Supplementary Material
Att-Adapter: A Robust and Precise Domain-Specific Multi-Attributes T2I Diffusion Adapter via Conditional Variational Autoencoder

A. Evaluation Details

Dataset and Preprocessing **FFHQ**: FFHQ [6] is a high-quality public face dataset. To use the data for finetuning Text-to-image diffusion models, we get a caption per image by using ChatGPT [1]. We preprocessed 20 facial attributes: 13 attributes for facial compositions, one attribute for age, and 6 attributes for race.

Specifically, we first obtain head pose information (pitch, roll, and yaw)[1] and use them to get frontal view face images; We extract 16,336 images out of 70,000 images, which are used as our fine-tuning data. Next, we use 68 landmarks per image from dlib [7] to obtain the information about facial composition. For example, to obtain 'the gap between eyes', we use the 39th and 42nd landmarks coordinates and compute the L1 norm. Min-max normalization is applied to make the range of the value within 0 to 1, which means that 1 and 0 respectively indicate the max and min values of a certain attribute from the fine-tuning dataset. The 13 attributes related to facial compositions are: 'the gap between eyes', 'width of eyes', 'height of eyes', 'width of nose', 'height of nose', 'width of mouth', 'height of mouth', 'width of face', 'height of face' (from glabella to chin), 'height between eyebrow and eye', 'height between nose and mouth', 'height between mouth and chin', and 'width of eyebrow'.

For age and race extraction, we use deepface [15] with a detector Retinaface [2]. The age is divided by 100 for normalization purposes. The six extracted features about the race are 'p(Asian)', 'p(Black)', 'p(Hispanic Latino)', 'p(Indian)', 'p(Middle Eastern)', 'p(White)'.

EVOX: We used a commercial grade vehicle image dataset from EVOX [2] as the experiment data. The dataset includes high-resolution images across multiple car models collected from 2018 to 2023, totally 2030 images.

[1] https://github.com/DCGM/ffhq-features-dataset

[2] The images used in this study are the property of EVOX Productions, LLC and are subject to copyright law. The appropriate licenses and permissions have been obtained to ensure the rightful use of these images in our study. For more information, see https://www.evoxstock.com/.

Predefined 30 text prompts for evaluation 'A smiling man', 'A smiling woman', 'A man surprised', 'A woman surprised', 'An angry man', 'An angry woman', 'A man crying', 'A woman crying', 'A sad man', 'A sad woman', 'A painting of a man', 'A painting of a woman', 'A marble sculpture of a man', 'A marble sculpture of a woman', 'A 3D model of a man', 'A 3D model of a woman', 'A man wearing earrings', 'A woman wearing earrings', 'A man with a crown', 'A woman with a crown', 'A man wearing a cap', 'A woman wearing a cap', 'A man wearing eyeglasses', 'A woman wearing eyeglasses', 'A man with rainbow hair', 'A woman with rainbow hair', 'A man with shadow fade hair style', 'A woman with ponytail', 'A man wearing a furry cat ears headband', 'A woman wearing a furry cat ears headband'.

Test data creation for the evaluation for the baselines (Latent Control). The 13 attributes that we preprocessed from FFHQ are used to measure single attribute performance. The multi-attributes that we used to measure are as followings: ('height_between_eyebrow_eye', 'width_of_face'), ('width_of_mouth','height_of_face'), ('width_of_nose','height_of_mouth'), ('height_between_mouth_chin','gap_between_eyes'), ('height_of_face','height_of_eyes'), ('width_of_eyes','width_of_face'), ('width_of_nose','height_of_eyes'), ('width_of_nose','gap_between_eyes'), ('width_of_eyebrow','height_between_nose_mouth'), ('height_of_nose','width_of_eyes'). We generate a postive and a negative pair from each attribute set and compare the CR and DIS scores as mentioned in the main paper.

Test data creation for the evaluation for the baselines (Absolute Control). For each prompt, we randomly sample 50 combinations of 20 attributes. As a result, the attribute combinations for testing contains 1,500 samples; 30 (text prompts) by 50 (20-dimensional attributes combinations per prompt). Specifically, for facial composition attributes, we first obtain the means and the standard devia-

tions of the attributes from the finetuning data. For example, the mean of 'gap between eyes' is 0.687 and the standard deviation is 0.087. We sample from 2 sigma region. For age, we sample from normal distribution of which mean and standard deviation is 30 and 10. For race, we uniformly sample from 0.8 and 1 to assign the biggest value to one of 6 races. Once the major race is determined, the values for other races are randomly assigned to be sum-to-one.

