DreamCache: Finetuning-Free Lightweight Personalized Image Generation via Feature Caching

Supplementary Material

Table S1. **Masked metrics** quantitative evaluation.

Method	MCLIP-I (↑)	MDINO (↑)
DreamBooth	0.868	0.712
Custom Diffusion	0.864	0.711
JeDI	0.876	0.751
BLIP-D	0.862	0.669
ELITE	0.861	0.681
Toffee-5M	0.874	0.803
Ours	**0.906**	**0.837**

S1. Synthetic Dataset

In this section, we describe our dataset generation pipeline, inspired by the success of BootPIG, with some modifications to ensure the pipeline adopts open-source models and is fully reproducible. Figure S1 provides an overview of the data creation process. We also show some examples of generated synthetic data in Figure S2.

We use the *lang-sam* pipeline[1] to segment both generated and reference images based on textual conditioning, using a combination of Grounding-DINO and SAM. For caption generation, we leverage the LLama 3.2 8B [2], with a carefully crafted prompt that aims to generate diverse and descriptive captions of concrete objects, placing them in various meaningful contexts. We filter the generated captions to ensure the dataset's diversity and remove duplicates or highly similar captions. We write a simple filtering script that counts the number of occurrences for each object/category and filter out redundant captions.

The filtered captions are then used to prompt SD-XL [5] with a Classifier-Free Guidance (CFG) scale of 3.5, employing 25 denoising steps to generate the images. Our entire data generation pipeline is reproducible, and we plan to release it alongside the code for DreamCache. Additionally, we will provide access to our generated dataset to encourage further research in this area.

S2. Additional Evaluations

S2.1. Masked Metrics

Recent studies [8, 9] emphasize the value of evaluating masked versions of image similarity metrics to eliminate potential interference from background elements, thus ensuring the evaluation focuses on the fidelity of the personalized object. We use Grounded-SAM [7] to segment both generated and reference images, subsequently computing the CLIP-I and DINO scores for these segments. The results for these masked metrics are reported in Table S1. DreamCache achieves a higher score on both metrics, demonstrating its superiority in subject preservation.

S2.2. Qualitative Results

In this section, we present additional qualitative generations produced by DreamCache (Figure S3). We conduct experiments using both synthetically generated subjects and real subjects from the Dreambooth dataset. Our results demonstrate that our method effectively follows complex text prompts. Interestingly, despite the absence of explicit training for subject modification (as seen in editing datasets), our approach successfully adapts and transforms the input subject in various contexts, rather than simply replicating the reference.

Pose and style variations To further illustrate the capability of our method in achieving substantial subject transformations, particularly in pose and style variation. Our model successfully modulates the cached features via textual prompts, enabling significant variations beyond mere content replication. Notably, as shown in Fig. S4 the dog reference subject was prompted into diverse poses, clearly showcasing that our approach avoids the "copy-paste" effect. The chair is transformed into a Van Gogh-stylized version, the dragon is stylized as Chinese painting, the elephant changes pose while dressed as a wizard, and the guitar can be transformed to ice. These examples underscore that the injected features are effectively guided and transformed by text instructions, enabling control over subject characteristics such as pose, appearance, and stylistic attributes.

Additional Qualitative Comparisons We also provide additional qualitative comparisons in Figure S4, including two reproducible open-source baselines: BLIP-D [3] and Kosmos-G [4].

S3. Additional Ablation Study

Impact of Encoding Timestep t The proposed reference encoding mechanism relies on selecting $t = 1$ as a fixed timestep during the encoding process. We validate this design choice in Table S3, showing that $t = 1$ yields the best performance. This finding aligns with the intuition that

