ObjectMate: A Recurrence Prior for Object Insertion and Subject-Driven Generation

Supplementary Material

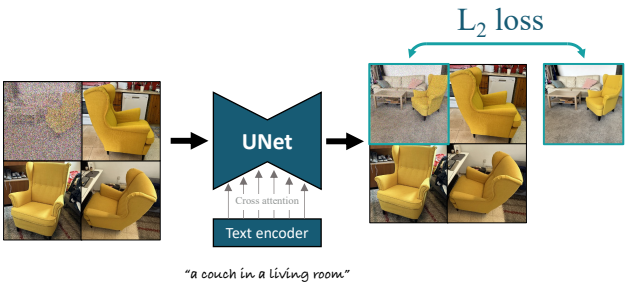

Figure 1. Subject-driven generation model's architecture.

A. Implementation details

Training. As detailed in Sec. 5, we train two separate models: one for object insertion and another for subject-driven generation. Fig. 7 in the main manuscript illustrates the architecture of our object insertion model. Additionally, App. Fig. 1 provides a diagram for the subject-driven generation model.

The primary difference between these architectures lies in how the input is integrated into the UNet. For object insertion, the scene description, background image and mask are concatenated along the channel axis with the noise input. In contrast, for subject-driven generation, the scene description is provided as a text prompt and incorporated into the UNet via standard cross-attention layers.

During object insertion training, we use an empty text prompt. The mask indicating the target object's location is the bounding box of the object rather than a precise mask.

k-Nearest Neighbors (kNN) search. For each detected object in our dataset, we compute retrieval-specific features designed for instance retrieval without local feature matching. This design makes them well-suited for large-scale kNN searches. Using the Python library ScaNN [5], we calculate the cosine similarity of features between all object pairs in the dataset. In the final dataset, we retain the top 5 nearest neighbors with similarity scores ranging from 0.93 to 0.975, as detailed in Section 4.

A.1. Classifier-Free Guidance

Following Brooks et al. [2], we apply classifier-free guidance (CFG) [6] to both text and image conditions. CFG is a widely used method to enhance the model's adherence to its conditioning inputs. This involves jointly training the model for both conditional and unconditional generation and leveraging both modes during inference.

Object insertion. In object insertion, we modify the training process by zeroing out the reference condition O in 10% of the training examples, while keeping the scene condition S (background images and masks) unchanged. During inference, the model's output is adjusted using the following formula:

$$\begin{aligned}\tilde{D_\theta}(x_t, O, S) =& D_\theta(x_t, \varnothing, S) \\ &+ \gamma_I \cdot (D_\theta(x_t, O, S) - D_\theta(x_t, \varnothing, S))\end{aligned}$$

Here, γ_I controls the influence of the reference condition, we empirically set $\gamma_I = 2$.

Subject-driven generation. For subject-driven generation, in 10% of the training examples, we zero out the reference condition O, and in another 10%, we use an empty prompt for the scene description S. During inference, the model's output is adjusted as follows:

$$\begin{aligned}\tilde{D_\theta}(x_t, O, S) =& D_\theta(x_t, \varnothing, \varnothing) \\ &+ \gamma_{txt} \cdot (D_\theta(x_t, O, S) - D_\theta(x_t, O, \varnothing)) \\ &+ \gamma_I \cdot (D_\theta(x_t, O, \varnothing) - D_\theta(x_t, \varnothing, \varnothing))\end{aligned}$$

Here, γ_{txt}, γ_I controls the strength of the text condition (scene description) and references condition respectively. We use constant values of $\gamma_I = 1.5$ and $\gamma_{txt} = 7.5$.

A.2. Dataset statistics

In Sec. 4 we use the train split of the datasets COCO [10], Open Images [8], and a web-based dataset with 48M images. We provide dataset statistics in App. Tab. 1.

B. Additional comparisons

Retrieval augmented models. As discussed in Sec. 2, several studies [1, 3, 4, 7, 9, 14] have used nearest neighbor (NN) retrieval to enhance generation fidelity. Specifically, [1, 3, 9, 14] retrieve the NNs based on the text prompt provided during inference to improve the generation of rare concepts. SuTI [4] and Instruct-Imagen [7] cluster images from the same URL and refine them using CLIP image similarity calculated at the whole-image level. Our approach differs in two key ways: (1) we employ an instance retrieval (IR) model that better distinguishes between identities with similar semantics compared to CLIP, and (2) we calculate

Dataset	# Images	# Objects	Detection type	# Examples with at least 1 NN	# Examples with at least 3 NNs
COCO	108,151	362,684	Human annotations	31,445 (8.7%)	17,119 (4.7%)
Open Images	1,743,042	8,067,907	Human annotations	471,091 (5.8%)	64,991 (2.4%)
Web-based	47,992,480	55,232,441	Object detection model	9,947,017 (18%)	4,550,770 (8.2%)

Table 1. Datasets statistics.

Method	Test-time tuning-free	Text-Align. (CLIP-T)	Identity (IR)
Dreambooth (SD-XL)	✗	0.306	0.674
DisenBooth	✗	0.301	0.728
BLIP-Diffusion	✓	0.288	0.664
Ours	✓	0.322	0.750
Ours, SD-XL	✓	0.304	0.757
Ours, SD-XL + Public data	✓	0.304	0.739

Table 2. Subject driven: comparison on public reproducible setup.

similarity at the object level rather than for the entire image. These differences result in object clusters with a higher likelihood of representing the same identity.

