U-VAP: User-specified Visual Appearance Personalization via Decoupled Self Augmentation

Supplementary Material

In Sec. A, we provide user studies for human preference evaluation. We demonstrate the ability of attribute combination in Sec. B and analyze the impact of constructed attribute-aware samples on personalization in Sec. C. The limitation of our method is discussed in Sec. D.

A. User Study

We conduct user studies to evaluate the proposed U-VAP in specified visual appearance personalization on seven different concepts with three baseline methods: Textual Inversion (TI) [1], DreamBooth [4], and ProSpect [5]. To obtain one set of image results, given a reference concept and the specified attribute, four images of a certain novel concept are generated by each method. Every volunteer is asked to select the best image based on two criteria to evaluate per image set:

- Criterion-I: the accuracy of attribute personalization.
- Criterion-II: the quality of novel concept generation on the premise of attribute accuracy.

We generated 24 sets of these images using different specified attributes and concepts, and a total of 55 volunteers participated in our study. As shown in Tab. S1, when evaluating the highest accuracy of attribute personalization, the images from our method were voted with a probability of 64.71%. The results by DreamBooth were chosen with the second maximum probability, which was only 15.42% Taking consideration of attribute accuracy, our images were voted with a probability of 53.84% for choosing the best quality of novel concept generation. However, the maximum probability of selecting images from another method is only 17.92% (TI). Compared with the other three baseline methods, images generated by our method were chosen with higher probability. These results indicate U-VAP generates novel concepts with accurately specified attributes, exhibiting better performance in human preference.

Table S1. User study on two criteria. Criterion-I: The accuracy of attribute personalization. Criterion-II: The quality of novel concept generation. The highest voting probability is in bold.

Methods	Criterion-I	Criterion-II
Ours	$\mathbf{67.50\%}$	$\mathbf{57.22\%}$
ProSpect	8.47%	10.69%
DreamBooth	15.42%	14.58%
TI	8.47%	17.92%

B. Attribute Combination

We demonstrated the ability of U-VAP to combine different learned attributes. As shown in Fig. S1, given two different reference concepts, U-VAP can combine their different attributes. For example, when given an image of a vase with a unique pattern and a backpack with a dog face pattern, we construct attribute-aware samples for each to learn the pattern of the former and the structure of the latter. Subsequently, U-VAP simultaneously learns two target attributes on the same model and represents them with different identifiers ("sks1" and "sks2").

In the inference phase, the user could directly write them in a prompt (as shown in Fig. S1: "a photo of sks1 pattern sks2 backpack") and generate a new concept. This new concept has the unique serrated pattern of the reference vase and the backpack's structure. Similarly, given a colored cat statue and an image of a colored kettle, we generate a new concept of an appearance with colored stripes and a kettle structure that is the same as the reference concept.

C. Attribute-aware Samples

Fig. S2 (a) illustrates a pair of attribute-aware sets constructed from the reference image and the target attribute by U-VAP. Specifically, when the target attribute is "color", different objects in the target attribute set exhibit the same color as the reference concept. We obtain the target attribute set by conditioning the initial concept-aware model with a target description ("a photo of tgt color ") Conversely, in the non-target attribute set generated by the non-target description, the concepts in each image have colors different from the reference concept. Although images in the target attribute set have the specified color of reference concept, some of them are not in line with the description entirely. For example, the initial model fails to generate "broom" or "coach" correctly, so we can not use it directly as a great attribute-aware model. Nonetheless, we only care about the target attribute (color of reference statute) of the target attribute set, and the error of other non-target attributes does not affect the learning of target identifier "tgt".

Using these constructed samples, we evaluate the influence of the number of attribute-aware samples on specific appearance learning, as shown in Fig. S2 (b). We generated a car, sunglasses, and a hat using the statue's color in the reference image, represented by "sks". Taking the example of "hat", when the number of attribute-aware samples (represented as n) is 4, the generated results barely match the

Figure S1. Given two different reference images, U-VAP can combine specific attributes from each concept and generate a new concept. Left: the combination of the pattern of the red vase and the specific backpack. Right: the combination of the color of the cat statue and the specific kettle.