B. Additional Related Works

Controllable Text-to-Image Generation T2I models [9, 11, 12, 14] generate high quality images from text prompts, using large pretrained visual language models like CLIP [10]. However, text only conditioning lacks fine-grained control as text often underspecifies visual details such as object type, perspective and style [5, 17]. To improve control, various approaches incorporate images as additional conditions. ControlNet [18] conditions image generation on inputs like depth, sketches, and semantic maps by training a copy of the diffusion model on these inputs, allowing spatial control. Similarly, T2I-Adapter [8] uses a lightweight adapter for external signals such as images, while preserving the pre-trained model's capabilities. IP-Adapter [16] adds controllability via cross-attention networks for both text and image inputs, enabling guidance with a reference image.

However, these methods [8, 10, 18] generate images based on a provided example (e.g., body pose) but lack precise control over specific attributes. For instance, with a set of reference faces showing nose widths from narrowest to widest, one might want to generate an image with a specific nose width. Our approach, Att-Adapter, enables precise attribute adjustments using numeric values derived from domain-specific data (e.g., nose width ranges), allowing more accurate modifications within targeted domains.

C. Application results: LoRA

LoRA can be used for personalizing pretrained Diffusion Models [3, 4, 13]. In this section, we show that Att-Adatper can be combined with the LoRA module that is finetuned for the appearance of a specific individual. Briefly, we used 22 images of a certain celebrity (Jennifer Lawrence) to finetune LoRA. After finetuning, we combine LoRA and Att-Adapter which are separately trained. The results are shown in Fig. 1. We can see that Att-Adapter can adjust the facial components of the generated image while the person identity is heavily affected by LoRA module. This shows that the wide applicability of Att-Adapter.

D. Exploring training settings.

In this section, we explore some of the important factors that could be needed during the training process for Att-Adapter.

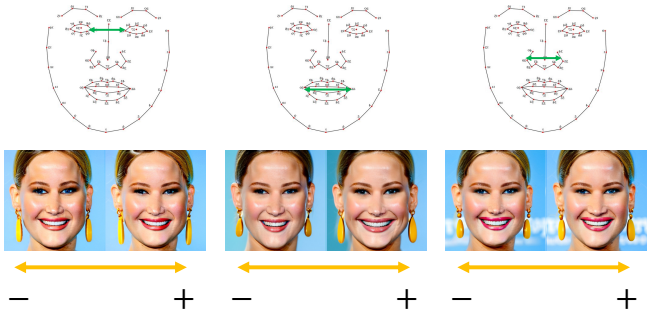

Figure 1. Qualitative experiments showing that LoRA and Att-Adapter can be combined during the sampling process.

Training iteration First we analyze the quantitative performance per iteration which is shown in Table 1. We can see the trade-off between the finetuned knowledge and pretrained knowledge as iteration goes on. At 100k, we can better maintain the pretrained knowledge. However, the performance w.r.t. finetuned knowledge is not good. After 200k, even though we lose some pretrained knowledge, we get better scores for the finetuned knowledge. It is not easy to determine which checkpoint is better among 200k and 300k as their performance gap is not conspicuous. During the experiments, we empirically used the model saved at 200k.

	Finetuned knowledge (↓)			Pretrained knowledge (↑)	
(Iter.)	Facial comp.	Age	Race	CLIP	ChatGPT
100k	21.1	9.8	2.48	**0.285**	**82%**
200k	**15.57**	8.0	**0.36**	0.283	78%
300k	15.92	**7.7**	**0.36**	0.282	77%

Table 1. Performance per iterations. We observe that the performance is empirically converged around 200k.

Embedding dimension of Z. Second, we explore the impact of dimensionality of Z on performance in Table 2. Interestingly, with 128 dimension, Att-Adapter tend to learn less the finetuned knowledge while maintaining more the pretrained knowledge. With 512 dimension, the model shows decent performance in both finetuned and pretrained knowledge. With 2048 dimension, the model shows comparable performance with the 512 setting; The age prediction gets slightly better, but slightly worse in most of the other measures.