Since SuTI and Instruct-Imagen have not released their models, we compare our results with those reported in their manuscripts. App. Fig. 10 compares results where SuTI uses 5 references and our model uses 3. Our approach consistently achieves better identity preservation. Additionally, App. Fig. 11 compares our results with SuTI where both models use either 1 or 3 references. App. Fig. 12 qualitatively compares our model with Instruct-Imagen, demonstrating superior preservation of fine object details.

Counterfactual object insertison. Similarly to ObjectDrop [15], we trained an object removal model using 2,000 counterfactual examples. We then used this model to synthesize the backgrounds for object insertion training. ObjectDrop's approach involves training an object insertion model by first removing objects from images and then reinserting them into their original positions. For comparison, we implemented this approach in our experiments.

When inserting objects into a scene, the ObjectDrop model pastes them and generates only their effects on the surroundings. While this ensures identity preservation, it does not allow for adjustments to the pose or lighting of the inserted objects. In contrast, our model incorporates these capabilities, enabling more realistic harmonization of the object with the scene. App. Fig. 8 highlights our model's superior performance in harmonizing lighting and pose.

Retrieval and DINO features. We conducted an ablation study to assess the importance of instance retrieval (IR) features in our model's performance. Specifically, we used DINO features to perform kNN search on the same image dataset used in our primary experiments. Subsequently, we trained a subject generation model using the retrieval results based on these features. Notably, DINO features tend to identify objects with only semantic similarities (as illustrated in Fig. 2), which substantially influences the downstream performance of the model. To complement the findings of the user study presented in the main manuscript, App. Fig. 6 provides qualitative evidence showing that our model achieves superior identity preservation compared to a model trained using DINO-based retrievals.

More results. We extend the qualitative comparisons presented in the main manuscript with the following figures:

- Fig. 5 complements the quantitative comparison between different retrieval features made in Fig. 9 of the main manuscript.
- Fig. 7 shows that using publicly available dataset and IR features outperforms current SOTA insertion method.
- Fig. 13 shows a creative application.
- Fig. 14 presents failure cases.
- Fig. 9, 15, and 16 show additional examples of object insertion.
- Fig. 17 and 18 present additional examples of subject-driven generation.

C. User study

To evaluate the performance of our models, we conducted a detailed user study on the CloudResearch platform. For the object insertion task, we had 50 participants, randomly selected, primarily from the United States. Each participant reviewed 25 examples drawn from our benchmark dataset comprising 136 examples. For each example, participants were presented with two images in random order: one generated by our model and another by a baseline model. Participants were asked to answer the following questions:

1. *Which image looks more realistic and natural?*
2. *In which image the subject is more similar to the reference?*

The responses to the first question were used to compute the *Composition* score, while the responses to the second

Figure 2. A screenshot of the user study questionnaire.

question contributed to the *Identity* score. The results of this study are presented in Tab. 4 of the main manuscript.

For the subject-driven generation task, 45 participants completed a similar questionnaire with the following questions:

1. *Which image matches the text prompt more?*
2. *In which image the subject is more similar to the references?*

In this evaluation we used the public benchmark DreamBench, which includes 30 unique objects and 25 textual prompts, resulting in a total of 750 examples. The results are summarized in Tab. 5 of the main manuscript. Fig. 2 shows a screenshot of the questionnaire.

D. Quantitative evaluation protocol

As outlined in Sec. 6, existing quantitative metrics, such as CLIP and DINO, primarily evaluate semantic similarity rather than the preservation of identity. To address this, we propose using the instance retrieval (IR) features from [12], which we demonstrate to be more closely aligned with user preferences for identity preservation (see Tab. 3 in the main manuscript). Below, we detail the evaluation protocol used in our approach.

Given a generated image I_g and a reference image of the subject I_{ref}, we begin by detecting the bounding box of the subject in I_g using [11] with the object's class name as input. The generated image I_g is then cropped to this bounding box, resulting in $\tilde{I}_g$. Next, we compute the IR features, denoted as $\mathcal{E}$, for both $\tilde{I}_g$ and I_{ref}. Specifically, these features are represented as $\mathcal{E}(\tilde{I}_g)$ and $\mathcal{E}(I_{ref})$, respectively. Finally, the IR identity preservation score is determined by calculating the cosine similarity between $\mathcal{E}(\tilde{I}_g)$ and $\mathcal{E}(I_{ref})$. The weights of the encoder $\mathcal{E}$ are publicly available to download from [13].

To validate this protocol, we analyzed user study responses regarding identity preservation (see Sec. C). Each response comprises a triplet $(I_{ref}, I_{g_1}, I_{g_2})$, where I_{g_1} is the output of our model, I_{g_2} is the output or one of the baselines, and $y \in \{1, 2\}$ indicates the user's choice for better identity preservation. For evaluating the validity of the metrics, the user responses serve as ground truth and we measure the accuracy of each metric in predicting user preferences. As presented in Tab. 3 of the main manuscript, IR demonstrates significantly improved performance over existing metrics, confirming the strong alignment between our automated evaluation method and human judgment.

E. Object insertion benchmark

We introduce a new benchmark for object insertion. The benchmark comprises a test set of 34 distinct objects, each captured in 4 different poses and scenes, representing variations such as indoor/outdoor settings and different times of day (e.g., daytime vs. nighttime). For each scene, we use a tripod-mounted camera to capture images both with and without the object. From each quadruplet of images, we extract 4 samples: a ground truth image (y), the background of the scene as a scene description (S), and 3 reference images (O). This results in a total of 136 samples. To the best of our knowledge, this is the first object insertion dataset that includes ground truth images and 3 reference views of the object. Fig. 3 shows an example of such quadruplet.