[1] https://github.com/luca-medeiros/lang-segment-anything

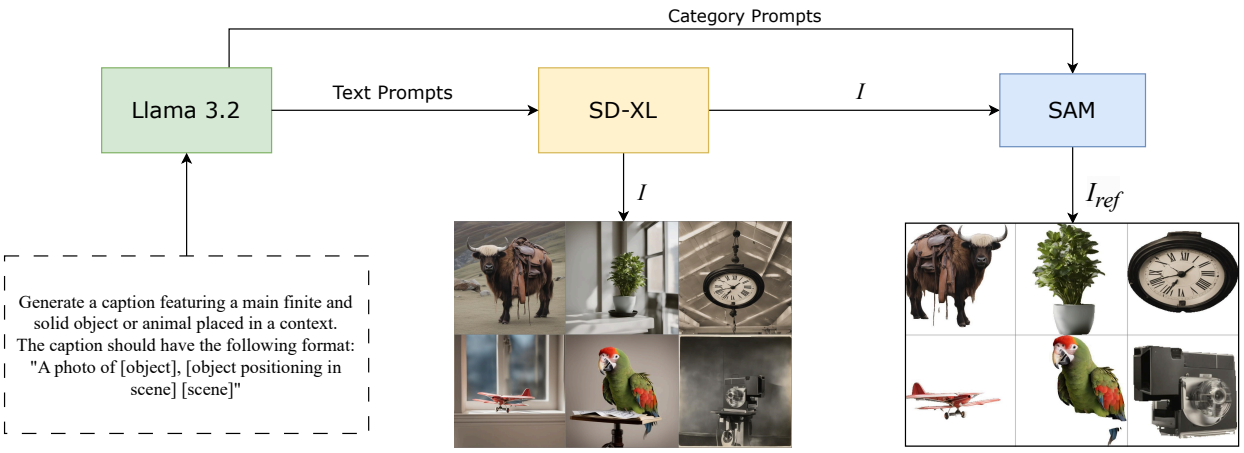

Figure S1. **Data synthesis pipeline** inspired by BootPIG [6].

Table S2. **Reference features** ablation study.

# Reference Features	CLIP-I (↑)	CLIP-T (↑)
Middle Features	0.778	0.312
Respective Features	0.810	0.298

Table S3. **Encoding timestep** ablation study.

Timestep	CLIP-I (↑)	CLIP-T (↑)
1	0.810	0.298
150	0.800	0.299
300	0.789	0.301

Table S4. **Decoder Layers** ablation study.

# Selected Layers	CLIP-I (↑)	CLIP-T (↑)
Every	0.811	0.296
Every second	0.810	0.298

Table S5. **Fixing the Reference U-net** ablation study

# Trained	CLIP-I (↑)	CLIP-T (↑)
✓	0.813	0.297
✗	0.810	0.298

less noisy features provide a more informative conditioning signal. Furthermore, this experiment highlights a significant limitation of reference U-Net-based methods that inject noisy features corresponding to different timesteps. These noisy features are less informative and contain fewer details compared to the low-noise, fixed-timestep references we use to condition the generation independently of the current timestep.

Impact of Multi-Resolution Features We also investigate the necessity of multi-resolution features for DreamCache's performance. In a variant of our method, we fixed the cached features to a single resolution (i.e., the bottleneck resolution of the U-Net,(8×8), after the encoding stage). Our experiments demonstrate that leveraging multiple resolutions significantly enhances performance compared to using a single fixed-resolution cached feature map, as shown in Table S2.

Impact of the number of layers in the decoder We also performed an experiment in which we insert our conditioning adapters in every decoder layer instead of every two layers. The results shown in Table S4, demonstrate that our choice is optimal, since inserting the adapters into all decoder layers brings negligible improvements while almost doubling the parameter count.

Impact of freezing the reference U-net In this ablation study we demonstrate that training the reference U-Net (like BootPIG) significantly increases parameters but brings negligible benefits, as shown in Table S5. This further justify our design choice.

S4. Sampling Space and Image Guidance

In our experiments we follow prior works [1, 8] and experiment with different types of guidance for image and text conditioning signal. The first and simpler *joint guidance* approach jointly drops text and image conditioning for the

Figure S2. **Synthetic dataset samples** generated via the process outlined in Fig. S1.

