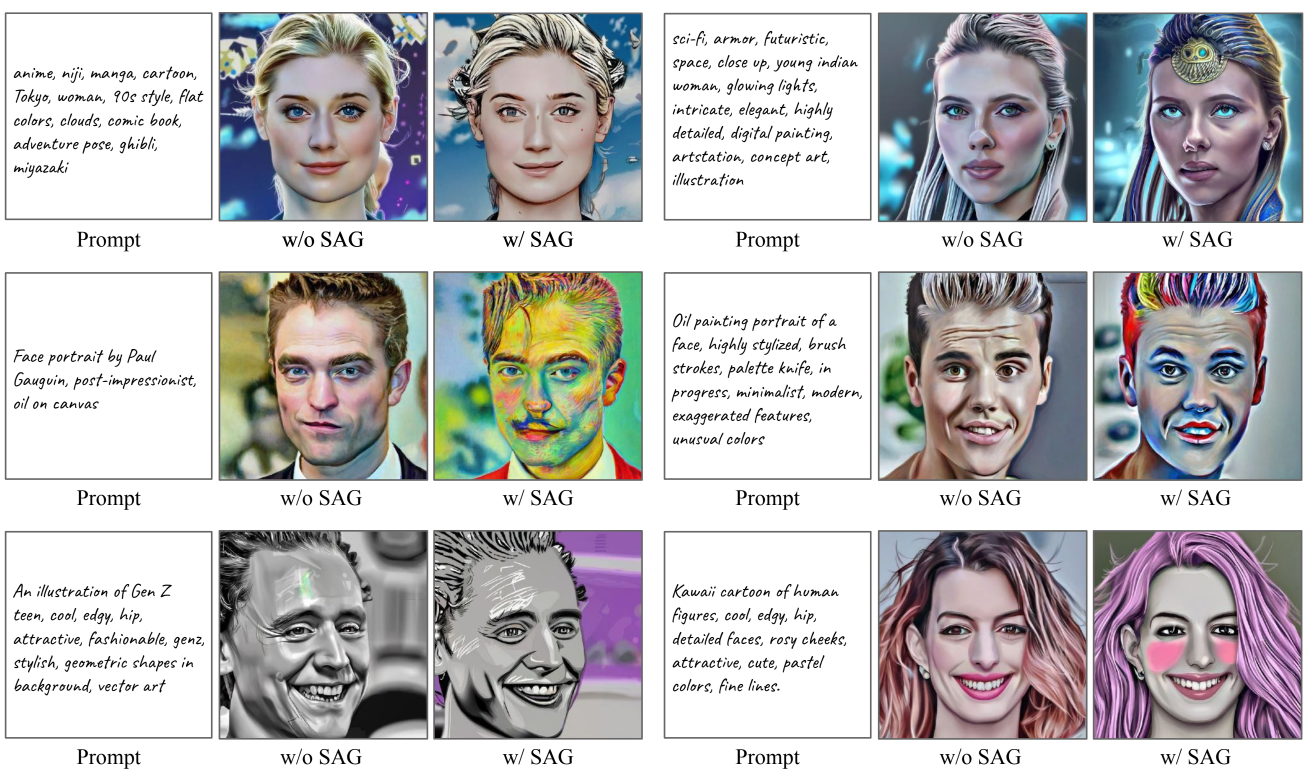

Figure 2. **More Results on SuTI.** Without SAG, while identity is successfully preserved, the outputs often fail to capture the details specified by the text prompts. In contrast, text alignment is much better with SAG, without sacrificing identity. Reference images are not provided to protect privacy.

