# BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing

**Anonymous Author(s)**
Affiliation
Address
email

# A Appendix

## A.1 Broader Impact

Image generation models are susceptible to be used as tools for generating false content or prompting misinformation. Subject-driven generation could be misused as a tool for generating fake image of individuals. To mitigate this issue, our model has been trained on generic objects where person-related subjects have been purposely removed from the training data. This makes the model weaker at generating fake images using person as subject control.

Our model is built using the pre-trained Stable Diffusion model trained on web-scraped datasets. Therefore, our model inherits some of its shortcomings, such as generating biased contents with social stereotypes, or other NSFW contents if used inappropriately. Our model's ability to precisely control the generation subject can help mitigate certain biases. We can use NSFW detectors to block potential inappropriate content from being generated. Nevertheless, we strongly caution against using our model directly in user-facing applications without a careful inspection of the model's output. Proper content moderation and regulation are highly advised to prevent undesirable consequence.

## A.2 Failure Cases Analysis

In Figure 1, we outline common failure cases of the model. Our model suffers from issues observed for prior subject-driven generation models as outlined in [1], including incorrect context synthesis, overfitting to training set. In addition, it subsumes some weakness of the underlying diffusion model, such as failing to address text prompts or generating fine-grained composition relations.

(a) Context-appearance entanglement | (b) Failing to address the text prompt. | (c) Incorrect spatial composition.

Subject images | A backpack on top of a **purple rug** in a forest. | Subject images | A **cube-shaped** bear plushie. | Subject images | A sneaker **on top of** a mirror.

Figure 1: Example failure generations. Subject images used for finetuning are shown on the left.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

### A.3 Competing Methods

We compare BLIP-Diffusion with fine-tuning based [2, 1] and retrieval-augmented [3] subject-driven generation models on the public DreamBench dataset [1]. We also compare qualitatively with the image editing method InstructPix2Pix [4]. We briefly introduce these methods below.

- *Textual Inversion* [2]: a fine-tuning method which optimizes a placeholder embedding to reconstruct the training set of subject images. It requires 3,000 training steps for learning a new concept, which takes around 30 minutes on an A100 GPU.
- *DreamBooth* [1]: a fine-tuning method similar to textual inversion. In addition to the placeholder embedding, it also optimizes parameters of the U-Net for a total budget of around 800 steps. We report intermediate results using 100 and 300 fine-tuning steps, while refer to metrics reported by the authors for full model comparison. Fine-tuning DreamBooth on a new concept costs around 6 minutes on an A100 GPU.
- *Re-Imagen*: a retrieval-augmented model, which takes the subject images as references and attend to them to generate new images. While the model requires no tuning, it significantly underperforms other models. The model is not publicly available, thus we do not have access to qualitative examples for comparison.
- *InstructPix2Pix*: an image editing model, which takes as input the source image and an editing instruction to generate edited images. Although it does not represent explicitly subjects, it can be used for applications such as subject re-contextualization and property modification. Therefore, we also include it for qualitative comparison. In particular, we experiment with both low (1.0) and high (1.5) image guidance scales, where a low image guidance scale preserves less the subject while promotes the text alignment; a high image guidance scale preserves better the original image yet is more likely to overlook the editing instruction.

### A.4 Evaluation Metrics

We adopt metrics proposed in DreamBooth [1] for evaluation, including DINO, CLIP-I and CLIP-T scores. Among them, DINO and CLIP-I scores are used to measure subject fidelity and CLIP-T is used to measure image-text alignment. DINO score is the average pairwise cosine similaity between the ViT-S/16 DINO embeddings of the generated and real images. CLIP-I score is the average pairwise CLIP ViT-B/32 image embeddings of the generated and real images. It is considered that DINO score is the preferred metric for measuring subject fidelity as it is sensitive to the differences between subjects of the same class. CLIP-T score is the average cosine similarity between prompt and image CLIP embeddings.

To better evaluate and compare subject-driven text-to-image models, it is suggested that these metrics should be considered jointly to avoid biased conclusion. For example, a model that naively copies the training set images will produce high DINO and CLIP-I scores with low CLIP-T scores. In the other case, a vanilla text-to-image generation model without subject knowledge, *e.g.* stable diffusion, will produce high CLIP-T scores with poor subject alignment. Both models are not considered desirable for the subject-driven text-to-image generation task.