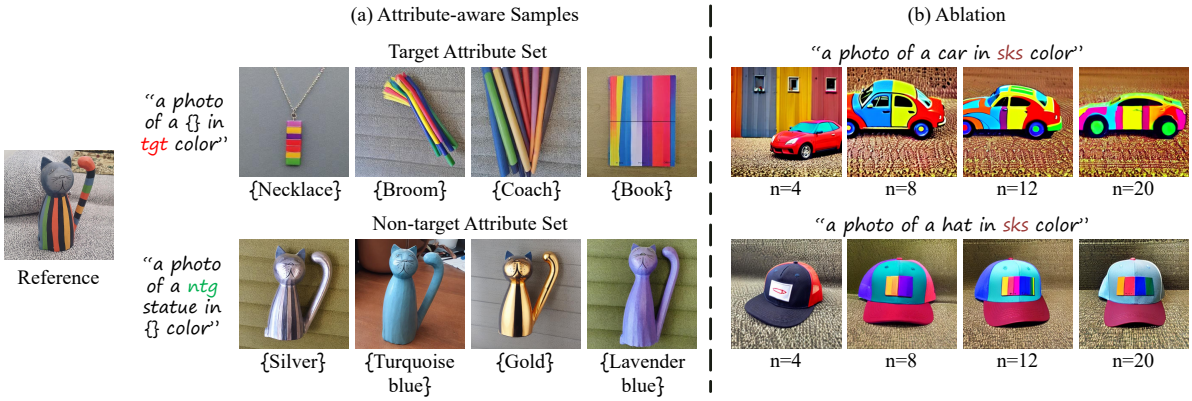

Figure S2. When specifying the color of the cat statue for personalization, the constructed attribute-aware samples are shown in (a). The target attribute set comprises various novel concepts with the reference color of the reference cat statue, while the non-target attribute set includes cat statues with different appearances. They are generated with target and non-target descriptions respectively. (b) demonstrates the generated results of various new concepts with the specified color using different quantities of attribute-aware samples. As the number of attribute-aware samples increases, the generated images more accurately embody the specified color.

reference concept's color. However, when n=8, the generated results meet the requirements. The generated concept has the same colorful stripe as the reference. Further increasing n, we observed minimal improvement in learning the specified appearance.

D. Limitation and Discussions

Because of the dependence on the capability of basic concept-aware personalization methods (such as DreamBooth (DB) [4]) in the pre-learning step, U-VAP could have poor performance in decoupling attributes of some cases, as shown in Fig. S3 (a). Given a photo of the colorful cat statue, we compare DB and U-VAP in generating a birdhouse in the specific color based on two different models, respectively. The original DreamBooth and U-VAP are based on Stable Diffusion (SD) v1.5 [3]. Additionally, we utilize SDXL [2] in DreamBooth and our method for further comparison. Although the original U-VAP performs better than DreamBooth, it fails to completely decouple the undesired tail of the cat statue. DreamBooth on SD v1.5 has limited ability to accurately decouple the target attribute, leading to low-quality constructed attribute-aware samples

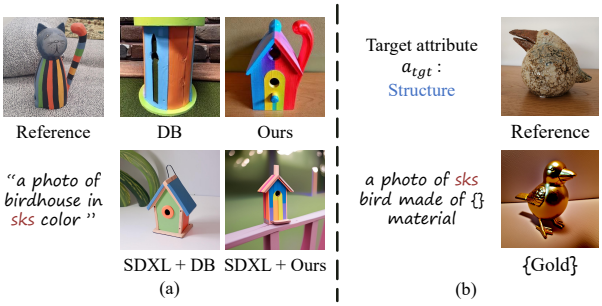

Figure S3. Bad cases. (a) DreamBooth (DB) [4] and U-VAP fails in generation of some concept such as "birdhouse" because of the limitation of SD-based personalization. Based on a better basic model like SDXL, U-VAP overcomes this limitation and achieves better decoupling (SDXL + Ours). (b) After personalizing the structure of the bird statue and bounding it with the identifier "sks", U-VAP could generate bad results because of the influence of the prior information of certain words ("bird").

Figure S4. Results by SDXL-based U-VAP.

and bad results at the end.

However, based on the SDXL model, which performs better capability in text-to-image generation compared to previous versions of Stable Diffusion, the results of both DreamBooth and U-VAP achieve better quality. We believe that with a better basic model like SDXL, the ability of attribute personalization of DreamBooth is improved and the efficiency of U-VAP can be further increased with high-quality attribute-aware samples.

In Fig. S4, we provide more results generated by SDXL-based U-VAP to demonstrate the quality of specified appearance personalization.

Furthermore, in Fig. S3 (b) we show another type of bad case where the structure of "bird" in the inference prompt suppresses the target attribute "structure" of the reference image in the result. This is because the prior information of certain words may have stronger dominance in semantics and influence the learned attributes in some cases.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China under No. 62102162, Beijing Science and Technology Plan Project under No. Z231100005923033.

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[2] 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. 2

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 2

[5] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 1