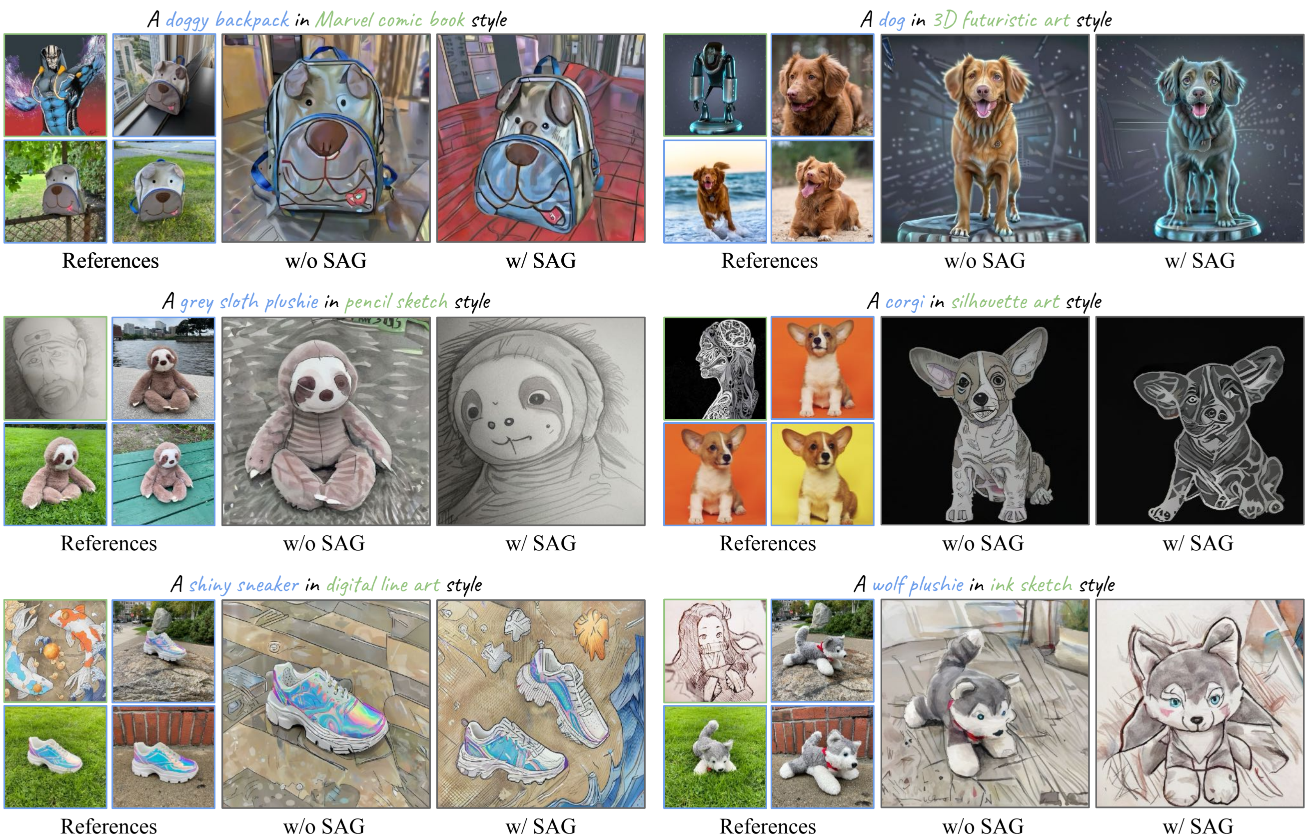

Figure 3. **More Results on DreamSuTI.** Given a SuTI fine-tuned on a given style, our SAG leads to better style alignment while being able to preserve subject identity.

