Yo'Chameleon: Personalized Vision and Language Generation

Supplementary Material

1. Full-model Finetuning vs. Soft Prompt

As discussed in the Introduction, our experiment reveals that soft prompt tuning can match the performance of full-model fine-tuning when trained with approximately 300 real images of a single concept. In this section, we provide details about that experiments.

In this experiment, we collected photos for three concepts: one person (300 images), one dog (500 images), and one cat (500 images). These images are "in-the-wild" and therefore exhibit significant diversity in appearance. To address this, we first roughly cropped the regions containing the target concepts, creating datasets for each concept at a resolution of 512×512. The concepts of interest are typically centered within the images. Our goal was to verify whether soft prompt tuning could achieve performance comparable to full-model fine-tuning, which is commonly used in personalized image generation (i.e., [1–3]).

For full-model fine-tuning, we fine-tune Chameleon [4] using the prompt "A photo of <sks>" with a learning rate of 1×10^{-7}, a batch size of 2, over maximum 1000 iterations. For soft prompt tuning, we used the prompt "<sks> is $<\text{token}_1>...<\text{token}_{16}>$. A photo of <sks>." with a learning rate of 1×10^{-4}, batch size of 4, for 15 epochs. In another words, a concept is represented by $k = 16$ latent tokens.

To evaluate general abilities, we used prominent benchmarks such as MMLU [5] for text-only generation, POPE [6], and MMBench [7] for visual question answering. For personalized abilities, we measured CLIP-Image Similarity [8] and Facial Similarity using the off-the-shelf ArcFace model [9] to compare generated images with the reference images.

The results are shown in Tab. 1 (first five rows, "300+ real images"). As the table demonstrates, full-model fine-tuning leads to catastrophic forgetting, with performance degradation ranging from 1–65% across tasks. Although fine-tuning improves personalized image generation metrics (e.g., CLIP-Image Similarity increases from 0.804 to 0.849), it significantly compromises the model's general abilities, such as text-only generation, where MMLU performance drops from 65.4 to 59.6. In contrast, soft prompt tuning achieves comparable performance in personalized image generation (e.g., Facial Similarity reaches 0.429) while maintaining general abilities nearly identical to the base model.

It is important to note that this experiment was conducted for research purposes only and has limited practical applicability, as users might not be able, or not willing to provide 300+ images of a concept. Nonetheless, this pilot study effectively demonstrates the advantages of soft prompt tuning over full-model fine-tuning: (1) it matches the performance of full-model fine-tuning in personalized tasks and (2) mitigates catastrophic forgetting.

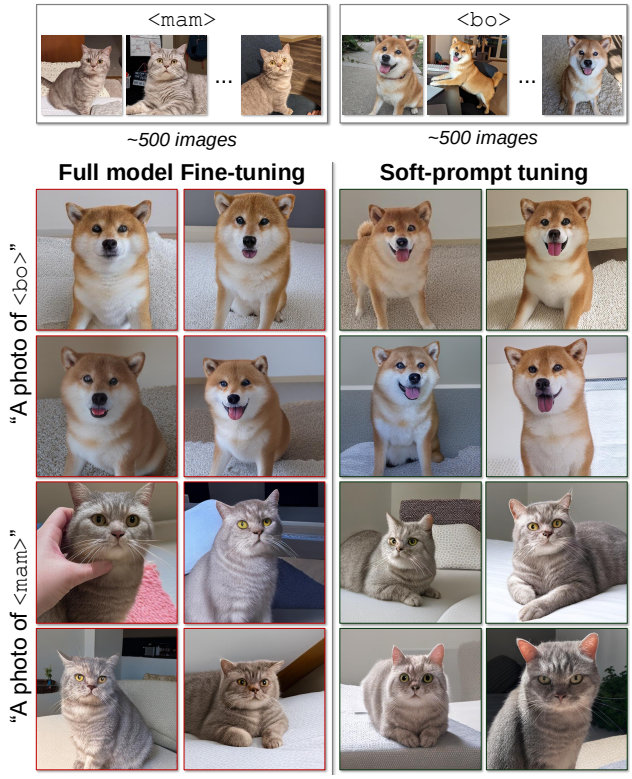

Figure 1. With 300+ real images, soft-prompt tuning can match the performance of full-model fine-tuning while retaining the model's overall abilities. We cannot show the facial results due to anonymity.

2. Additional Ablation Studies

Along with the ablation studies presented in the main paper, we provide an additional ablation study on (1) the number of "soft-positive" images and (2) Evaluation for Catastrophic Forgetting . These studies could not be included in the main paper due to space limitations.