References

[1] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *arXiv preprint arXiv:2204.11824*, 2022.

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[3] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

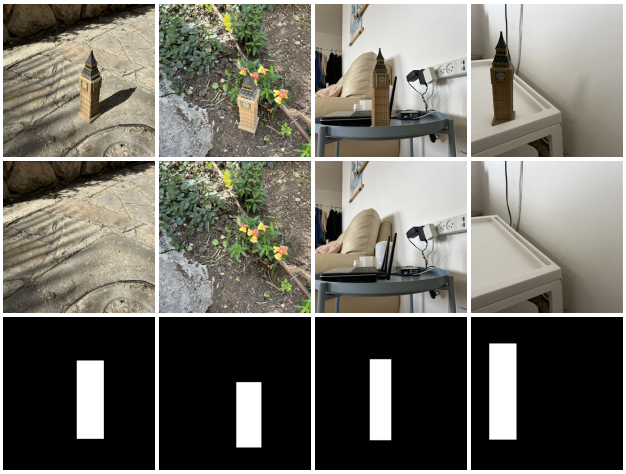

Figure 3. Example of a quadruplet from out test set. From each quadruplet we extract 4 samples, where one object is used as the ground truth and the remaining 3 serve as the reference condition.

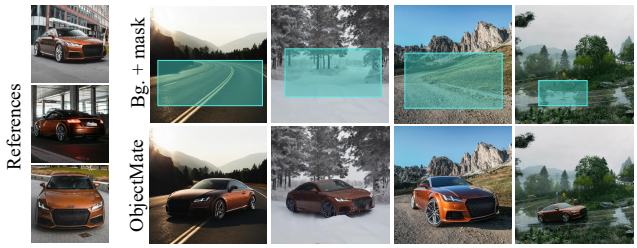

Figure 4. Inserting the same object into different scenes.

[4] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR, 2020.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[7] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024.

[8] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[9] Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. Memory-driven text-to-image generation. *arXiv preprint arXiv:2208.07022*, 2022.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[12] Shihao Shao and Qinghua Cui. 1st place solution in google universal images embedding. *arXiv preprint arXiv:2210.08473*, 2022.

[13] Shihao Shao and Qinghua Cui. 1st solution in google universal image embedding. https://www.kaggle.com/datasets/louieshao/guieweights0732, 2023.

[14] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

[15] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *Computer Vision – ECCV 2024*, pages 112–129, Cham, 2024. Springer Nature Switzerland.

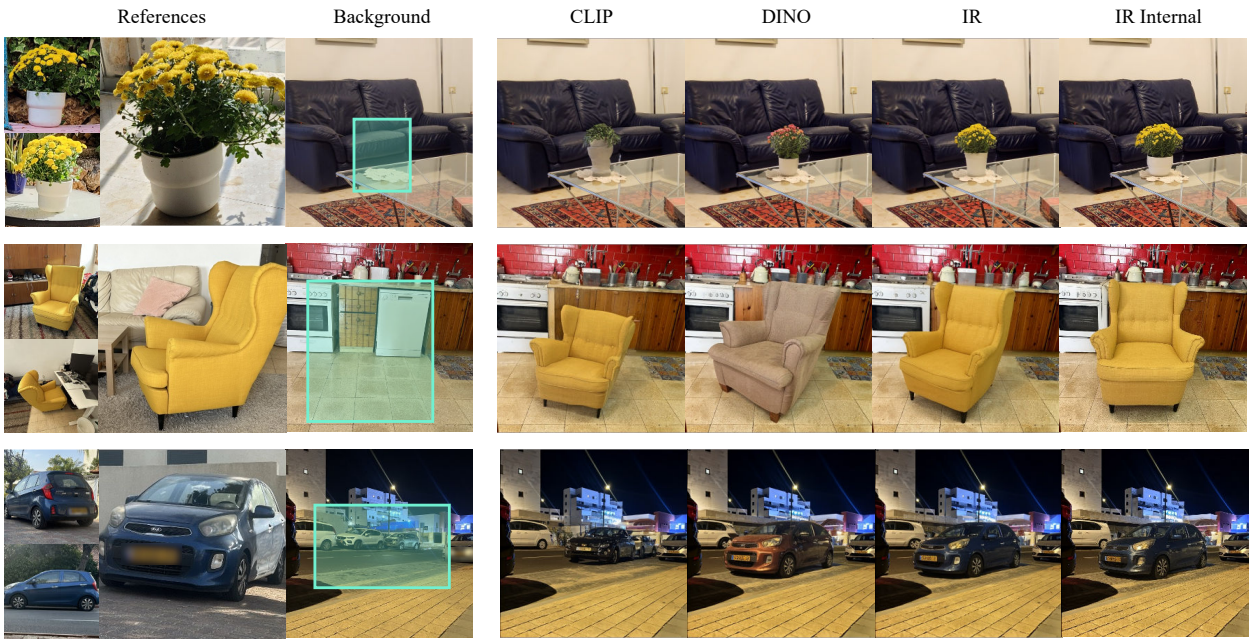

Figure 5. **Ablation study on the importance of IR features for object insertion.** Using CLIP or DINO features for instance retrieval during object insertion training is insufficient to achieve identity preservation. Using specialized instance-retrieval (IR) features achieve much stronger results. In addition, the publicly available IR model from [12] is comparable to our internal model.