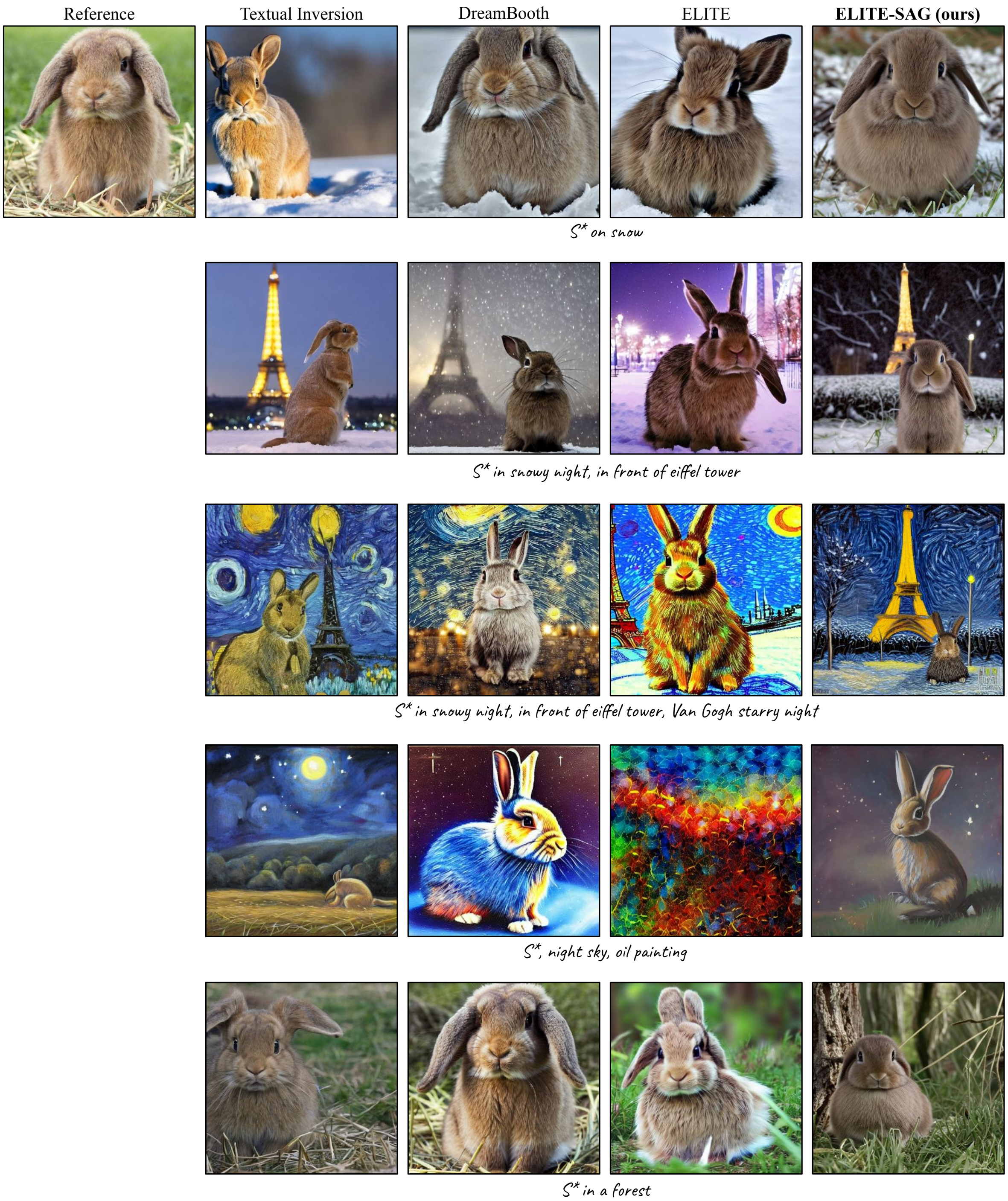

Figure 4. **More Comparison Using ELITE-SAG.** While existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.

