# Supplementary Materials of *InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning*

Jing Shi* Wei Xiong* Zhe Lin Hyun Joon Jung
Adobe Inc.
$*$ Equal Contribution
{jingshi,wxiong,zlin,hjung}@adobe.com

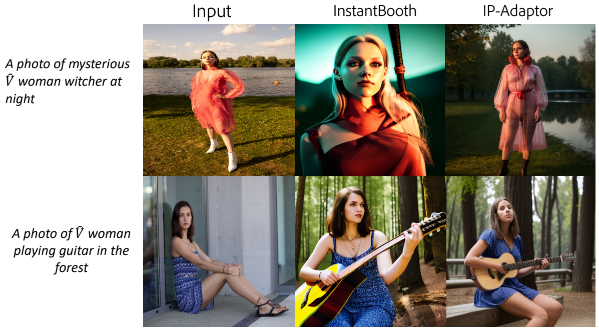

Input
InstantBooth
IP-Adaptor
A photo of mysterious $\hat{V}$ woman witcher at night
A photo of $\hat{V}$ woman playing guitar in the forest

Figure 1. Comparison between InstantBooth and IP-Adaptor.

## 1. Adjust the adapter weight $\beta$ and concept renormalization factor $\alpha$

Tab. 1 shows different compositions of $\beta$ and $\alpha$. The results indicate that larger $\beta$ or $\alpha$ can both contribute to better identity preservation but weaker language comprehension ability. We finally choose the model with $\beta = 0.3$, $\alpha = 0.4$ as a trade-off.

## 2. Comparison with IP-Adapter

We compare our model with IP-adaptor on the person category. Results shown in Fig. 1 indicate our model's stronger capability on identity preservation.

## 3. Influence of dataset

Our model is trained on domain-specific object categories such as person and cat. To analyze the impact of dataset, we further train a generic InstantBooth on the dataset composed of general objects (10M images from a wide variety of object categories). As shown in Fig. 2, the generic model generalizes well to the testing images, indicating that InstantBooth still outperforms ELITE when trained on general objects regarding identity preservation and image quality.

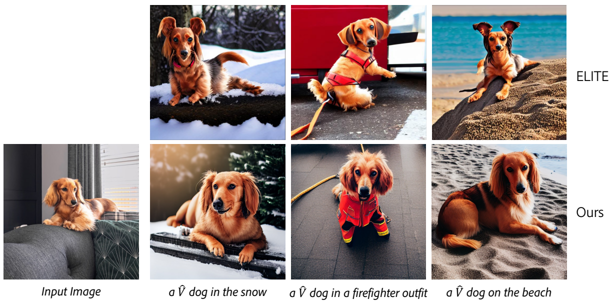

ELITE
Ours
Input Image
a $\hat{V}$ dog in the snow
a $\hat{V}$ dog in a firefighter outfit
a $\hat{V}$ dog on the beach

Figure 2. Results of our model trained on images of general object categories (dog as a test example), with comparison to ELITE.

| $\beta$ | $\alpha$ | Align ↑ | Reconstruct↑ |
|---|---|---|---|
| 0.3 | 0.1 | 0.3242 | 0.6631 |
| 0.3 | 0.3 | 0.3232 | 0.7002 |
| 0.3 | 0.4 | 0.3140 | 0.7329 |
| 0.3 | 0.5 | 0.3032 | 0.7544 |
| 0.3 | 1.0 | 0.2087 | 0.7905 |
| 0.5 | 0.1 | 0.3127 | 0.7051 |
| 0.5 | 0.3 | 0.3076 | 0.7480 |
| 0.5 | 0.5 | 0.2874 | 0.7778 |
| 1.0 | 1.0 | 0.2017 | 0.7944 |

Table 1. Hyper-parameter adjustment for adapter weight $\beta$ and concept renormalization factor $\alpha$.

## 4. Quantitative results for cats

We collect 20 cat subjects in different species and each consists of five images of the same cat. For Textual Inversion/DreamBooth/Ours, the language alignment score is 0.2620/0.3027/0.2922, the reconstruct score is 0.8369/0.8540/0.8369. The quantitative performance of our model is on par with DreamBooth and better than Textual Inversion on cat.

