HybridBooth: Hybrid Prompt Inversion for Efficient Subject-Driven Generation
** *Supplementary Material***

Shanyan Guan[1⋆], Yanhao Ge[1,3⋆], Ying Tai[2✉], Jian Yang[2]
Wei Li[1], and Mingyu You[3✉]

[1] vivo Mobile Communication Co., Ltd
[2] School of Intelligence Science and Technology, Nanjing University
[3] College of Electronic and Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
{guanshanyan, halege}@vivo.com yingtai@nju.edu.cn myyou@tongji.edu.cn
https://sites.google.com/view/hybridbooth

1 Using UNet or MLP structure for the Prompt Regressor?

Recall that in Sec 4.2, we use the downsample and middle blocks of the UNet [5] as the prompt regressor. Alternatively, previous direct-regression-based methods adopt the multi-layer perception (MLP) structure. We compare the effect of these two kinds of structures. Specifically, we use the MLP structure the same as FastComposer [8].

Qualitative Results. The qualitative comparison results are shown in Fig. 1. We take DreamBooth as the reference method for more clear comparison. In the first line of Fig. 1, we visualized the learned subject word embedding 'S*'. We can observe that using UNet structure can better implant detailed subject characteristics, *e.g.* lips, hairstyle, forehead. From the editing results shown in the second line of Fig. 1, the UNet structure is better aligned with the input image in terms of identity and also with the given guidance prompt. In short, the UNet structure used in HybridBooth provides better subject and prompt fidelity.

Quantitative Results. From the quantitative results reported in Tab. 1, we can observe that using UNet structure as the prompt regressor consistently outperforms using MLP structure in both metrics related to subject-fidelity (CLIP-I, DINO-I) and prompt-fidelity (CLIP-T).

2 Qualitatively analyzing `HybridBooth`

Limited by the space of the manuscript, we only qualitatively analyzed the contributions of the propose `HybridBooth` in Sec. 5.4, including multi-grained feature merging (DINO and CLIP), mask regularization, number of word embedding.

⋆ Equal contributions. Ying Tai and Mingyu You are corresponding authors.

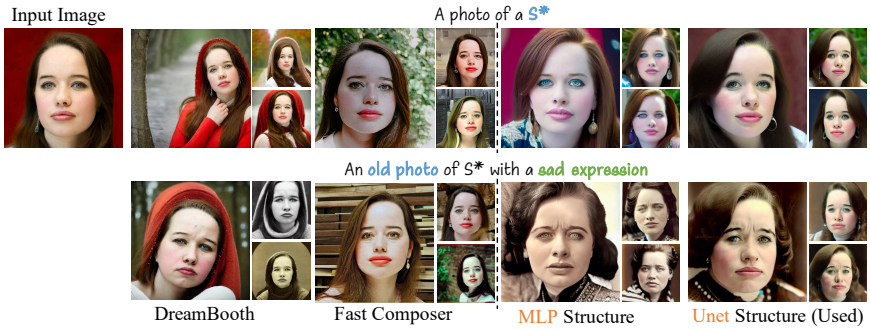

Fig. 1: Qualitatively comparing UNet (Used) or MLP structure in the prompt regressor.

Table 1: Quantitative Evaluation on the CelebA-HQ. **First Block**: Optimization-based methods. **Second Block**: Direct-regression-based methods. **Third Block**: `HybridBooth` and its ablations.

Method	CelebA-HQ			DreamBooth-dataset			Iter. Step$^{\downarrow}$
	CLIP-T $^{\uparrow}$	CLIP-I $^{\uparrow}$	DINO-I $^{\uparrow}$	CLIP-T $^{\uparrow}$	CLIP-I $^{\uparrow}$	DINO-I $^{\uparrow}$	
Textual Inversion [2]	0.164	0.612	0.236	0.183	0.663	0.462	5000
DreamBooth [6]	**0.251**	0.564	0.376	0.251	0.785	0.674	1000
Custom Diffusion [4]	0.237	0.675	0.398	0.245	0.801	0.695	200
ELITE [7]	0.169	0.592	0.311	<u>0.255</u>	0.762	0.652	1
FastComposer [8]	0.201	0.782	<u>0.581</u>	-	-	-	1
MLP Structure	0.210	0.763	0.567	0.221	0.755	0.638	5
UNet Structure (used)	0.246	0.865	0.644	0.261	0.865	0.755	5

Here, we further provide qualitative evaluation results in Fig. 2. We take DreamBooth and FaseComposer [8] as optimization- and direct-regression-based reference methods, respectively. From Fig. 2, we can observe that:

- Only using CLIP feature can effectively preserve semantic attributions, but it cannot preserve pixel-related information, *e.g.* texture details and eye contact. By combining with the DINO feature, this problem can be effectively addressed.
- From Column 6, only using one word cannot fully represent the subject's characteristics.
- All variants of `HybridBooth` outperform reference methods (DreamBooth and FastComposer). This verifies the advantages of the hybrid framework in subject-concept learning.

3 What if using ID feat for facial images?

As shown in Tab. 2, replacing DINOv2 features with identity features (extracted by ArcFace [1]) resulted in worse performance. However, merging both features

Fig. 2: More visualizations on the contributions of the HybridBooth. The input prompt is 'A photo of a S*'

Table 2: Quantitative studies on ID feature for facial images (on CelebeA-HQ)

Method	CLIP-T ↑	CLIP-I ↑	DINO-I ↑
Custom Diffusion [33]	0.237	0.675	0.398
ELITE [45]	0.169	0.592	0.311
FastComposer [46]	0.201	0.782	0.581
FastComposer with DINOv2 feat.	0.231	0.811	0.602
Replace DINOv2 feat. with ID feat.	0.240	0.853	0.634
DINOv2 feat. **and** ID feat.	0.248	0.870	0.648

improved the performance, indicating that identity features complement DINOv2 features.

4 Data Augmentation for the Optimization Methods.

`HybridBooth` only needs **one** subject image. However, the compared optimization baselines usually need $3-5$ images in their implementation. Therefore, we augment the input subject image to obtain 5 different images, in order to compare as fairly as possible. Specifically, we remove the background of the subject image using InSPyReNet [3]. Then we apply affine transformations with random rotation ($-10-10$ degrees), translation ($0.2-0.5$), and scale ($0.6-1.0$). Finally, we use the Stable Diffusion inpainting model to generate the background with a prompt of "portrait photo of a person, realistic, professional".

References

1. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019) 2
2. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2022) 2
3. Kim, T., Kim, K., Lee, J., Cha, D., Lee, J., Kim, D.: Revisiting image pyramid structure for high resolution salient object detection. In: Proceedings of the Asian Conference on Computer Vision. pp. 108–124 (2022) 3

4. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) 2
5. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 1
6. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023) 2
7. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: ICCV. pp. 15943–15953 (2023) 2
8. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. CoRR:2305.10431 (2023) 1, 2