### A.5 Pre-training Datasets

For multimodal representation learning, we use the same pre-training data as by BLIP-2, totaling 129M images. This includes COCO [5], Visual Genome [6], CC3M [7], CC12M [8], SBU [9] and 115M images from LAION400M [10]. We also employ the synthetic captions created using CapFilt method [11] for web images. We refer interested readers to Section 3.4 in the BLIP-2 paper [12] for details of the data bootstrapping configurations.

For subject representation learning, we use a subset of OpenImage-V6. We filter the data using the annotations provided by the dataset. In particular, we discard a sample if it satisfies one of the following cases: (i) a group of objects of the same class appear in the image; (ii) the image is taken from inside of the subject; (iii) the object is of aspect ratios larger than 2; (iv) objects occupy a too large (0.8) or too small (0.3) area relative to the image; (v) human-related subject, including boy, girl, person, man, mammal, woman, human body, human head, human hair, human arm, human face, human leg, human hand, human foot, human eye, human mouth, human nose, human ear, clothing, suit; (vi) cluttered objects, including tree, plant, houseplant, desk, table, poster and billboard. This results in 292K images for subject representation learning.

### A.6 Fine-tuning, Inference and Evaluation on DreamBooth Dataset

For all fine-tuning experiments, we use AdamW [13] optimizer with constant learning rate 5e-6 and no warm-up steps. We use batch size 3, adam beta1 0.9, adam beta2 0.999, adam epsilon 1e-8 and weight decay 0.01. We fine-tune models on a single A100 (40Gb) GPU and select checkpoints manually based on a set of validation prompts. We report the number of iterations for each subject on DreamBench below, on average 76 steps, taking around 40 seconds to complete on a single A100.

For inference, we use PNDM scheduler [14] for 100 denoising steps. We use a fixed guidance scale 7.5 for all experiments.

Table 1: Number of fine-tuning steps for DreamBench subjects.

| | | | | | |
|---|---|---|---|---|---|
| backpack | 110 | backpack-dog | 110 | bear-plushie | 110 |
| bowl | 40 | can | 70 | candle | 80 |
| cat | 40 | cat2 | 50 | clock | 120 |
| colorful-sneaker | 80 | dog | 50 | dog2 | 50 |
| dog3 | 40 | dog5 | 20 | dog6 | 40 |
| dog7 | 50 | dog8 | 40 | duck-toy | 60 |
| fancy-boot | 50 | grey-sloth-plushie | 70 | monster-toy | 120 |
| pink-sunglasses | 90 | poop-emoji | 90 | rc-car | 120 |
| red-cartoon | 70 | robot-toy | 110 | shiny-sneaker | 80 |
| teapot | 120 | vase | 120 | wolf-plushie | 80 |

Table 2: Average metrics for each subject on DreamBench in zero-shot setup.

| **Subject** | backpack | backpack-dog | bear-plushie | berry-bowl | can | candle |
|---|---|---|---|---|---|---|
| **DINO** | 0.452 | 0.467 | 0.634 | 0.750 | 0.540 | 0.395 |
| **CLIP-I** | 0.782 | 0.712 | 0.739 | 0.792 | 0.641 | 0.710 |
| **CLIP-T** | 0.320 | 0.310 | 0.304 | 0.257 | 0.314 | 0.316 |
| **Subject** | cat | cat2 | clock | colorful-sneaker | dog | dog2 |
| **DINO** | 0.760 | 0.703 | 0.402 | 0.680 | 0.780 | 0.730 |
| **CLIP-I** | 0.835 | 0.854 | 0.735 | 0.769 | 0.849 | 0.831 |
| **CLIP-T** | 0.306 | 0.286 | 0.303 | 0.298 | 0.310 | 0.307 |
| **Subject** | dog3 | dog5 | dog6 | dog7 | dog8 | duck-toy |
| **DINO** | 0.558 | 0.705 | 0.763 | 0.656 | 0.641 | 0.665 |
| **CLIP-I** | 0.747 | 0.788 | 0.867 | 0.817 | 0.816 | 0.840 |
| **CLIP-T** | 0.310 | 0.313 | 0.288 | 0.309 | 0.307 | 0.287 |
| **Subject** | fancy-boot | grey-sloth-plushie | monster-toy | pink-sunglasses | poop-emoji | rc-car |
| **DINO** | 0.538 | 0.632 | 0.490 | 0.599 | 0.494 | 0.569 |
| **CLIP-I** | 0.800 | 0.755 | 0.734 | 0.836 | 0.689 | 0.761 |
| **CLIP-T** | 0.291 | 0.315 | 0.293 | 0.308 | 0.307 | 0.281 |
| **Subject** | red-cartoon | robot-toy | shiny-sneaker | teapot | vase | wolf-plushie |
| **DINO** | 0.697 | 0.534 | 0.668 | 0.451 | 0.471 | 0.463 |
| **CLIP-I** | 0.826 | 0.787 | 0.759 | 0.804 | 0.786 | 0.737 |
| **CLIP-T** | 0.263 | 0.315 | 0.294 | 0.314 | 0.262 | 0.327 |