## 5. Data collection, data/code release, and evaluation metrics

Both the person and cat images are collected from our internal database. Due to the restriction, we cannot release our dataset or pre-trained model weights. However, we may release the pre-trained weights once we have retrained our model on public datasets like Open Images.

For the *Reconstruction* metric, if there are multiple generated images and multiple ground-truth images, then we calculate the average similarity between each pair of generated image and input image. For example, if there are 4 generated images and 5 input images, then we calculate the average similarity of 20 pairs of generated image and input image.

## 6. Details of Test Prompts

The specific prompts that we use to obtain the quantitative results for person are as follows.
*"a photo of $\hat{V}$ [class noun] in the swimming pool"*
*"a photo of $\hat{V}$ [class noun] in New York"*
*"a photo of $\hat{V}$ [class noun] on the moon"*
*"a photo of $\hat{V}$ [class noun] in the kitchen"*
*"a photo of $\hat{V}$ [class noun] in mountain with aurora"*
*"a photo of $\hat{V}$ [class noun] in the library"*
*"a painting of $\hat{V}$ [class noun] in Van Gogh style"*
*"a manga drawing of $\hat{V}$ [class noun]"*
*"a colorful graffiti of $\hat{V}$ [class noun]"*
*"a $\hat{V}$ [class noun] funko pop"*
*"a pencil drawing of $\hat{V}$ [class noun]"*
*"a Ukiyo-e painting of $\hat{V}$ [class noun]"*
*"a photo of $\hat{V}$ [class noun] holding a corgi on a bench"*
*"a photo of $\hat{V}$ [class noun] playing guitar in the forest"*
*"a photo of $\hat{V}$ [class noun] shaking hands with Joe Biden"*
*"a photo of mysterious $\hat{V}$ [class noun] witcher at night"*
*"a photo of $\hat{V}$ [class noun] riding a bicycle under Eiffel Tower"*

## 7. Limitations and Future Work

While our model exhibits strong performance and fast speed, it still has several limitations. First, we have to separately train the model for each object category. This limitation has been addressed by training the model with more data of different categories together. We have shown some visual results in Fig. 2. We plan to continue improving the model on more object categories to solve some failure cases. Second, due to the current design of the adapter, it can only accept a single concept to provide the identity details. In addition, the current visual performance is limited by the synthesizing quality of the pretrained Stable Diffusion, such as the hand artifacts in some results. We plan to use stronger pretrained text-to-image models as the backbone to obtain better visual results.

## 8. More Visual Examples

Fig. 3 and Fig. 4 show more visual examples of our model.