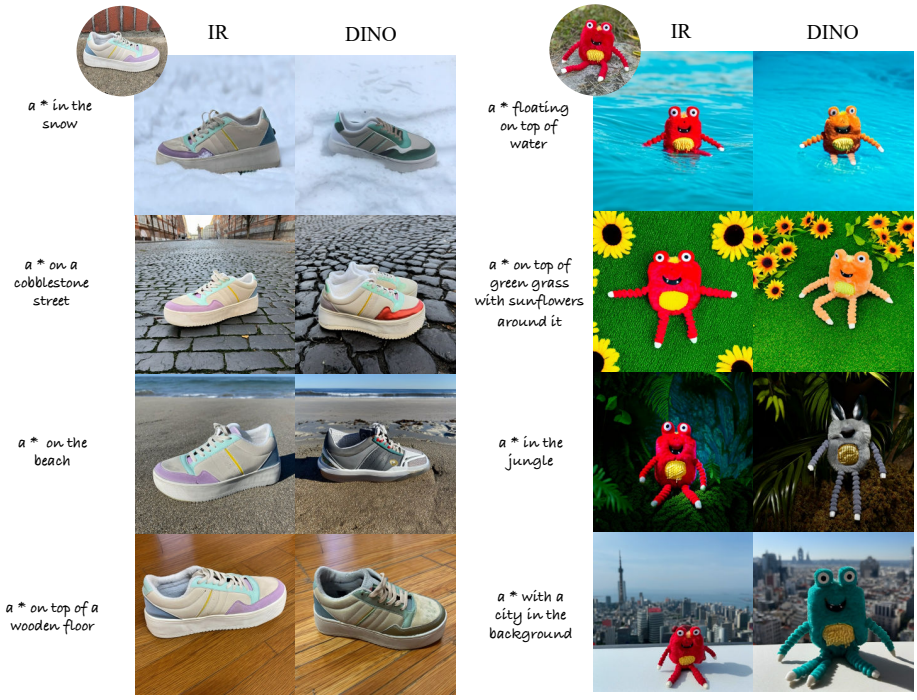

Figure 6. **Ablation study on the importance of IR features for subject generation.** Our subject generation model, denoted as IR, demonstrates superior identity preservation compared to a model trained using DINO-based retrievals.

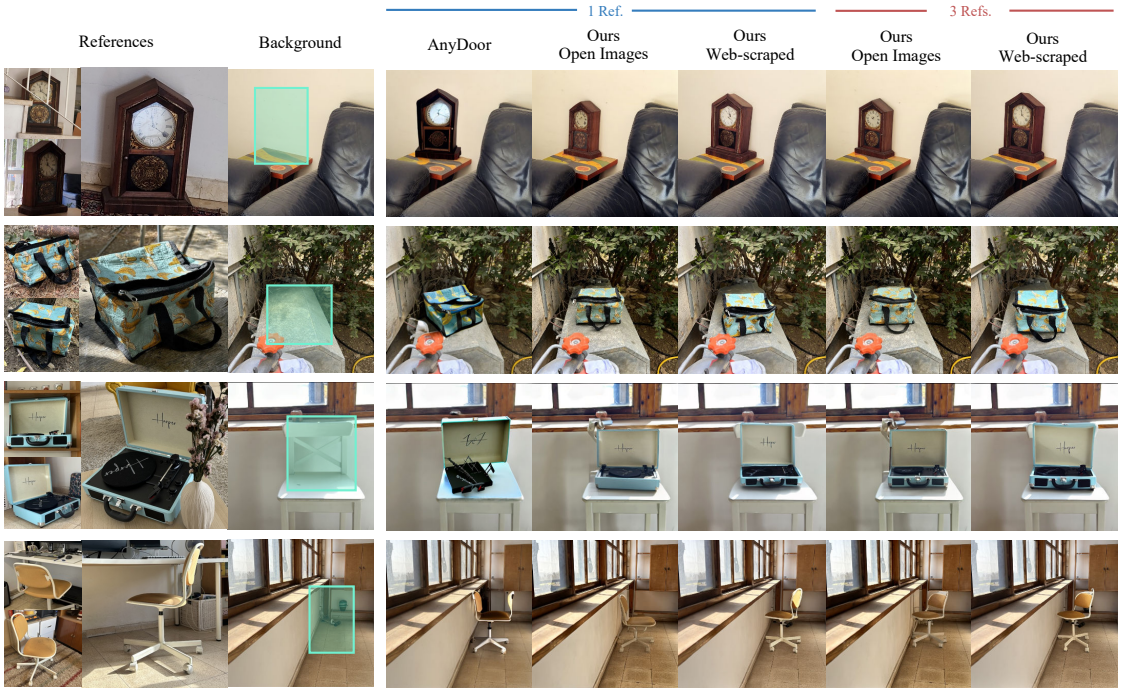

Figure 7. **Ablation study on data sources.** We compare the effectiveness of different data sources for training. Training on Open Images with publicly available IR features and on a web-scraped dataset using our internal IR model both outperform the current state-of-the-art insertion model, AnyDoor.