In Table 2 and 3, we report average metrics across 10 experiment runs for each subject in the dataset, in zero-shot and fine-tuning setups, respectively.

Table 3: Average metrics for each subject on DreamBench in fine-tuning setup.

| **Subject** | backpack | backpack-dog | bear-plushie | berry-bowl | can | candle |
|---|---|---|---|---|---|---|
| **DINO** | 0.551 | 0.639 | 0.693 | 0.808 | 0.618 | 0.519 |
| **CLIP-I** | 0.839 | 0.760 | 0.752 | 0.829 | 0.695 | 0.752 |
| **CLIP-T** | 0.320 | 0.317 | 0.307 | 0.254 | 0.313 | 0.311 |
| **Subject** | cat | cat2 | clock | colorful-sneaker | dog | dog2 |
| **DINO** | 0.806 | 0.747 | 0.479 | 0.739 | 0.821 | 0.793 |
| **CLIP-I** | 0.869 | 0.864 | 0.784 | 0.805 | 0.860 | 0.841 |
| **CLIP-T** | 0.306 | 0.284 | 0.305 | 0.320 | 0.313 | 0.307 |
| **Subject** | dog3 | dog5 | dog6 | dog7 | dog8 | duck-toy |
| **DINO** | 0.573 | 0.727 | 0.834 | 0.672 | 0.723 | 0.699 |
| **CLIP-I** | 0.751 | 0.801 | 0.891 | 0.823 | 0.823 | 0.838 |
| **CLIP-T** | 0.312 | 0.311 | 0.280 | 0.310 | 0.310 | 0.284 |
| **Subject** | fancy-boot | grey-sloth-plushie | monster-toy | pink-sunglasses | poop-emoji | rc-car |
| **DINO** | 0.649 | 0.717 | 0.566 | 0.625 | 0.627 | 0.651 |
| **CLIP-I** | 0.827 | 0.780 | 0.743 | 0.826 | 0.784 | 0.775 |
| **CLIP-T** | 0.299 | 0.322 | 0.292 | 0.312 | 0.290 | 0.288 |
| **Subject** | red-cartoon | robot-toy | shiny-sneaker | teapot | vase | wolf-plushie |
| **DINO** | 0.788 | 0.626 | 0.757 | 0.484 | 0.628 | 0.599 |
| **CLIP-I** | 0.882 | 0.803 | 0.804 | 0.819 | 0.812 | 0.760 |
| **CLIP-T** | 0.262 | 0.316 | 0.297 | 0.331 | 0.261 | 0.325 |

### A.7 Zero-shot Subject-driven Image Manipulation

Our model is able to extract subject features to guide the generation. In addition to applications of subject-driven generations and editing, we show that such pre-trained subject representation enables intriguing and useful applications of zero-shot image manipulation, including subject interpolation and subject-driven style transfer.

*Subject Interpolation.* It is also possible to blend two subject representation to generate subjects with a hybrid appearance. This can be achieved by traversing the embedding trajectory between subjects. In Figure 2, we create bilinear interpolations among 4 different subject representations, and render the interpolated subject in a novel context. As the figure shows, the subject appearance blends along the trajectory and fits naturally with the environment. This is useful when multiple subjects are used as reference to guide the generation. For example, subject interpolation can be used in joint with subject-driven style transfer to create hybrid style from multiple guiding subjects.

*Subject-driven Style Transfer.* When provided with a subject, the model can encode the appearance style of it and transfer to other subjects. We refer such an application as subject-driven style transfer. In Figure 3 and 4, we generate stylized reference subjects with the aid of edge-guided ControlNet. The styles are hinted by the guiding subjects. Specifically, we feed BLIP-2 with guiding subjects and their category texts, *e.g.* fire, flower, glass, vase, ball, bread, to extract the subject representation. In this application, guiding subjects serve as alternative of textual prompts to specify styles. This is useful especially when a style is non-trivial to describe by natural languages accurately.