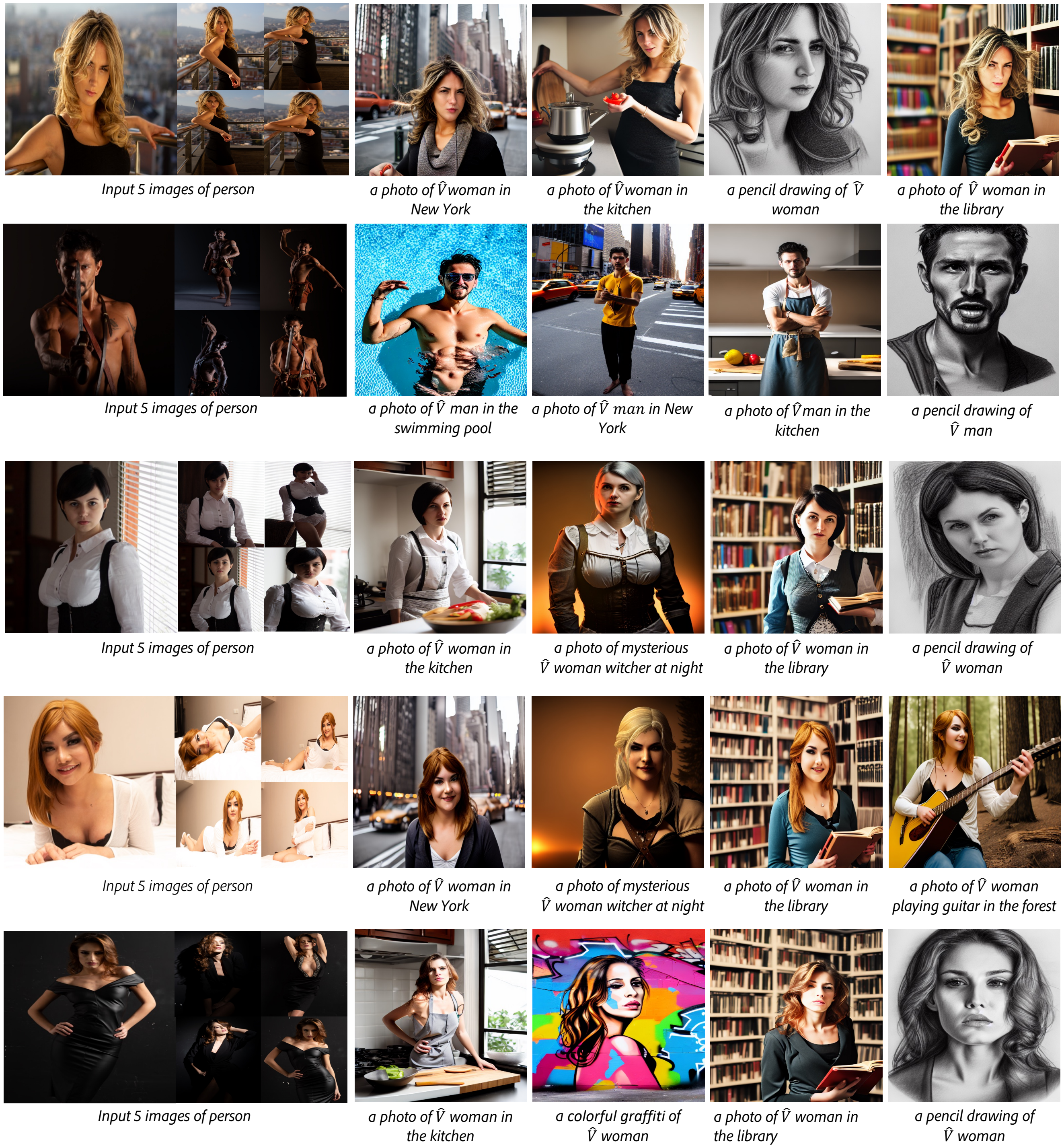
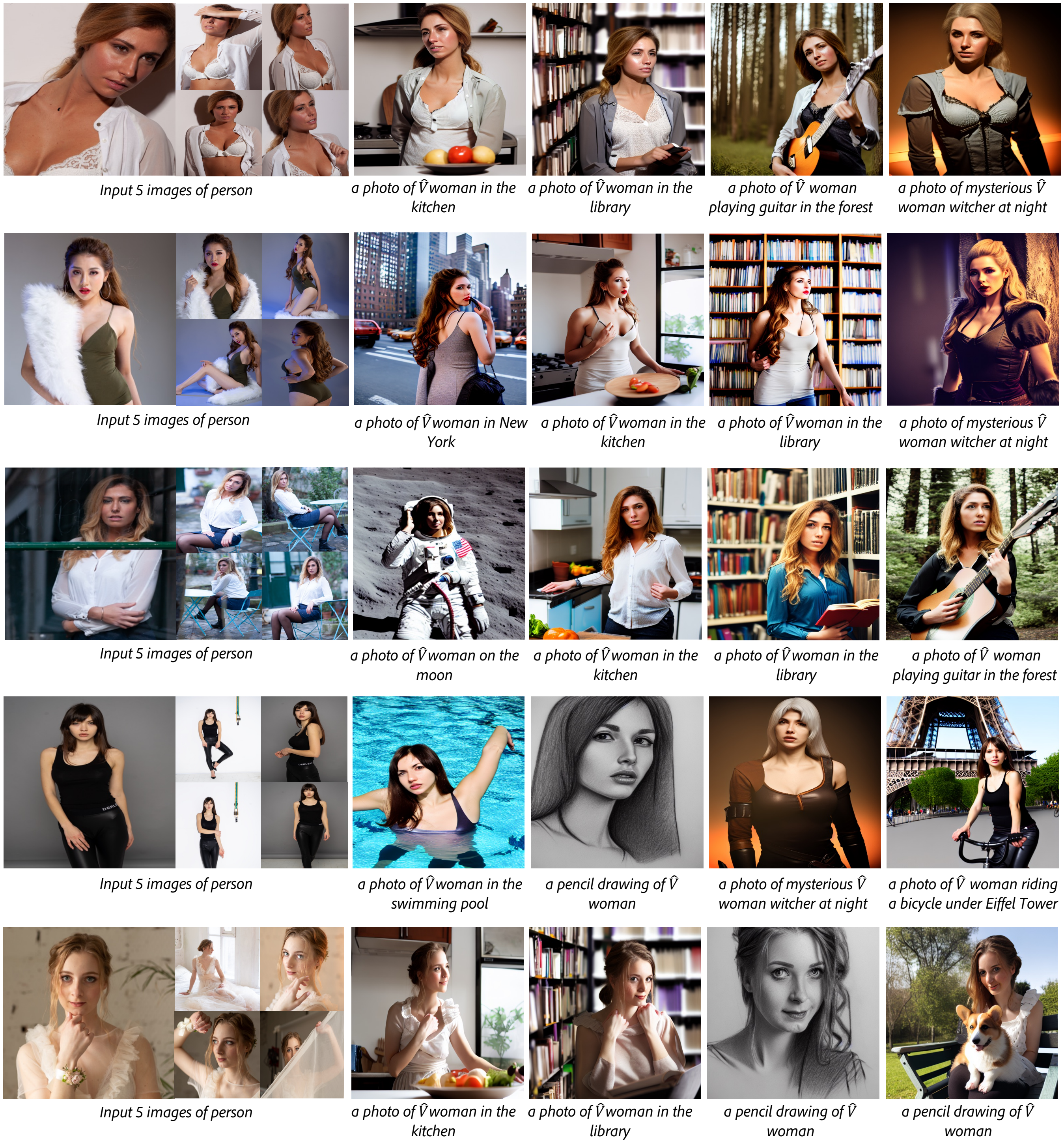