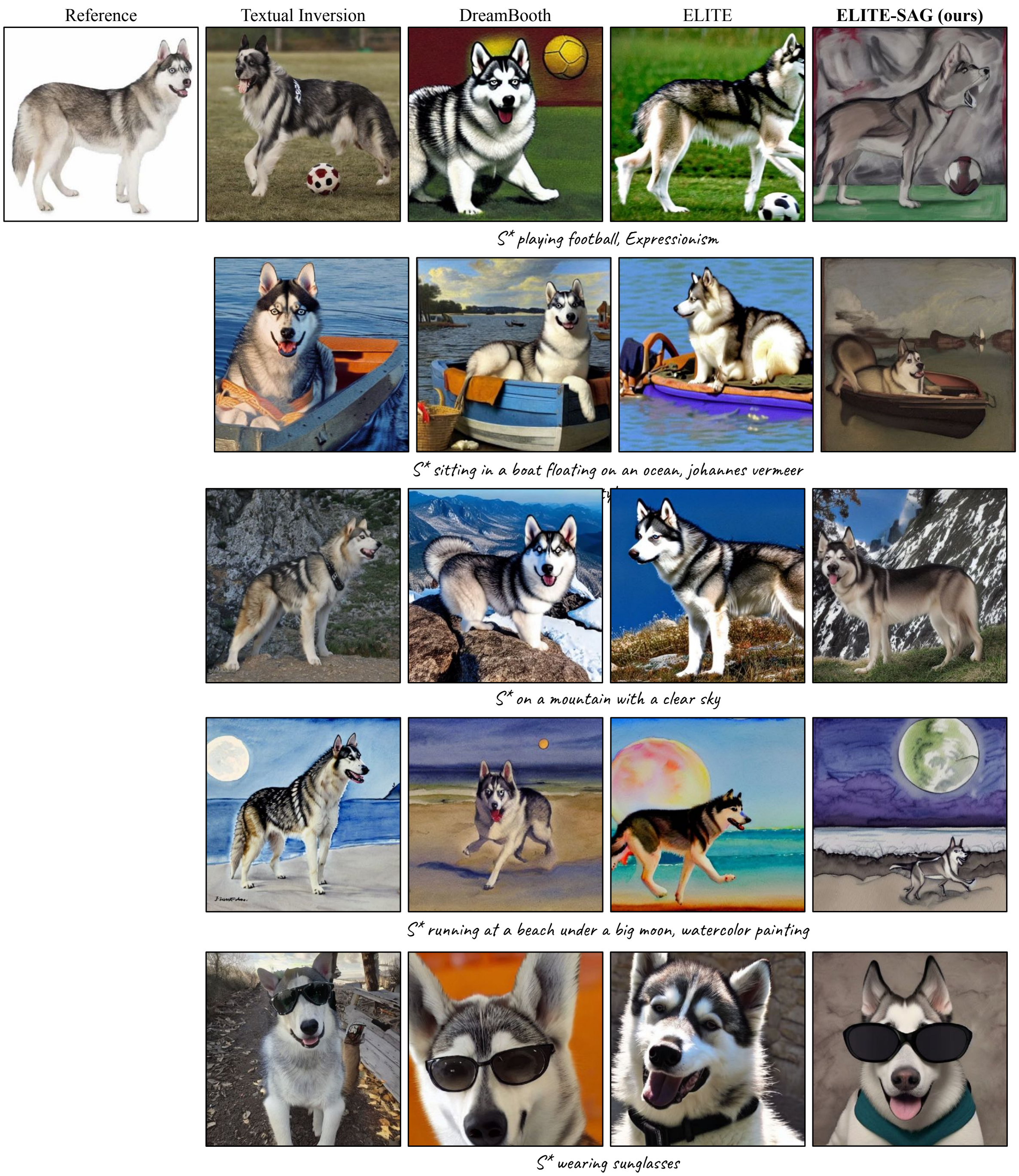

Figure 5. **More Comparison Using ELITE-SAG.** While existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.