### A.8 Additional Qualitative Results and Subject Fidelity Showcasing

In Figure 5 to 7, we provide additional qualitative results on DreamBench subjects and prompts. We show the reference subject image in the first column. In the rest columns, we provide generated renditions. To showcase subject fidelity and photorealism, we purposely mix one genuine subject image in and leave for interested readers to figure out. Read the captions to verify.

Figure 2: Zero-shot subject interpolation. We interpolate subject representation and use the same denoising and decoder network for generation. The intermediate subject representation naturally blends the subject appearance, while fitting coherently into the new context.

# References


[1] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[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.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[5] 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.

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[7] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

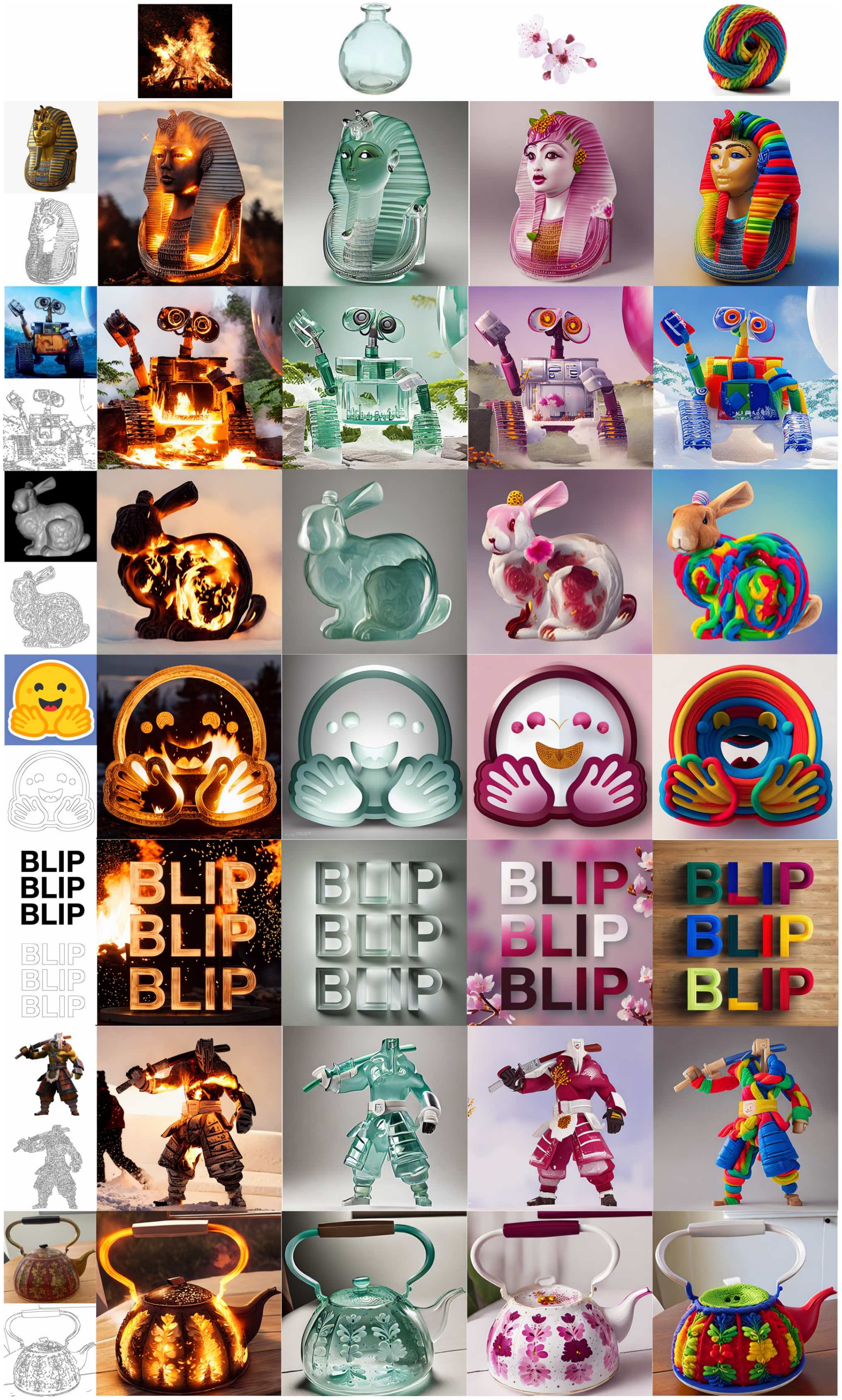


Figure 3: Zero-shot subject-driven stylization. We show guiding subject images on top. In the rest rows, we show reference subjects and their canny maps on left, and stylized reference subjects by column.