Input 5 images of person
a photo of $\hat{V}$ woman in New York
a photo of $\hat{V}$ woman in the kitchen
a pencil drawing of $\hat{V}$ woman
a photo of $\hat{V}$ woman in the library
Input 5 images of person
a photo of $\hat{V}$ man in the swimming pool
a photo of $\hat{V}$ man in New York
a photo of $\hat{V}$ man in the kitchen
a pencil drawing of $\hat{V}$ man
Input 5 images of person
a photo of $\hat{V}$ woman in the kitchen
a photo of mysterious $\hat{V}$ woman witcher at night
a photo of $\hat{V}$ woman in the library
a pencil drawing of $\hat{V}$ woman
Input 5 images of person
a photo of $\hat{V}$ woman in New York
a photo of mysterious $\hat{V}$ woman witcher at night
a photo of $\hat{V}$ woman in the library
a photo of $\hat{V}$ woman playing guitar in the forest
Input 5 images of person
a photo of $\hat{V}$ woman in the kitchen
a colorful graffiti of $\hat{V}$ woman
a photo of $\hat{V}$ woman in the library
a pencil drawing of $\hat{V}$ woman

Figure 3. More visual results of our model on the person category.



Input 5 images of person

a photo of $\hat{V}$ woman in the kitchen

a photo of $\hat{V}$ woman in the library

a photo of $\hat{V}$ woman playing guitar in the forest

a photo of mysterious $\hat{V}$ woman witcher at night

Input 5 images of person

a photo of $\hat{V}$ woman in New York

a photo of $\hat{V}$ woman in the kitchen

a photo of $\hat{V}$ woman in the library

a photo of mysterious $\hat{V}$ woman witcher at night

Input 5 images of person

a photo of $\hat{V}$ woman on the moon

a photo of $\hat{V}$ woman in the kitchen

a photo of $\hat{V}$ woman in the library

a photo of $\hat{V}$ woman playing guitar in the forest

Input 5 images of person

a photo of $\hat{V}$ woman in the swimming pool

a pencil drawing of $\hat{V}$ woman

a photo of mysterious $\hat{V}$ woman witcher at night

a photo of $\hat{V}$ woman riding a bicycle under Eiffel Tower

Input 5 images of person

a photo of $\hat{V}$ woman in the kitchen

a photo of $\hat{V}$ woman in the library

a pencil drawing of $\hat{V}$ woman

a pencil drawing of $\hat{V}$ woman

Figure 4. More visual results of our model on the person category.