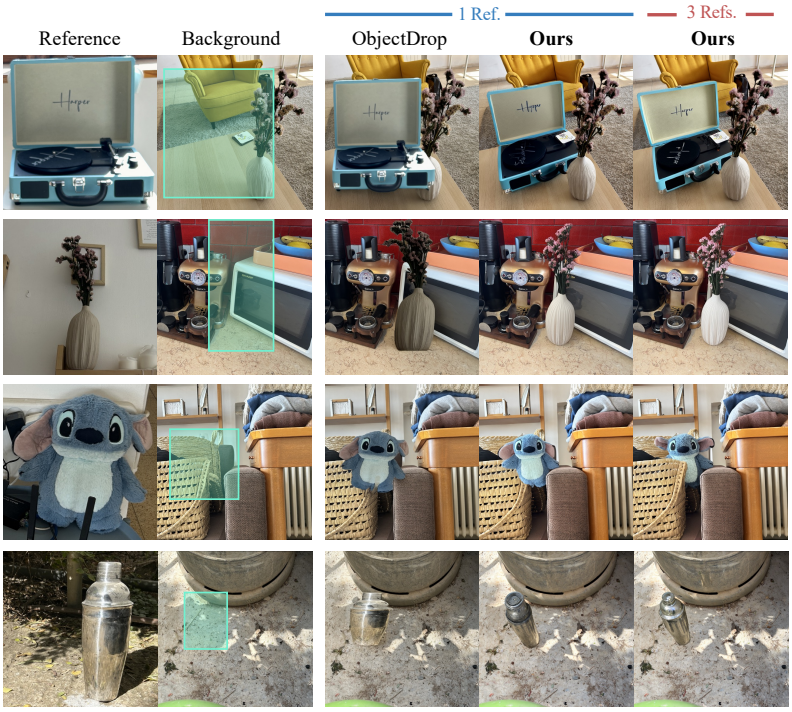

Figure 8. **Comparison with counterfactual object insertion.** We compare to a model similar ObjectDrop. Our model is able to realistically harmonize the object's pose and lighting, while the counterfactual model pastes the object without adjustments.

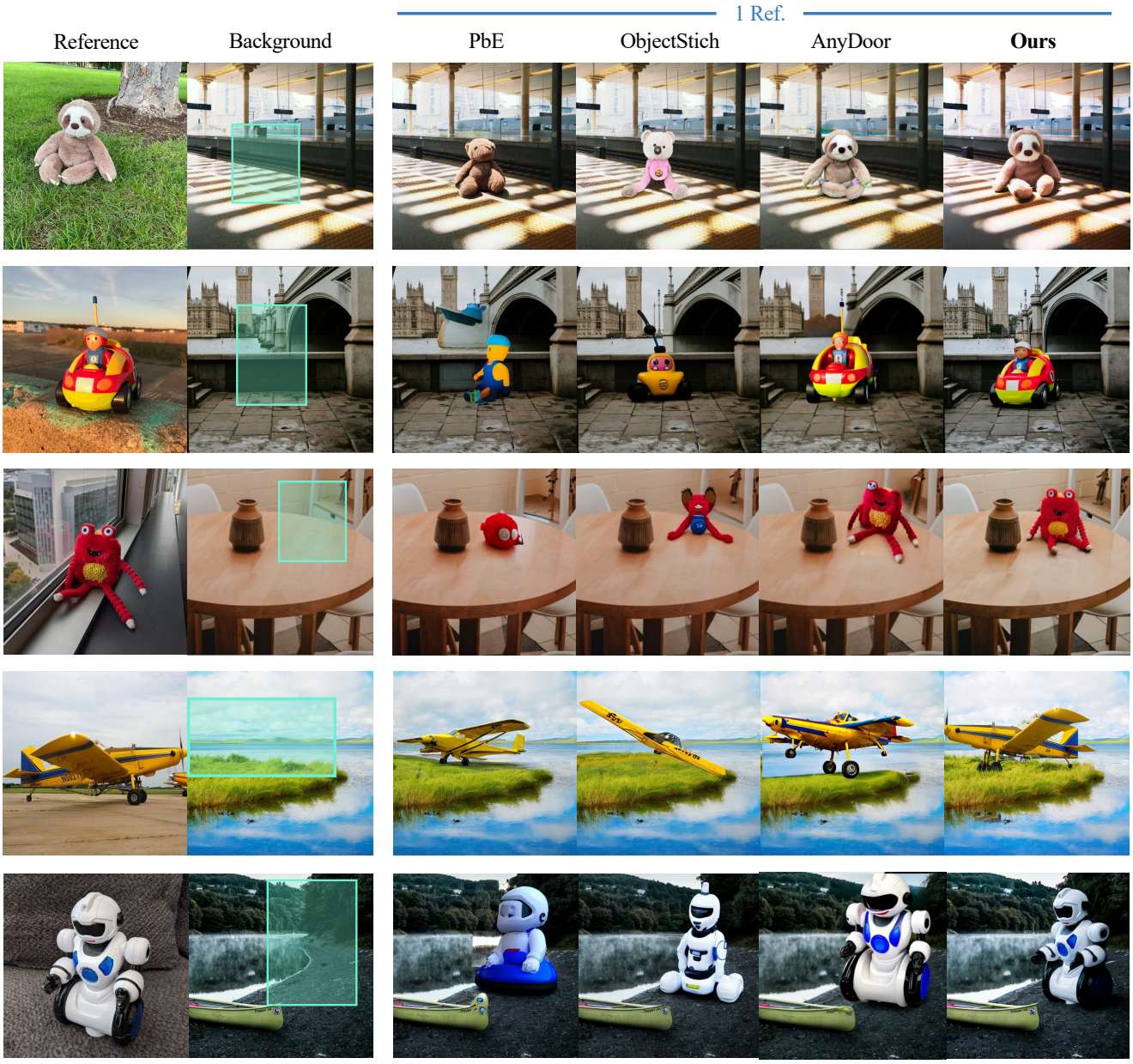

Figure 9. Additional in-the-wild object insertion results.