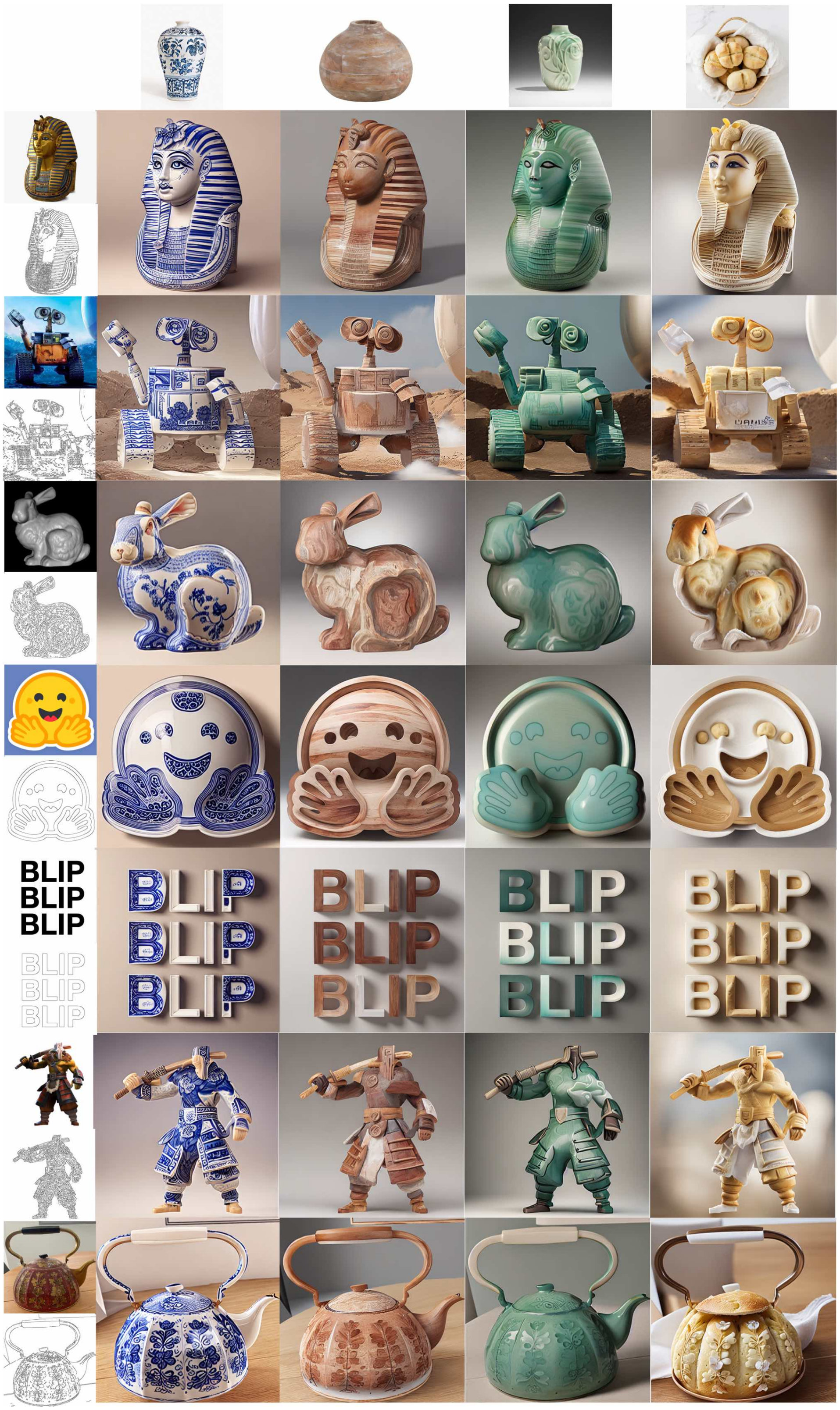


Figure 4: (Cont.) Zero-shot subject-driven stylization. We show guiding subject images on top. In the rest rows, we show reference subjects and their canny maps on left, and stylized reference subjects by column.

