

Supplementary Materials: DreamBooth++: Boosting Subject-Driven Generation via Region-Level References Packing

Anonymous Authors

A MORE QUALITATIVE RESULTS

We present an extended visual analysis comparing DreamBooth++ with the established baseline methods using the DreamBooth dataset. Our results, displayed in Figure 1, illustrate that DreamBooth++ is capable of generating images that not only preserve the identity of the subjects but are also photorealistic and maintain high fidelity across various contexts. This superior performance underscores the enhancements integrated into DreamBooth++, making it a robust solution for subject-driven image generation.

Further demonstrating the versatility and robustness of our model, we provide additional results from the CustomConcept101 dataset [1] in Figure 2 and Figure 3. This dataset is notably diverse, featuring 101 different objects, each associated with a unique set of prompts. The results of CustomConcept101 dataset highlight DreamBooth++’s ability to generalize well across varied subjects and prompts, showcasing its effectiveness in adapting to different artistic styles and contextual demands.

B DETAILED ABLATION STUDY

We incorporate various layout of packed reference images into our training process by randomizing the position and number of random anchors mentioned in the main text. Some of the packed examples can be shown in the Figure 4. In our experiments, we explored the effects of varying the proportions of reference images packed into different layouts during the training process. This involved adjusting the distribution of original images, 2x2, and 3x3 packed configurations. Specifically, we tested ratios of 1:1:1, 1:2:2, 1:3:3, 1:5:5, and 1:7:7, to examine how these configurations influence the model’s performance in terms of subject fidelity, prompt fidelity, and image diversity. The results, illustrated in Table 1, demonstrate how different packing strategies can optimize the training efficiency and effectiveness of our DreamBooth++ model.

different proportions of packed examples. Our study explores how different ratios of packed examples influence model performance, particularly focusing on subject fidelity, prompt fidelity, and image diversity. The experiments vary the proportions of original, 2x2, and 3x3 packed images, assessing the effects on key metrics. As depicted in Table 1, increasing the proportion of packed images generally improves subject fidelity. The highest CLIP-I score is observed at a 1:3:3 ratio, indicating better subject representation, whereas the CLIP-T scores suggest that prompt fidelity does not uniformly benefit from higher proportions of packed configurations. These findings demonstrate the trade-offs between enhancing subject detail and maintaining prompt coherence as packing density increases.

different strength of Regularization. We investigated the effect of different strengths of text-guided prior regularization by adjusting the parameter λ . This study aims to understand how changes in λ

Table 1: Quantitative comparison of subject fidelity (DINO), prompt fidelity (CLIP-I, CLIP-T), and diversity (LPIPS) across different proportions of packed examples. Proportions are described as the ratio of original images to 2x2 and 3x3 packed configurations (e.g., 1:1:1 represents an equal number of original, 2x2 packed, and 3x3 packed images).

Proportion	DINO↑	CLIP-I↑	CLIP-T↑	LPIPS↑
1:1:1	0.674	0.806	0.246	0.756
1:2:2	0.695	0.817	0.242	0.745
1:3:3	0.673	0.826	0.242	0.750
1:5:5	0.681	0.812	0.243	0.756
1:7:7	0.675	0.809	0.242	0.755

Table 2: Impact of varying the strength of text-guided prior regularization on subject fidelity (DINO), prompt fidelity (CLIP-I, CLIP-T), and image diversity (LPIPS). The parameter λ controls the regularization strength.

λ	DINO↑	CLIP-I↑	CLIP-T↑	LPIPS↑
0	0.686	0.833	0.229	0.735
10	0.673	0.826	0.242	0.750
50	0.626	0.790	0.253	0.779
100	0.597	0.763	0.255	0.787

affect the semantic consistency and subject fidelity of generated images. As shown in Table 2, increasing λ generally enhances prompt fidelity (CLIP-T) at the cost of reduced subject fidelity (DINO and CLIP-I). This balancing of visual and textual accuracy highlights the nuanced role of λ in model’s performance.

C LIMITATION

Despite DreamBooth++ achieving promising results, it exhibits certain limitations, which we detail through failure cases shown in Figure 5. The first issue involves objects being truncated at the edges of the image. This occurs when some reference images used in the references packing process have extreme aspect ratios, making it challenging to avoid cutting off parts of the subject. The second limitation concerns the failure in complex semantic parsing, which includes semantic coupling and ignoring specified semantics in prompts. These issues may arise due to insufficiently robust priors for these contexts or the inherent difficulty of generating images that align both the subject and a rarely co-occurring specified concept. To mitigate these issues, we tested DreamBooth++ with the more powerful base model, Stable Diffusion XL [2], as shown in Figure 7, which demonstrated improved performance in managing complex semantic tasks.