Non-isotropic gaussian prior. Lastly, we compare the performance of using isotropic/non-isotropic gaussian for prior distribution. For isotropic gaussian, we estimate a scalar. For non-isotropic setting, we estimate a vector (i.e., the diagonal term of the covariance matrix.) The results are shown in Table 3. At 100k, we can see that the non-isotropic

	Finetuned knowledge (↓)			Pretrained knowledge (↑)	
(Dim.)	Facial comp.	Age	Race	CLIP	ChatGPT
128	20.54	9.3	0.808	**0.285**	**83%**
512	**17.42**	8.5	**0.799**	0.282	80%
2048	17.53	**7.7**	0.987	0.282	75%

Table 2. Performance comparisons given different Z dimensions. For each setting, the checkpoint saved at 150k iterations is used.

setting is better than the isotropic setting for learning the finetuned knowledge. We conjecture that this is because the higher degree of freedom of non-isotropic gaussian (than the isotropic gaussian) could be beneficial at fitting at some points. At 200k and 300k, however, the difference gets smaller and both settings become comparable. We empirically used the isotropic gaussian prior setting for our main experiments.

	Finetuned knowledge (↓)			Pretrained knowledge (↑)	
(Iter.)	Facial comp.	Age	Race	CLIP	ChatGPT
100k	**-0.39**	**-0.4**	**-0.663**	0.0	0%
200k	+0.39	**-0.2**	+0.207	-0.001	-1%
300k	**-0.23**	**-0.1**	+0.076	0.0	**+2%**

Table 3. Performance comparisons of two settings; 1. isotropic gaussian for the prior, 2. non-isotropic gaussian for the prior distribution. The values in the table are obtained by subtracting the values of the second setting from the values of the first setting, i.e., (informally) iso value − noniso value.

E. Additional comparison of Att-Adapter and LoRA

The better performance of Att-Adapter over the baseline in race attribute can be found in Fig. 2. This can be observed by comparing the fourth and the sixth columns of (a), which shows that LoRA is confused of generating White and Hispanic.

Fig. 3 shows the advantage of our method straightforward. Each column shows the result from the different conditioning value for the given attribute. For each macro column, we can observe from the center two rows that both Att-Adapter and LoRA show good performance given the within-domain conditioning values, i.e., $[0, 1]$. However, only Att-Adapter can deal with the negative or greater-than-one conditioning. This is because LoRA is only trained with dealing with the discretized and tokenized string identifiers of 0,1,...,9. For example, given -0.55 from the leftmost column, our preprocessor for discretizing makes the value -5. We guess LoRA ignores '-' sign, and '5' is taken, which yields the eyes openness to the similar extent with the third column (i.e., 0.59). Similarly, given 1.10 in the fourth column from the left, our discretizer makes it 11, which becomes '1' and '1' after tokenized. We can see that LoRA recognizes the two '1's as '1' by comparing the eye openness with the second column (i.e., 0.14, 1 after discretized). On the other hand, as shown in the first row, Att-Adapter can extrapolate to the attribute values beyond $[0, 1]$, to the unseen domain.

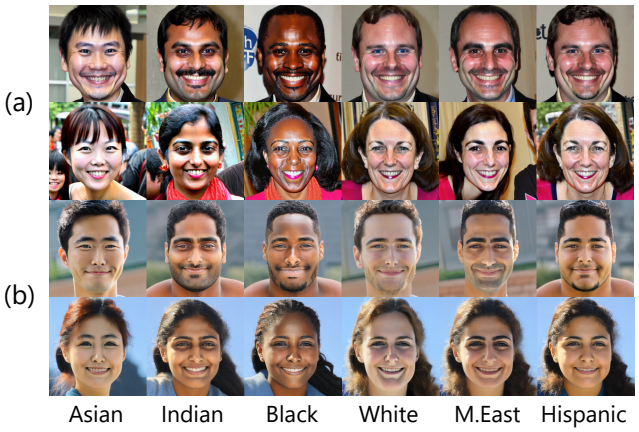

Figure 2. Qualitative Baseline comparisons on race. From the top, (a) LoRA, and (b) Att-Adapter. The prompts of "A photo of a woman with smiling" and "A photo of a man with shadow fade hair style" are used for the woman images and the man images.

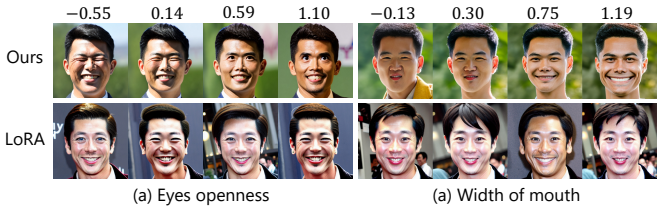

Figure 3. Extrapolation comparisons with LoRA showing the strength of Att-Adapter. A prompt of "A stylish man smiling" is used with 'Asian' condition.