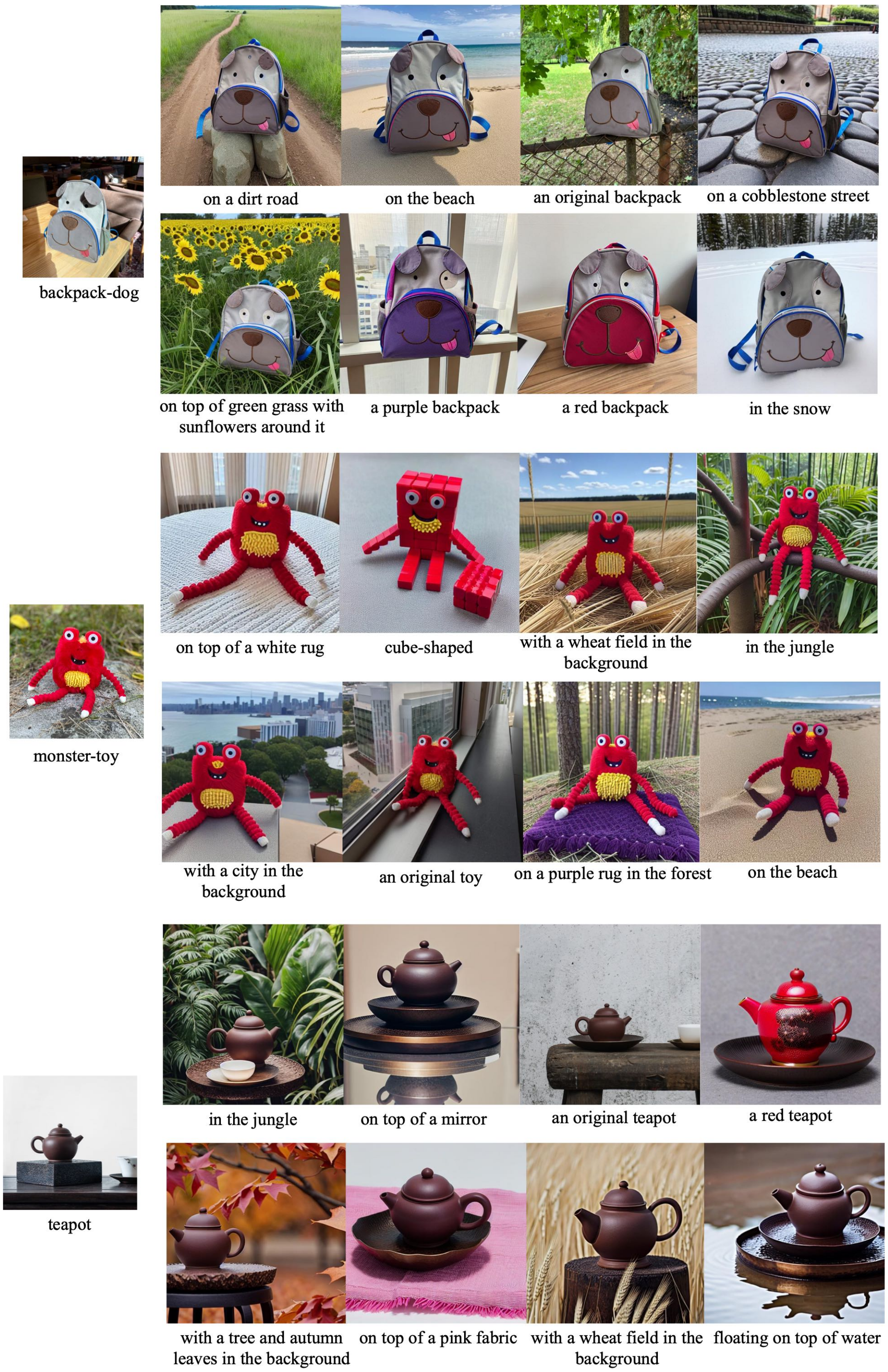


Figure 5: Additional qualitative results using DreamBench subjects and prompts. To showcase subject fidelity and photorealism, we mix one genuine subject image in the generations for readers to figure out. Zoom-in and read the captions to verify.