Training samples

DB++

DB

OFT

LoRA

on a
cobblestone
street

with the
Eiffel Tower

with a tree
and autumn
leaves

with the
Eiffel Tower

on a
cobblestone
street

on the
beach

cube
shaped

floating on
top of water

purple



Figure 1: More qualitative results on DreamBooth dataset.

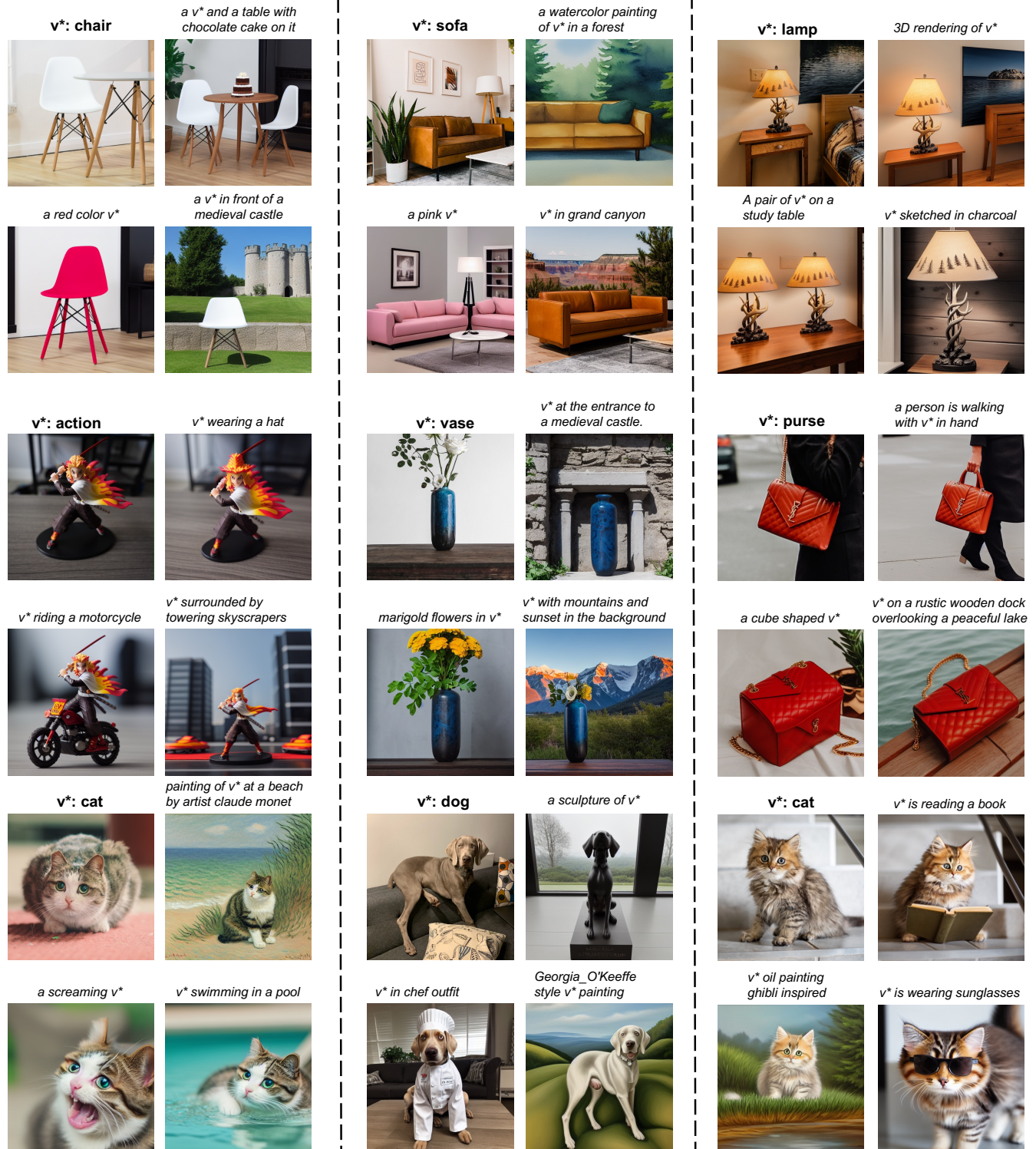


Figure 2: DreamBooth++ Results on CustomConcept101 Dataset [1].