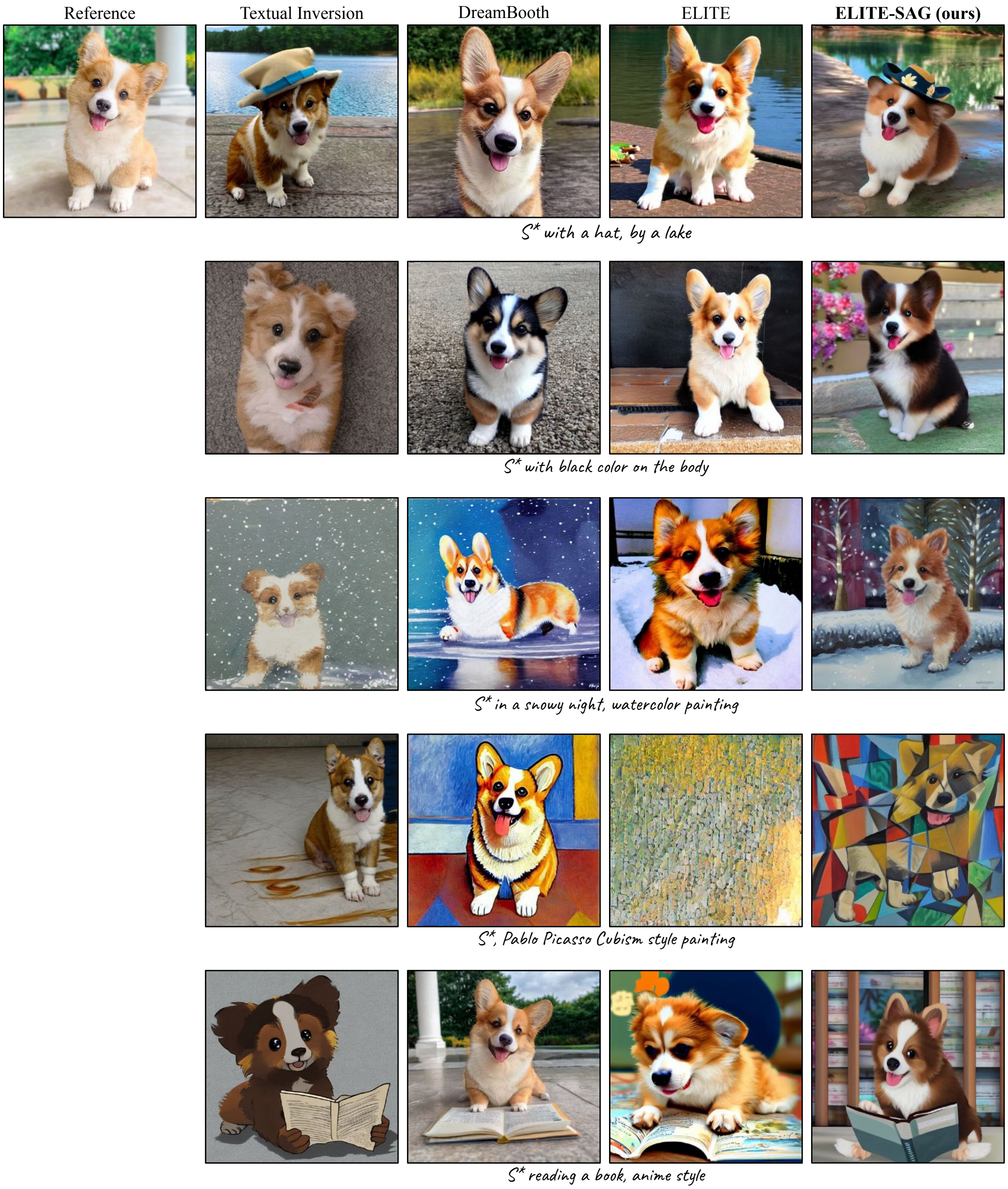

Figure 6. **More Comparison Using ELITE-SAG.** While existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.

