ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models – Supplementary Materials

YUXIN ZHANG and WEIMING DONG, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China
FAN TANG, Institute of Computing Technology, CAS, China
NISHA HUANG, School of Artificial Intelligence, UCAS, China and MAIS, Institute of Automation, CAS, China
HAIBIN HUANG and CHONGYANG MA, Kuaishou Technology, China
TONG-YEE LEE, National Cheng-Kung University, Taiwan
OLIVER DEUSSEN, University of Konstanz, Germany
CHANGSHENG XU, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China

ACM Reference Format:
Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models – Supplementary Materials. *ACM Trans. Graph.* 42, 6, Article 246 (December 2023), 2 pages. https://doi.org/10.1145/3618342

In the supplementary material, we present the quantitative evaluation of other attributes, the examples of the user study, and preliminary.

1 QUANTITATIVE EVALUATION OF OTHER ATTRIBUTES

Table 1 shows the quantitative evaluation of attribute-aware generation campared with Textual Inversion (TI) [Gal et al. 2023], DreamBooth [Ruiz et al. 2023] and InST [Zhang et al. 2023]. For material, style, and layout, we selected 8 concepts as references. Each concept comes with three results.

2 USER STUDY

A screenshot of our user study web pages is shown in Fig. 1. Options A and B show the results generated by our method and by one of the comparative image style transfer methods. The comparative method tested in each question and the order of the options are both random.

Authors' addresses: Yuxin Zhang; Weiming Dong, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, zhangyuxin2020@ia.ac.cn, weiming.dong@ia.ac.cn; Fan Tang, Institute of Computing Technology, CAS, China, tangfan@ict.ac.cn; Nisha Huang, School of Artificial Intelligence, UCAS, China and MAIS, Institute of Automation, CAS, China, huangnisha2021@ia.ac.cn; Haibin Huang; Chongyang Ma, Kuaishou Technology, China, huanghaibin03@kuaishou.com, chongyangma@kuaishou.com; Tong-Yee Lee, National Cheng-Kung University, Taiwan, tonylee@ncku.edu.tw; Oliver Deussen, University of Konstanz, Germany, oliver.deussen@uni-konstanz.de; Changsheng Xu, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, csxu@nlpr.ia.ac.cn.

0730-0301/2023/12-ART246
https://doi.org/10.1145/3618342

Table 1. CLIP-based evaluations. The best results are in **bold**, and the second best results are underlined. **Baseline*: the reference image.

Metric	Text Similarity↑			Image Similarity↑		
Method	**ProSpect**	DreamBooth	TI	**ProSpect**	DreamBooth	TI
Material	**0.2243**	0.1878	<u>0.2125</u>	<u>0.7424</u>	**0.7701**	0.5598

Metric	Text Similarity↑		Image Similarity↑	
Method	**ProSpect**	InST	**ProSpect**	InST
Material	**0.3011**	0.2840	**0.6632**	0.6117

Metric	Text Similarity↑		Image Similarity↑	
Method	**ProSpect**	Baseline*	**ProSpect**	Baseline
Material	**0.2478**	0.0982	**0.6977**	1

3 PRELIMINARY

The diffusion model continuously adds noise to the initial data distribution x_0 and finally makes the data distribution into independent Gaussian distributions. The forward diffusion process is defined as:

$$\begin{aligned} q\left(x_t \mid x_{t-1}\right) &= \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right), \\ q\left(x_{1:T} \mid x_0\right) &= \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right). \end{aligned} \tag{1}$$

$q(x_t)$ can be derived by reparameterization:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_{t-1} = \ldots = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} z, \tag{2}$$

where x_t denotes the intermediate latent map at a time step t, z denotes the added noise, β_t denotes the standard deviation, and $\alpha_t = 1 - \beta_t$ denotes the noise intensity. The standard deviation β_t of the noise added at each time step is specified and increases as t increases. The mean value of the noise added at each time step is adjusted according to β_t, to ensure that x_T converges stably to $\mathcal{N}(0,1)$. From $\mathbf{x_t} = \sqrt{\overline{\alpha_t}} \mathbf{x}_0 + \sqrt{1-\overline{\alpha_t}} \mathbf{z}$ can get that $\mathbf{q}\left(\mathbf{x}_\mathrm{t} \mid \mathbf{x}_0\right) = \mathbf{N}\left(\mathbf{x}_\mathrm{t}; \sqrt{\bar{\alpha}_\mathrm{t}} \mathbf{x}_0, (\mathbf{1} - \bar{\alpha}_{\ mathrmt}) \mathbf{I}\right)$ As noise is added, x_t gradually approaches pure Gaussian noise $\mathbf{x}_0 = \frac{\mathbf{1}}{\sqrt{\alpha_\mathrm{t}}}\left(\mathbf{x}_\mathrm{t} - \sqrt{\mathbf{1}-\overline{\alpha}_\mathrm{t}} \mathbf{z}_\mathrm{t}\right)$. We speculate that the attribute tendency of diffusion is to add noise standard deviation β_t gradually.

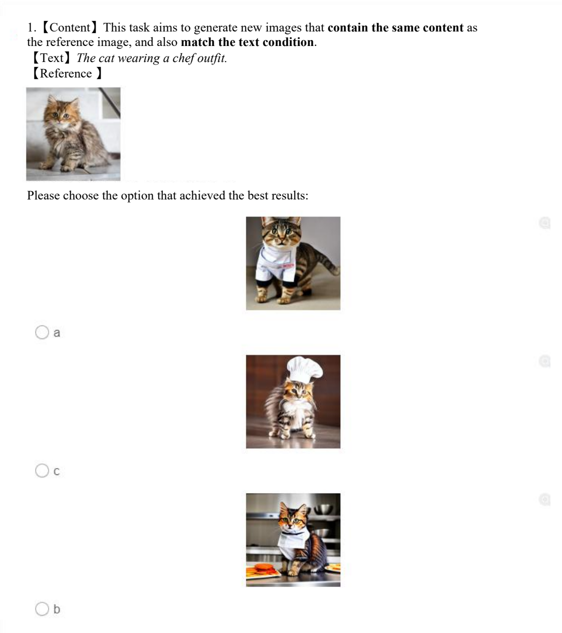

Fig. 1. Screenshot of our user study web page.

The Fourier transform is a classic transformation widely used in digital image processing. It transforms a signal from the time domain into the frequency domain, facilitating the identification of subtle features and the processing of challenging components. Grayscale images consist of discrete points in two dimensions, and the Two-Dimensional Discrete Fourier Transform (2D-DFT) is commonly used in image processing to obtain the frequency spectrum of an image that reflects its degree of grayscale variation. The center of the Fourier spectrum represents the low-frequency signal, whereas higher frequencies are represented by points closer to the edge. High-frequency signals typically correspond to edges and noise in the image, while the smooth areas of the image correspond to low-frequency signals. We can easily manipulate the high-frequency or low-frequency information of the image in the frequency domain to complete operations such as image denoising, image enhancement, and edge extraction. The Discrete Fourier Transform (DFT) of an image is formulated as:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi(ux/M + vy/N)}, \tag{3}$$

where M and N denote the length and height of the image, respectively. $F(u,v)$ denotes the frequency domain image, and $f(x,y)$ represents the time domain image. The range of u is $[0, M-1]$, and the range of v is $[0, N-1]$. The Inverse Discrete Fourier Transform (IDFT) of an image is formulated as:

$$f(x,y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u,v) e^{j2\pi(ux/M + vy/N)}. \tag{4}$$

REFERENCES

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations (ICLR)*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-Based Style Transfer with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10146–10156.