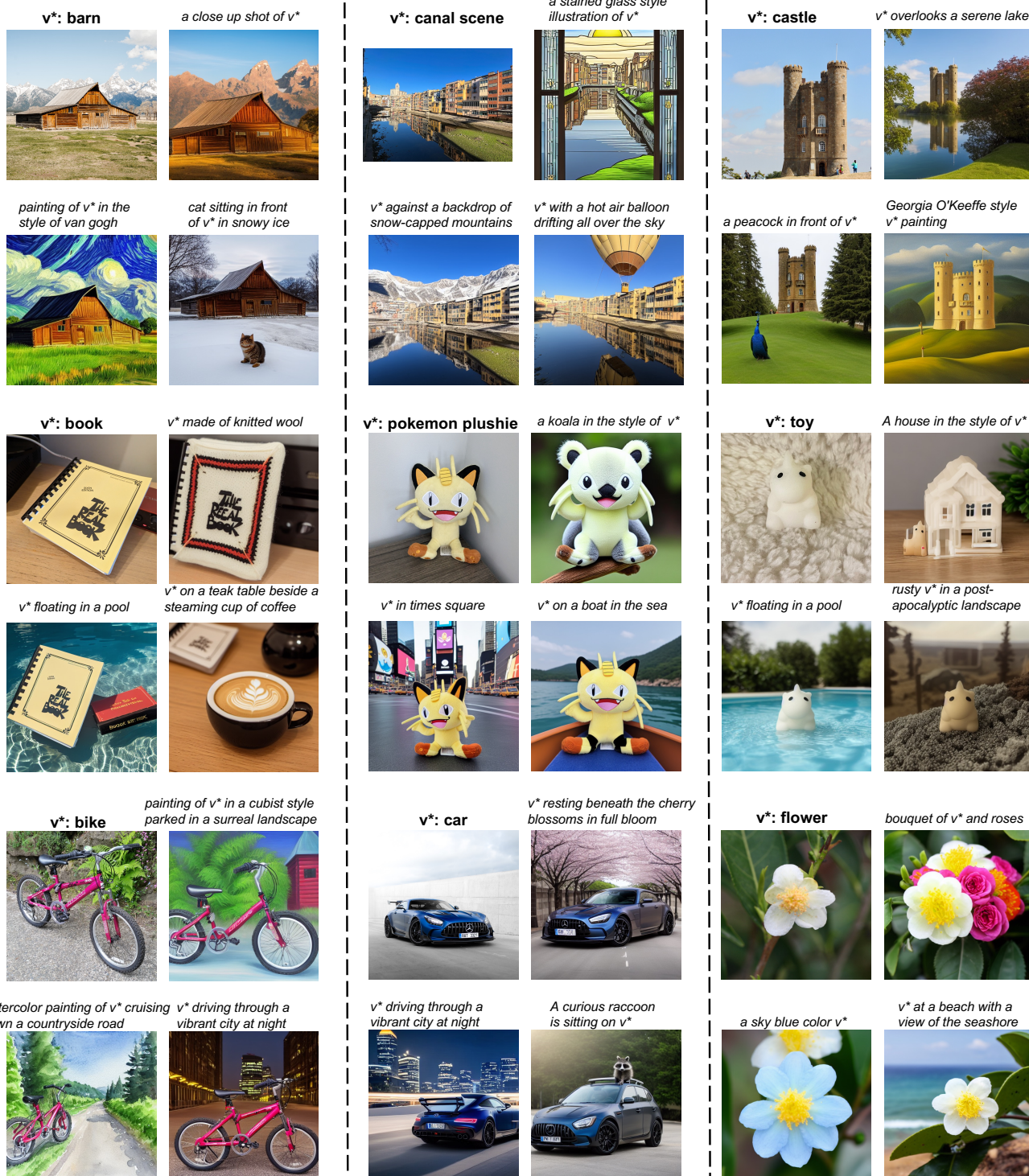


Figure 3: Continued Results on CustomConcept101 Dataset [1].

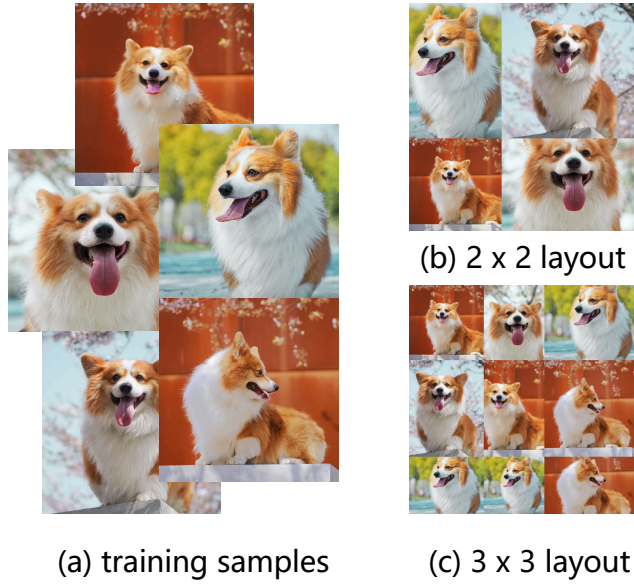


Figure 4: Examples of our Data Re-formulation with diverse layout configurations.