2.1. Number of Soft-Positive Images

Similar to other ablation studiesin the main paper, this study aims to analyze the effect of varying the number of "soft-positive" images used during concept training. We vary the

Settings	General Abilities (↑)					Personalized Image Gen. (↑)	
	POPE [6]			MMBench [7]	MMLU [5]	CLIP-I [8]	Facial Sim [9]
	pop	rand	adv	en			
Random	*0.500*	*0.500*	*0.500*	*0.25*	*0.25*	∼0.3-0.5	∼ 0.001
Original Chameleon [4]	0.702	0.504	0.656	0.57	0.52	0.423	0.001
300+ real images (3 concepts)							
Soft Prompt (16 tokens)	0.702 (same)	0.504 (same)	0.656 (same)	0.57 (same)	0.50 (−3.8 %)	0.803 (+0.380)	0.427 (+0.426)
Full-model Fine-tuning (iter=300)	0.561 (−20.1%)	0.497 (−1.4%)	0.534 (−18.6%)	0.46 (−19.3%)	0.21 (−59.6%)	0.804 (+0.381)	0.429 (+0.428)
Full-model Fine-tuning (iter=500)	0.500 (−28.8%)	0.500 (−0.8%)	0.500 (−23.8%)	0.45 (−21.1%)	0.18 (−65.4%)	0.849 (+0.426)	0.429 (+0.428)
3-5 images (10 concepts)							
Soft Prompt (16 tokens)	0.702 (same)	0.504 (same)	0.656 (same)	0.57 (same)	0.51 (−1.9%)	0.742 (+0.319)	0.225 (+0.224)
Full-model Fine-tuning (iter=300)	0.500 (−28.8%)	0.500 (−0.8%)	0.500 (−23.8%)	0.45 (−21.1%)	0.20 (−61.5%)	0.748 (+0.325)	0.242 (+0.241)

Table 1. Soft-Prompt Tuning vs. Full-Model Fine-Tuning. Overall, soft-prompt tuning matches the performance of full-model fine-tuning for personalized abilities while retaining the original model's general capabilities.

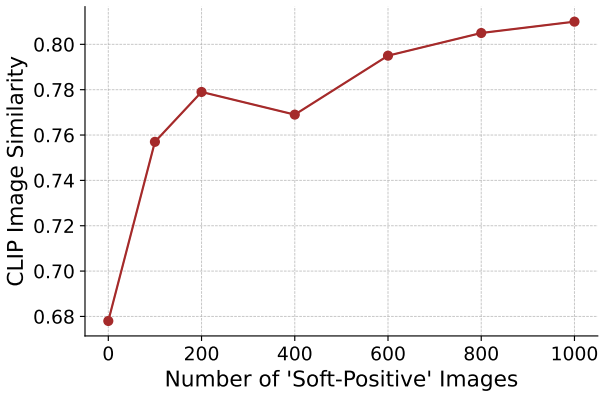

Figure 2. Ablation studies on the number of "soft-positive" images. Generally, increasing the number of "soft-positive" images helps to boost performance.

number of "soft-positive" images from 0 to 1000, where 0 indicates no "soft-positive" images were used during training, and 1000 indicates that 1000 "soft-positive" images were included.

The results are presented in Fig. 2. As shown, incorporating "soft-positive" images significantly improves performance compared to training with only positive images (e.g., 0.68 vs. 0.76+). Overall, increasing the number of "soft-positive" images enhances performance, with saturation observed around 1000 images when training with a soft prompt of token length $k = 16$ tokens.

2.2. Catastrophic Forgetting

Similar to the evaluation in Sec. 1, for general abilities, we utilized prominent benchmarks such as MMLU [5] for text-only generation, POPE [6], and MMBench [7] for visual question answering. For personalized abilities, we evaluated CLIP-Image Similarity [8] and Facial Similarity using the off-the-shelf ArcFace model [9] to compare generated images with the reference images.

The results are presented in Tab. 1 (last three rows). As shown, full-model finetuning leads to catastrophic forgetting across all benchmarks, with performance drops ranging from 1% to 61.5%. In contrast, using soft prompts preserves the model's general performance across nearly all benchmarks while achieving personalized abilities comparable to full-model finetuning (e.g., CLIP-Image Similarity is 0.742 vs. 0.748).

3. Data Augmentation Details

Here, we provide details about the data augmentation process for the ablation studies in the main paper. There are two main approaches for creating augmented training data: (A) Using positive images only, and (B) Using "Soft-Positive" Images (Ours).

Augmentation with Positive Images Only. The objective of this approach is to increase the diversity of training data when only 3–5 images of a subject are available. Inspired by [10], given an input image (e.g., a photo of a cat), we first obtain the corresponding object mask (e.g., the segmentation mask of the cat) using a pretrained SAM [11]. Subsequently, we randomly resize the subject (ranging from 30–100%) within a 512×512 image. This resized subject is then paired with a randomly selected background caption from a background library to inpaint the background (e.g., "A field of lavender flowers") using StableDiffusion-XL [12]. Fig. 3A illustrates this process.

The background library contains 100 captions, all generated by GPT-4o [13] and later human-audited. Table 2 lists 10 randomly selected examples of these captions. All augmented images generated through this process are treated as positive images and are given equal weight as positive samples during training.