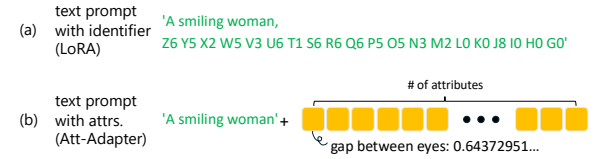

Figure 4. Visualization of the input setting of LoRA baseline. All the continuous attributes are discretized and named as special tokens. For example, the attribute 'gap between eyes' and its value 0.64 becomes 'Z6'. Multi-attributes are linearly converted and added consecutively.

	Finetuned knowledge (↓)			Pretrained knowledge (↑)	
(λ.)	Facial comp.	Age	Race	CLIP	ChatGPT
0.0	69.43	8.8	1.3	**0.298**	**97%**
0.5	28.96	8.5	0.50	0.293	94%
0.7	17.96	**8.0**	0.38	0.288	89%
0.9	**15.57**	**8.0**	**0.36**	0.283	78%

Table 4. Performance reports with different λs. The lower the λ is, the stronger the pretrained knowledge affects the generation process while relatively weakening the influence of Att-Adapter.

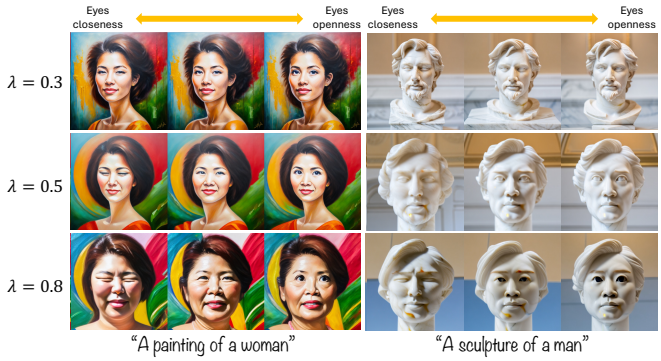

Figure 5. Qualitative explorations on the effects of λ.

F. Additional Analysis

F.1. λ during the sampling

As shown in Eq. 2 in the main paper, λ controls the strength of Att-Adapter when merging the information from different input conditions. Intuitively, by reducing λ, we can expect that Att-Adapter will affect the generation process less. This can be verified by looking at Fig. 5. When $\lambda = 0.3$, we can see that the effect of text prompt is strong while the attribute effect is weak. For example, the sculptures does not look like the given attribute, 'Asian'. When $\lambda = 0.8$, the effect of Att-Adapter (where the finetuning knowledge passes through) gets stronger as we can see that the results in the third row have more human-like texture and face-cropped view. The quantitative results can be found in Table. 4. We can first see that pretrained knowledge, measured by CLIP and ChatGPT is maintained better with lower λ. The higher the λ is, on the other hand, we can expect the stronger effects of the fine-tuned knowledge, in exchange for losing the pretrained knowledge.

F.2. Robustness for correlated attributes.

As shown in Table 2 in the main paper, Att-Adapter outperforms baselines in disentangling attributes due to joint attribute learning. To provide further evidence, we performed additional analysis on correlated attributes such as "mouth width" and "eye openness", which typically co-vary through "smiling" (wider mouth coinciding with narrower eyes). As shown in Fig. 6 left, Att-Adapter can manipulate

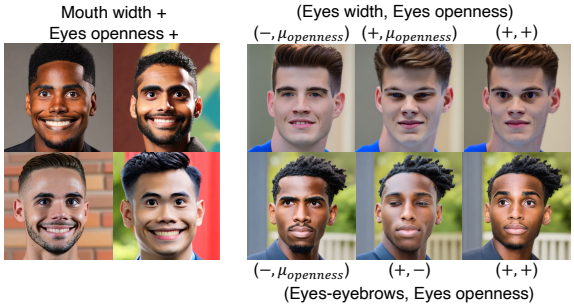

Figure 6. Disentanglement of correlated attributes by Att-Adapter.

them independently (wider mouth + and wider eye openness +). More examples showing independent controls over the correlated attributes can be seen in Fig. 6 right.