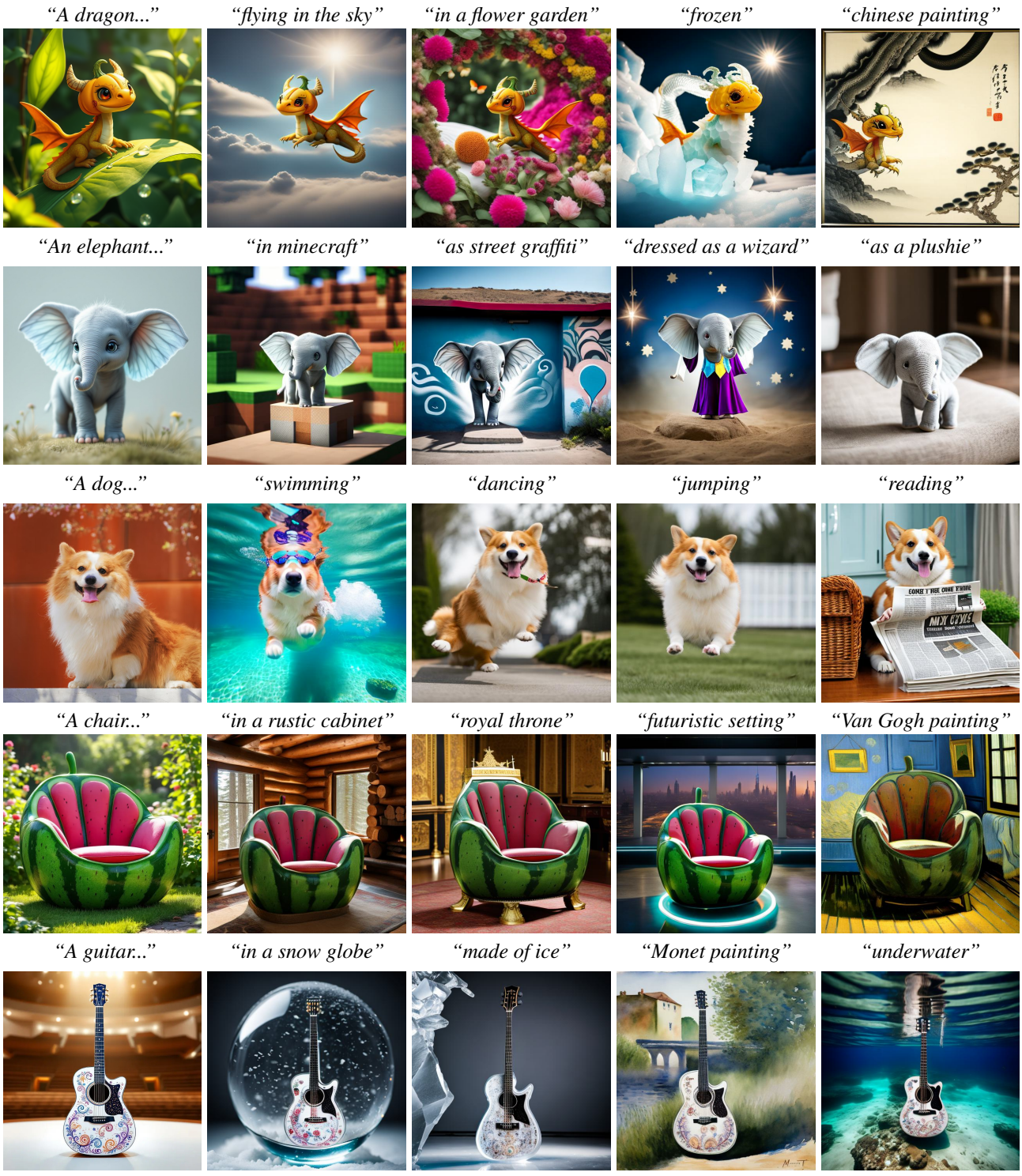

Figure S3. **Personalized generations by DreamCache.** The proposed method is able to adapt to different text prompts and leverage diffusion prior to perform appearance and style editing of the personalized content. We also notice how the background interference is completely absent in generated images due to our design choice of caching masked reference features.