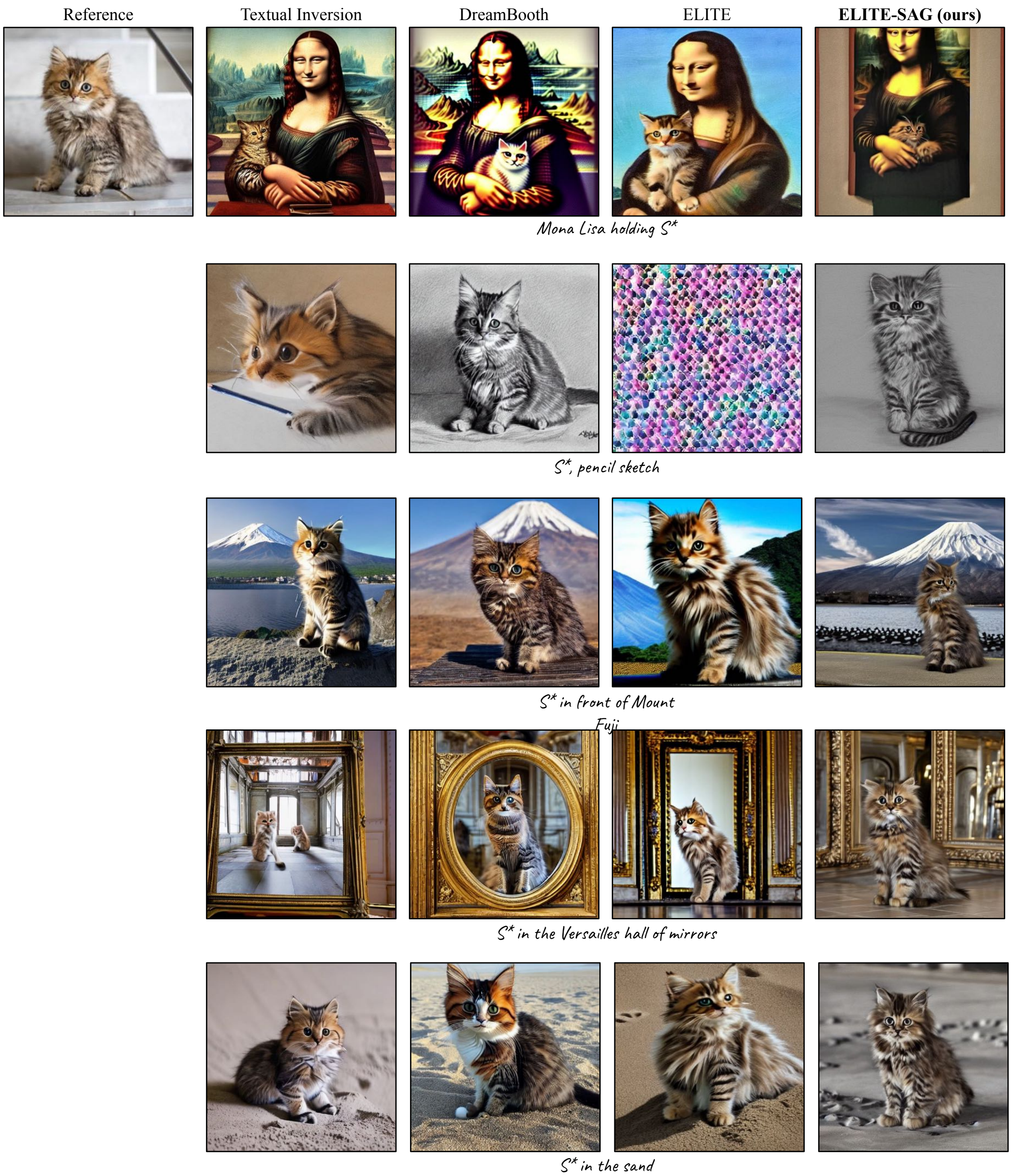

Figure 7. **More Comparison Using ELITE-SAG.** While existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.