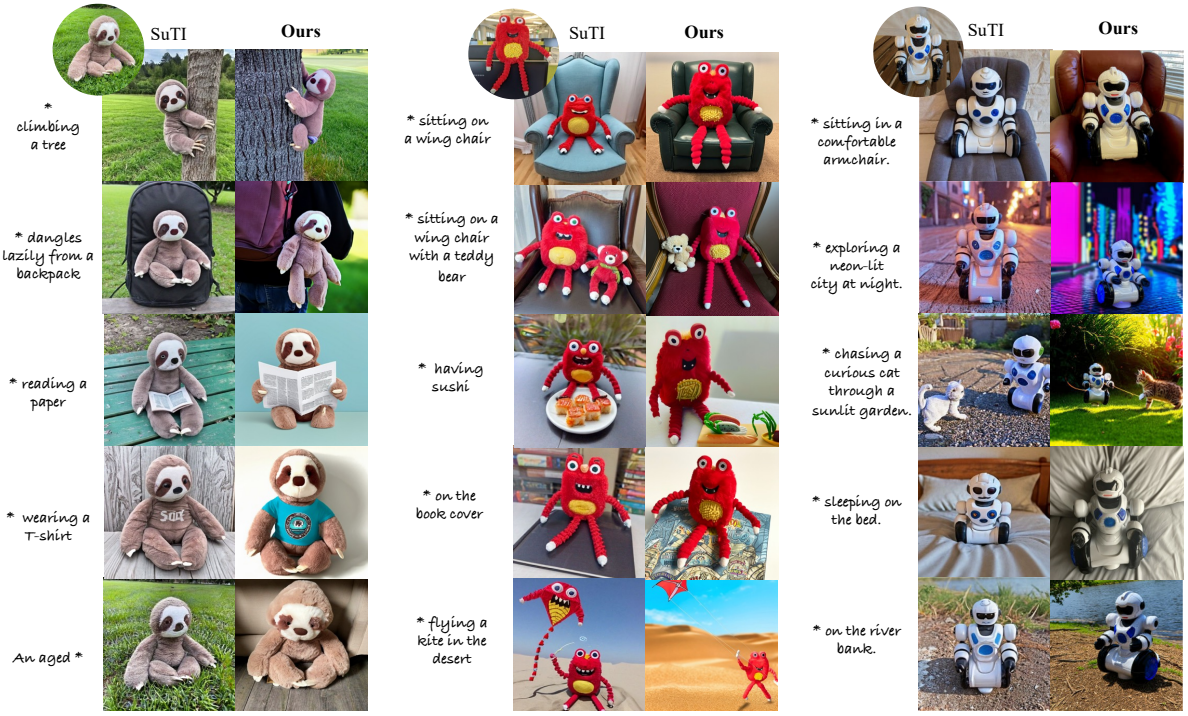

Figure 10. **Comparison with SuTI.** Our method better preserves the fine details of the subjects. SuTI uses semantic features (CLIP) for retrieval, while we use specialized instance-retrieval features. This makes our paired data more suitable for identity preservation. Results of SuTI are taken from their manuscript. Here, SuTI uses 5 references, while we use 3.

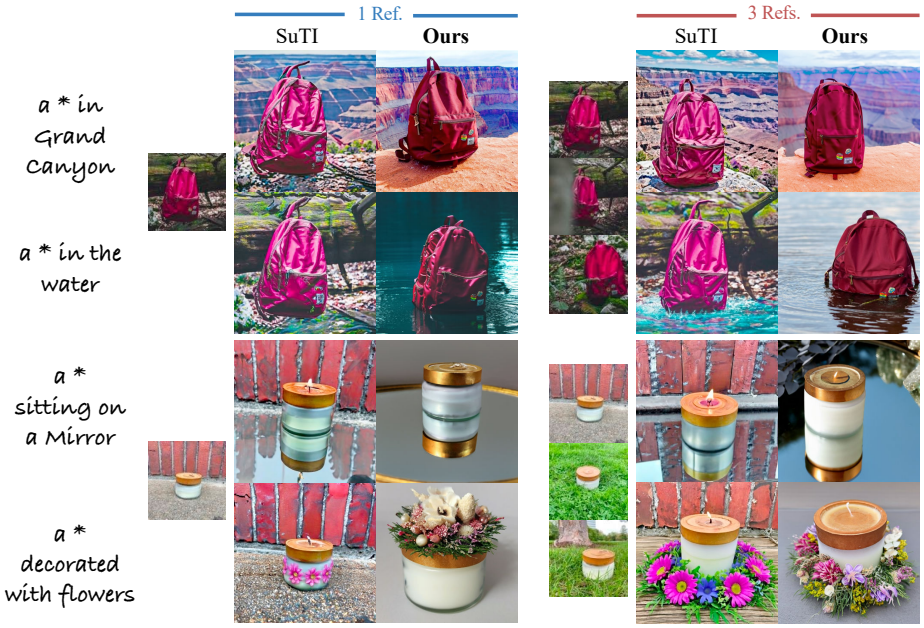

Figure 11. **Comparison with SuTI.** Our model demonstrates superior capability in preserving fine details of the object, regardless of whether 1 or 3 reference images are provided by the user. Results of SuTI are taken from their manuscript.