Augmentation with "Soft-Positive" Images (Ours). In this approach, input images are used to retrieve the top N most similar images from LAION-5B [14]. These retrieved

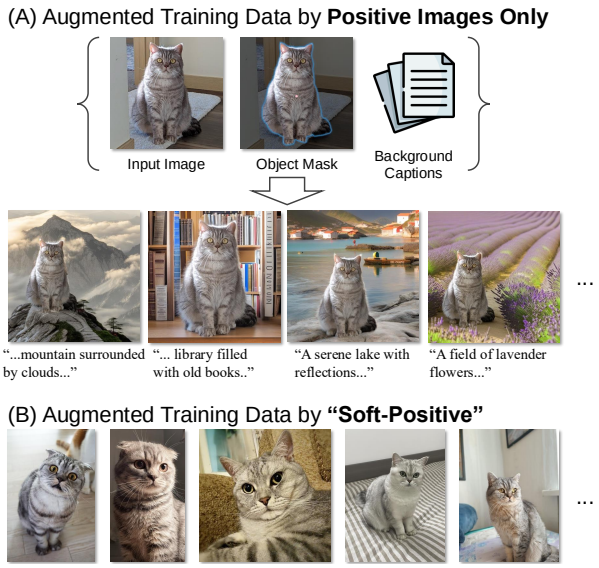

Figure 3. Comparison of data augmentation methods. Using 'Soft-Positive" images can increase both diversity and realism of the training data.

images are referred to as "soft-positive" images. The retrieved images are ranked, and an adaptive prompt length strategy is applied to describe them: the more similar a soft-positive image is to the input image, the more tokens are allocated to describe it. Fig. 3B provides some examples of these "soft-positive" images.

Comparisons. A key limitation of (A) Augmentation with Positive Images Only is that while the backgrounds vary, the foreground subject remains the same, which might restrict diversity in terms of the subject's pose or other variations. In contrast, the "soft-positive" images not only provide diverse background information but also add variations in the foreground, such as pose and angle.

Additionally, it is important to note that augmented images are generated content, whereas "soft-positive" images are real images. Training on real distributions can lead to more realistic results compared to training on generated (synthetic) distributions.

4. Limitation

Our method is not without limitations. The first limitation arises when dealing with objects that have intricate details (e.g., text on a cup or characters on a keyboard). Examples of such cases are shown in Fig. 4(a).

The second limitation is that, like other personalization methods [1, 3, 15], our method's performance is constrained by the capabilities of the base model. For instance, as [15] highlights, personalized Vision-Language Models like LLaVA [16] can still produce hallucinations (e.g., providing an incorrect date of birth for a person when asked). Similarly, our approach inherits the limitations of its underlying models, in this case, Chameleon/Anole [4, 17]. While these models perform reasonably well in generating object-centric images (e.g., "A photo of a dog"), they struggle with generating images involving multiple concepts (e.g., "A photo of a dog and a cat," as shown in Fig. 4(b)). Consequently, we were unable to test our approach on multiple personalized concepts effectively.

Lastly, although we achieved encouraging results in personalizing for individuals (e.g., facial similarity of 0.2xx), there remains a significant gap when it comes to personalizing human faces. For reference, the recommended threshold for facial recognition similarity is around 0.4–0.5, highlighting considerable room for improvement in this area.

Figure 4. Limitations. (a) lacks of details; (b) Generate multiple subjects

- A serene beach with golden sand and clear blue water.
- A vibrant sunset over a calm ocean.
- A snowy village during a peaceful winter evening.
- A quiet library filled with old books and wooden shelves.
- A crowded street in an ancient Asian market.
- A colorful spring garden in full bloom.
- A field of lavender flowers swaying in the breeze.
- A cozy coffee shop with a warm atmosphere and soft light.
- A stark, icy landscape with glaciers and frozen seas.
- A lush green valley surrounded by towering mountains.

Table 2. Sample of 10 out of 100 captions used for generating the background with Stable Diffusion-XL [12]

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. In *CVPR*, 2023. 1, 3

[2] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024.

[3] 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. In *arXiv*, 2022. 1, 3

[4] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. In *arXiv*, 2024. 1, 2, 3

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 1, 2

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *EMNLP*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics. 1, 2

[7] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *arXiv*, 2023. 1, 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *arXiv*, 2021. 1, 2

[9] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 2

[10] Yuheng Li, Haotian Liu, Yangming Wen, and Yong Jae Lee. Generate anything anywhere in any scene. In *arXiv*, 2023. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *arXiv*, 2023. 2

[12] 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. In *arXiv*, 2023. 2, 3

[13] OpenAI. Gpt-4o system card. In *arXiv*, 2024. 2

[14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *arXiv*, 2022. 2

[15] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. In *NeurIPS*, 2024. 3

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[17] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. In *arXiv*, 2024. 3