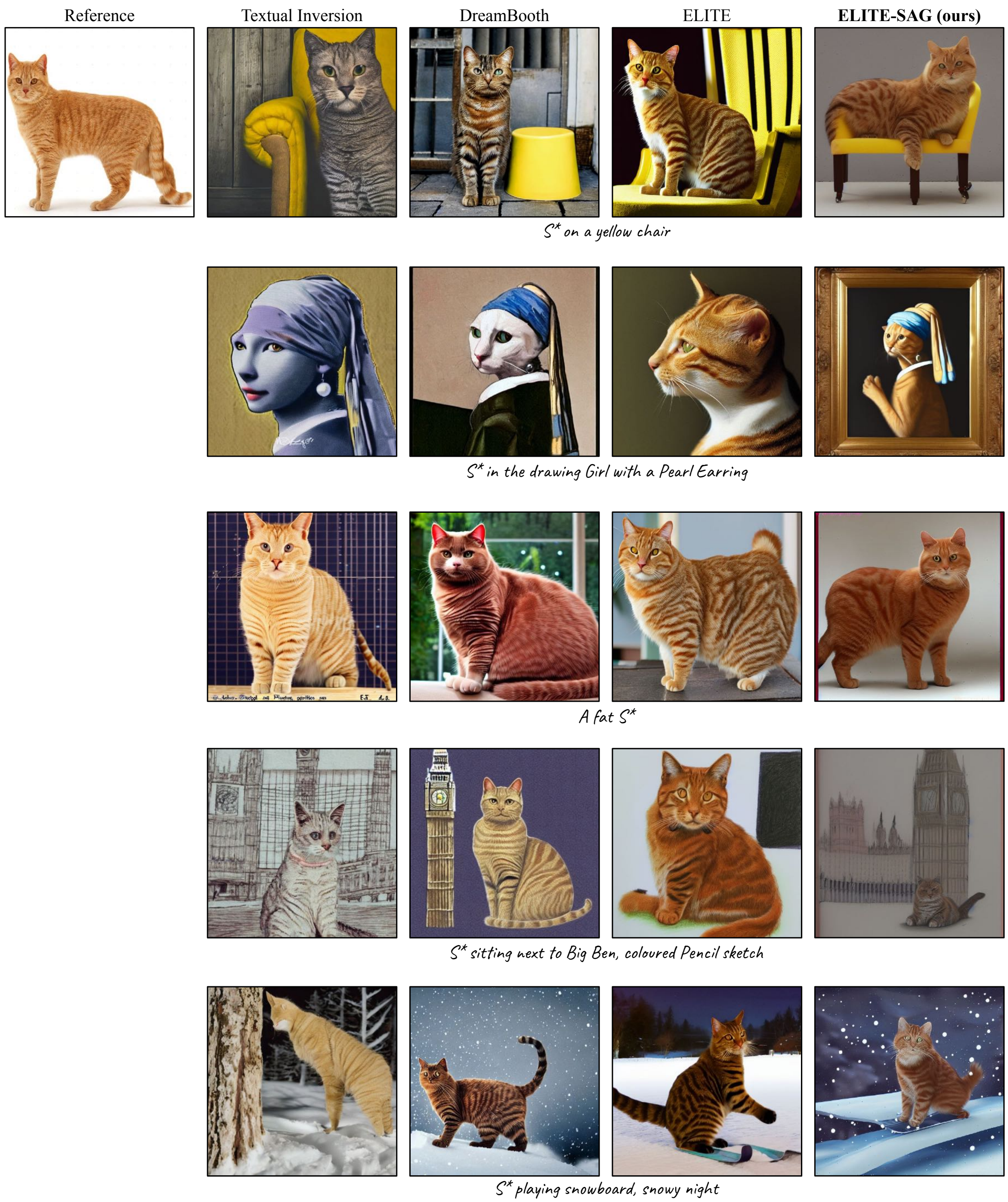

Figure 8. **More Comparison Using ELITE-SAG.** While existing works generally produce reasonable results, they often experience content ignorance or insufficient subject fidelity. This observation is especially obvious in complicated prompts, where the model has to comprehend complex relation between subjects. In contrast, ELITE-SAG achieves a balance between prompt consistency and subject fidelity.