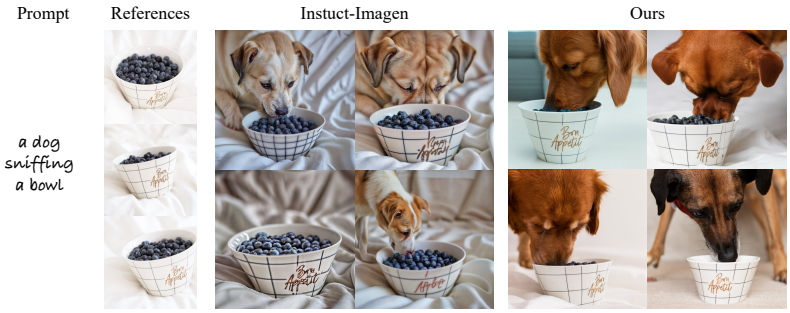

Figure 12. **Comparison with Instruct-Imagen.** Our method better preserves the fine details of the bowl (e.g., text decoration). Instruct-Imagen uses similar data to SuTI, which is based on semantic clustering. Results of Instruct-Imagen are taken from their manuscript.

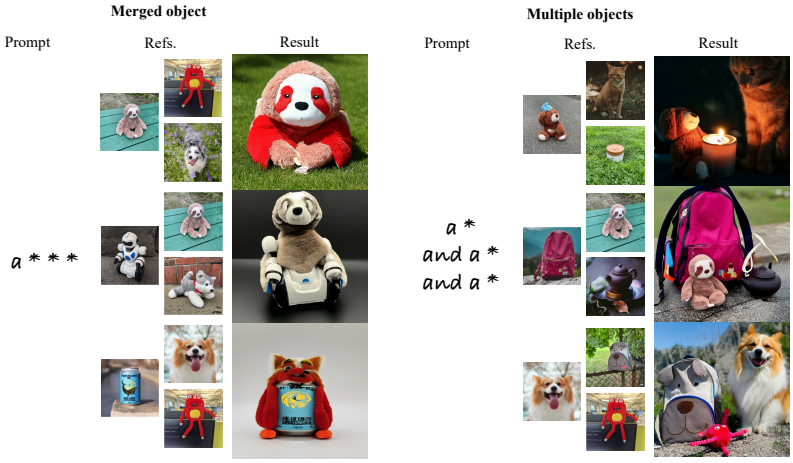

Figure 13. **Creative application.** We test the model's generalization by providing it with three references of *different* objects. This setup represents a significant deviation from the training distribution, where the model received three references of the same object. Remarkably, the model demonstrates an ability to generalize beyond its training data by either synthesizing the references into a single unified object or generating the three objects separately.

(a) (b)

Figure 14. **Limitations. (a)** This study primarily focuses on preserving subject identity, which may result in quality variability in scenarios that require changing some of the subject's properties, such as changes in color or shape. **(b)** Given that the training data is predominantly composed of real photographs, the model occasionally generates photos of paintings when the prompt specifies an artistic style.

Figure 15. Additional object insertion comparisons on our benchmark with the provided ground truth.