Figure 5: Failure modes in DreamBooth++. (a) shows the truncation of subjects due to extreme aspect ratios in reference images, and (b) depicts challenges in parsing complex semantics, such as merging unrelated elements or overlooking specific prompt details.

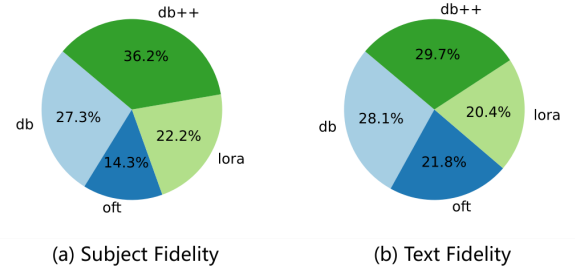


Figure 6: User Study Results. Pie charts illustrating participant preferences in (a) Subject Fidelity and (b) Text Fidelity for images generated by DreamBooth++ (db++), DreamBooth (db), LoRA, and OFT. DreamBooth++ leads in both categories, reflecting its enhanced ability to produce images that are both visually accurate and textually coherent.

D INTEGRATION WITH STABLE DIFFUSION XL

In this section, we extended our experiments to include DreamBooth++ implementation on a more capable base model—Stable Diffusion XL [2], which has shown improvements in managing complex semantic tasks. The integration of DreamBooth++ with the Stable Diffusion XL model demonstrates significant enhancements in high-resolution image generation. This combination not only addresses the limitations identified in our standard model configuration but also excels in producing images with greater detail and clarity. Figure 7 exhibits DreamBooth++’s enhanced capabilities when applied to better performing base model, underscoring its adaptability and effectiveness for advanced image generation tasks.

E USER STUDY

We conducted a user study to evaluate DreamBooth++ compared to DreamBooth (DB), LoRA, and OFT, focusing on subject fidelity and text fidelity as described in the original DreamBooth study [3]. Each pair, randomly selected from the comparison models. Participants were asked to choose which image better preserved the subject’s identity and which image better adhered to the text prompt, or if they were indistinguishable. The results were tallied to determine how often images from each model were preferred. These outcomes are displayed in Figure 6. It is noteworthy that OFT excelled in objective metrics but was less favored in our subjective evaluations. This may suggest that while OFT effectively prevents overfitting, it could be at the cost of capturing finer details that users value in visual comparisons. This result clearly showing that DreamBooth++ was frequently chosen as producing superior results, highlighting its effectiveness in producing high-fidelity, contextually accurate images in real-world applications.

REFERENCES

- [1] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 1931–1941.
- [2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion

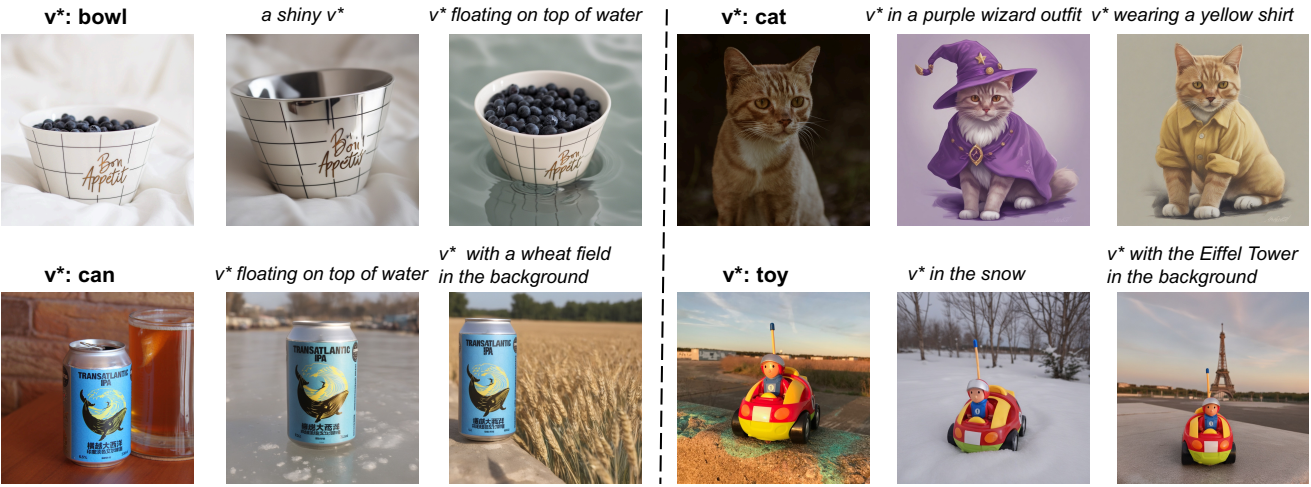


Figure 7: Enhanced Image Generation with DreamBooth++ on Stable Diffusion XL.

Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.