F.3. Scalability validation

We validated scalability by measuring resource usage with increasing number of attributes. With a batch size of 16 and latent dimension $Z \in \mathbb{R}^{1024}$, using one attribute requires around 23,704MB of VRAM and 97MB of storage. Each additional attribute increases VRAM by only about 338MB (1.4% of the initial VRAM) and storage by just 0.004MB (0.004% of the initial storage). These results support our claim of handling numerous attributes with negligible increases in memory usage.

F.4. Ablation Study (CVAE v.s. Dropout Regularization)

We explore the effect of dropout in preventing overfitting by conducting additional comparisons. Results in the Table below confirm using dropout (the best rate of 0.25 is reported) indeed reduces identity similarity which indicates improved regularization. However, even at this optimal dropout rate, performance remains worse than our CVAE-based approach. These confirms the effectiveness of CVAE in regularizing Att-Adapter and prevent overfitting.

	Naive		
	w.o. drop	with drop	Ours
ID sim (↓)	28.5%	23.6%	**6.2%**

F.5. Challenging attributes

In this section, we show that Att-Adapter can be used beyond frontal-view images. In order to show this, we extend our training dataset from the frontal-view face images to the entire FFHQ dataset. We also additionally add three attributes; yaw, pitch, and roll[3]. The results are shown in Fig. 7. Interestingly, even though Att-Adapter is not specifically designed for understanding 3D domain, we can see that the left-right rotation (i.e., yaw) and the up-down rotation (i.e., pitch) can be controlled. However, we observed that roll is not controllable. To improve this, we believe

[3]https://github.com/DCGM/ffhq-features-dataset

additional facial dataset with diverse roll information is required as FFHQ face-cropped dataset is face aligned. We also think it would be interesting and powerful if 3D domain knowledge could be aggregated in Att-Adapter which is beyond our research scope.

Additionally, we show that Att-Adapter can control two attributes simultaneously in Fig. 8 and Fig.9.

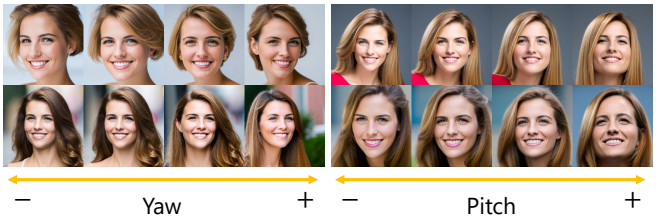

Figure 7. Qualitative results on additional attributes control. The results are generated by taking a prompt of "A smiling woman".

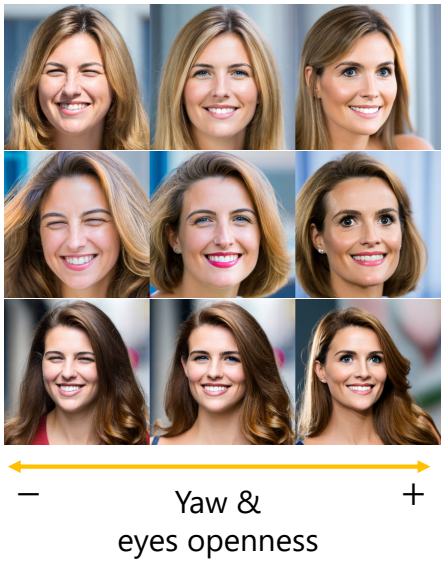

Figure 8. Additional two-attributes controlling examples. A prompt of "A smiling woman" is used.

Figure 9. Additional two-attributes controlling examples. A prompt of "A smiling woman" is used.

F.6. Dependency on Quality of Attribute Annotations and Paired v.s. Unpaired Data Performance

Att-adapter is specifically designed for scenarios where explicit paired data is unavailable but attribute annotations are present. Such scenarios frequently occur in practice, for example, product images are associated with metadata and manual annotations. In contrast, other methods like ConceptSlider can be effective in scenarios lacking explicit attribute annotations by relying on paired data. Adapting Att-Adapter to such scenarios is nontirivial issue, which can be an interesting future direction.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 1

[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. *arXiv preprint arXiv:2208.01618*, 2022. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[5] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. Underspecification in scene description-to-depiction tasks. *arXiv preprint arXiv:2210.05815*, 2022. 2

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[7] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. 1

[8] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[9] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[13] 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. 2

[14] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[15] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 1

[16] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2

[17] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 2

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2