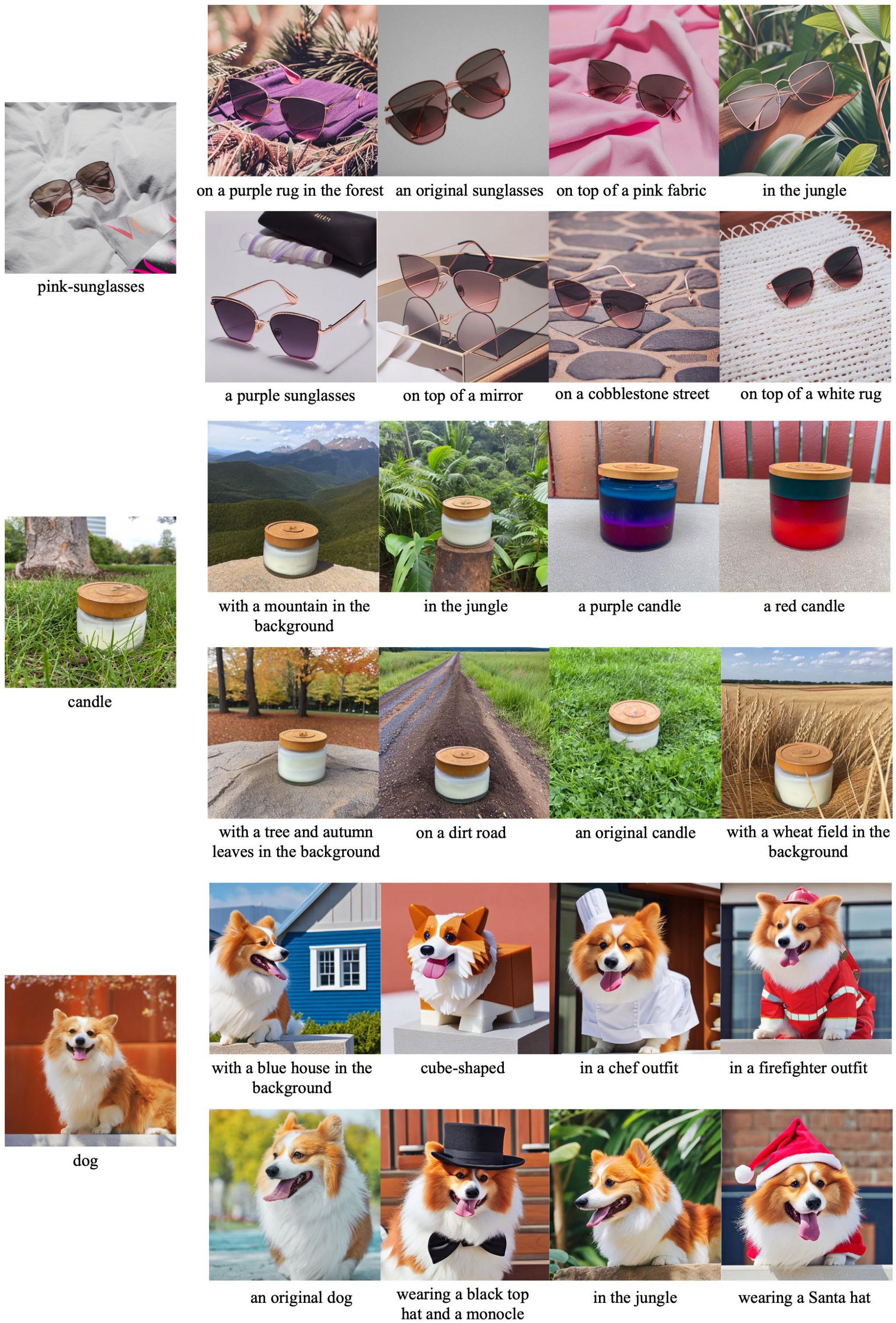


Figure 6: (Cont.) Additional qualitative results using DreamBench subjects and prompts. To showcase subject fidelity and photorealism, we mix one genuine subject image in the generations for readers to figure out. Zoom-in and read the captions to verify.


Figure 7: (Cont.) Additional qualitative results using DreamBench subjects and prompts. To showcase subject fidelity and photorealism, we mix one genuine subject image in the generations for readers to figure out. Zoom-in and read the captions to verify.

[9] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[10] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[14] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022.