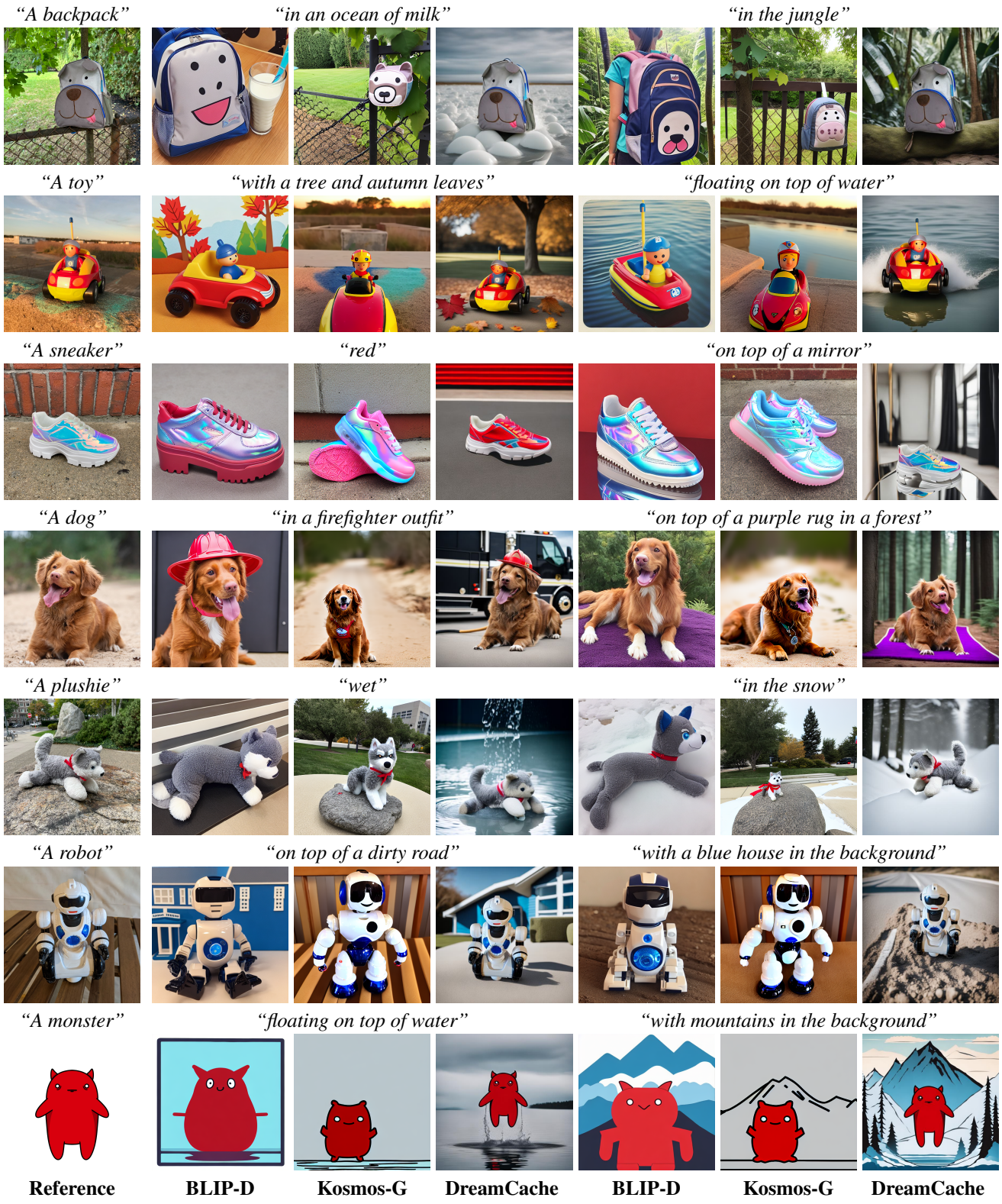

Figure S4. **Visual comparison.** Personalized generations on sample concepts. DreamCache preserves reference concept appearance and does not suffer from background interference. BLIP-D [3] and Kosmos-G [4] cannot faithfully preserve visual details from the reference.

unconditional prediction:

$$\tilde{e_\theta}(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) + s \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, \varnothing, \varnothing))$$

Where $\tilde{e_\theta}(z_t, c_I, c_T)$ represents the adjusted prediction at denoising step t conditioned on textual conditioning c_T and the image conditioning c_I. $e_\theta(z_t, \varnothing, \varnothing)$ denotes the unconditional prediction, and s is the guidance scale. The second approach, that we call *combined guidance* decouples text and image allowing for a more flexible balance between the two conditioning modalities:

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned}$$

Our experimental findings suggest that using a higher image guidance scale better preserves the content of the reference image, but reduces editability of the subject. Conversely, decreasing image guidance results in more flexible editing of the reference subject at the expense of reduced subject fidelity. Figure S5 illustrates these findings on the DreamBooth dataset, comparing the joint and combined guidance strategies.

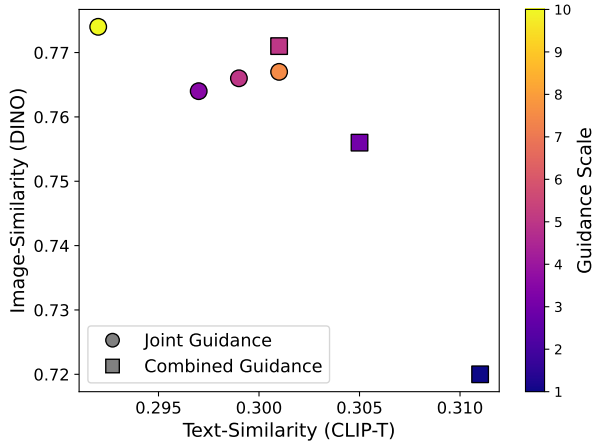

Figure S5. Sampling Space Exploration. For Combined Guidance, we leave the text scale $c_T = 7.5$ and we vary the image scale c_I.

S5. Broader Impact

DreamCache allows users to customize the subject of their images, focusing on individual elements such as animals or objects. However, it is crucial to recognize that, like other generative models and image editing tools, this technology has the potential to be misused for creating misleading content. Addressing these ethical risks is an essential and ongoing focus in the field of generative modeling, particularly in relation to deepfake creation. Techniques such as watermarking or content detection are particularly necessary to prevent misuse of this technology.

References

[1] 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. 2

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[3] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1, 5

[4] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 1, 5

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[6] Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. *arXiv preprint arXiv:2401.13974*, 2024. 2

[7] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 1

[8] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6786–6795, 2024. 1, 2

[9] Yufan Zhou, Ruiyi Zhang, Kaizhi Zheng, Nanxuan Zhao, Jiuxiang Gu, Zichao Wang, Xin Eric Wang, and Tong Sun. Toffee: Efficient million-scale dataset construction for subject-driven text-to-image generation. *arXiv preprint arXiv:2406.09305*, 2024. 1