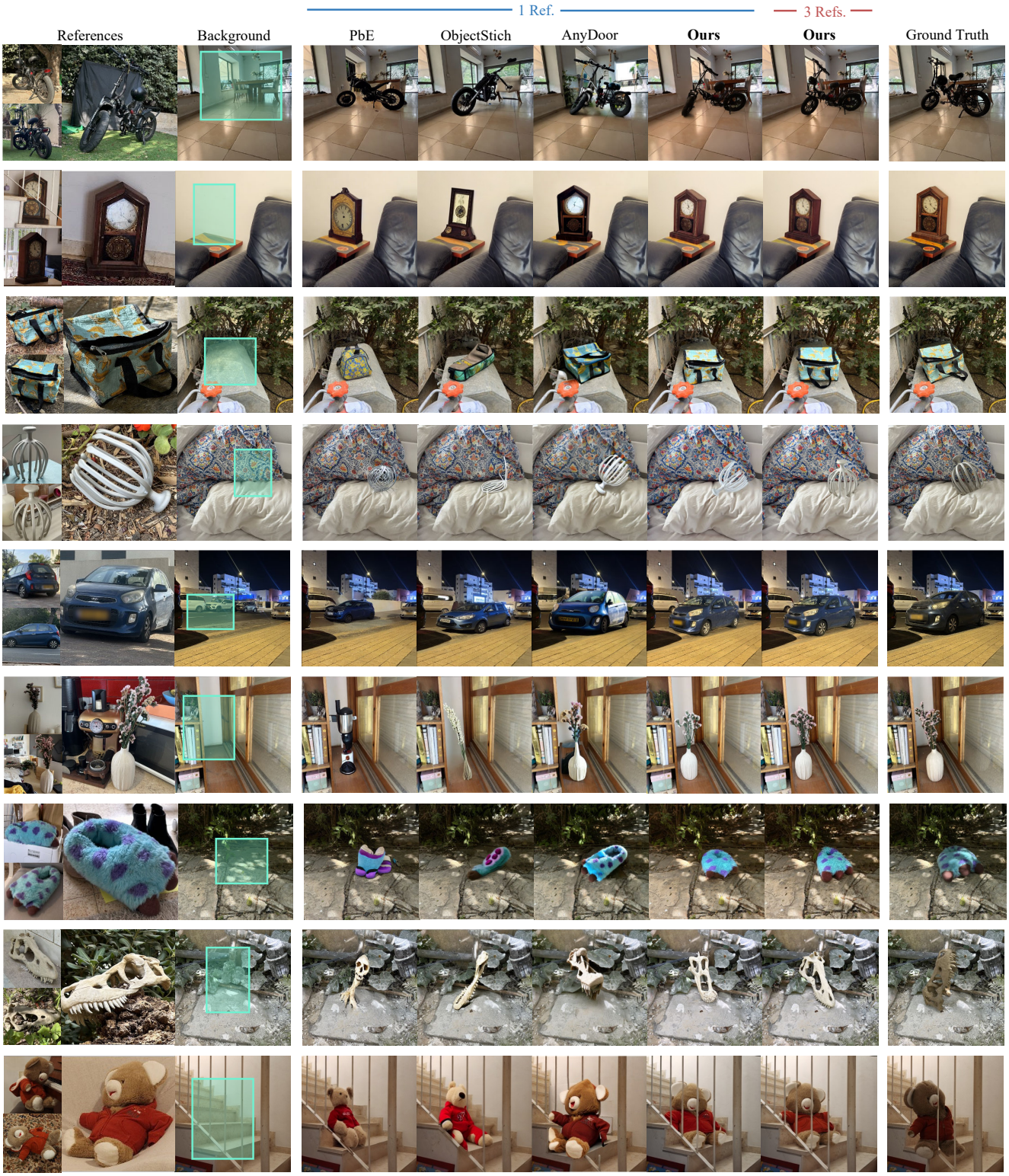

Figure 16. Additional object insertion comparisons on our benchmark with the provided ground truth.

Figure 17. Additional subject-driven generation comparisons.

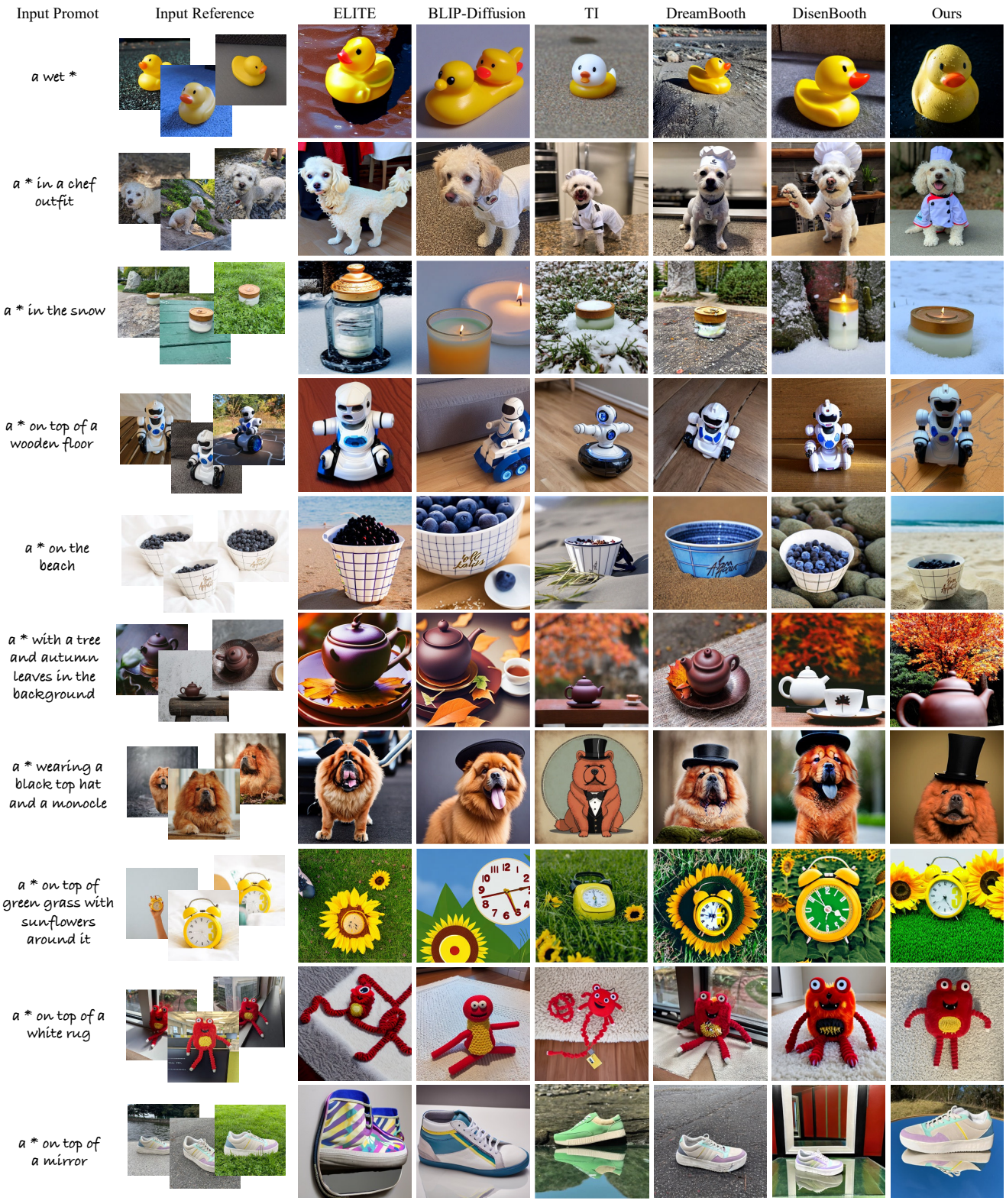

Figure 18. Additional subject-driven generation comparisons.