Supplementary Materials: ITI-GEN: Inclusive Text-to-Image Generation

Cheng Zhang [1] Xuanbai Chen [1] Siqi Chai [1] Chen Henry Wu [1] Dmitry Lagun [2]
Thabo Beeler [2] Fernando De la Torre [1]

[1] Carnegie Mellon University [2] Google

Contents

A. Ethical and Social Impacts

One important consideration is the potential impact on privacy and data protection. In order to generate inclusive images, ITI-GEN relies on reference images that are often sourced from publicly available datasets. However, the utilization of these images raises concerns about privacy and the potential for unintended consequences, such as the misuse of personal data. It is crucial to consider ways to mitigate these risks, such as data anonymization or obtaining explicit consent from individuals whose images are used.

While ITI-GEN's directional loss avoids directly measuring the distance between the prompts and the reference images, it is possible that the reference images used to represent certain attributes may themselves contain biases or inaccuracies. To address this concern, it will be important to carefully evaluate the quality and representativeness of the reference images used in the model and to develop strategies for identifying and correcting biases when they arise.

Inclusive image generation has the potential to promote greater representation and diversity in various industries, which could in turn promote greater social equality and reduce discrimination. However, it is also possible that the technology could be misused or weaponized to promote negative or harmful stereotypes. Therefore, it will be important to consider the potential risks and benefits of ITI-GEN carefully for mitigating negative outcomes.

B. Additional Related Work and Comparisons

In this section, we provide a more comprehensive comparison between ITI-GEN and related methods.

Bias Mitigation Methods in Text-to-Image Generation. As mentioned in Section 2 of the main paper, ITI-GEN uses images as guidance while existing approaches focus on debiasing the prompts. Two concurrent works, Prompt Debiasing [7] and Fair Diffusion [11] require the category names of the target attributes for learning fair prompts. However, we argue that, for some attributes, attribute names might be hard to specify using language (*e.g.*, skin tone, different levels of brightness). ITI-GEN learns tokens without gradient propagation through the original text-to-image models, making it more efficient in both training and deployment.

Personalization. Both ITI-GEN and personalized text-to-image generation methods [18, 12] are inspired by prompt tuning [15, 20]. However, they are fundamentally different, as introduced in Section 2 of the main paper. We compare with custom diffusion [18] in Table 1 of the main paper mainly to provide a justification for whether the personalization methods [18, 25, 12] can be used in inclusive text-

to-image generation. Specifically, we attempt to provide different numbers of reference images for Custom Diffusion [18] and select the best results to report. Moreover, unlike personalization methods that use diffusion losses to train the special tokens, the tokens learned by ITI-GEN are generalizable between different models.

Disentanglement. It is worth mentioning that the aggregation of multiple inclusive tokens learned with separate reference datasets in marginal distributions can implicitly disentangle attribute learning. However, we emphasize that the primary goal of ITI-GEN is *not* to achieve feature (or attribute) disentanglement [17]. Please see Section 4.3 and Figure 11 of the main paper for a detailed discussion.

Image-to-image Translation and Editing. As mentioned in Section 3.3 of the main paper, the goal of our work is to promote inclusiveness or diversity but not for image editing. In image-to-image translation or editing tasks, it is required to edit the desired attribute while keeping other features of the image intact. However, we *do not* have such a requirement for ITI-GEN. For example, in Figure 4, Figure 7, Figure 10, and Figure 11 of the main paper, while there are subtle changes to the clothing or background in the images, ITI-GEN *already achieves inclusiveness for the intended attribute. We show examples with the same random seeds in these figures mainly for a better comparison.*

C. Reference Images Preparation

In this section, we provide more details on the construction of reference image sets to complement Section 4.1 of the main paper. We use the following datasets as resources.

CelebA [21] is a benchmarked face attributes dataset and each image with 40 binary attribute annotations. We experiment with these binary attributes and their combinations.

FAIR Benchmark (FAIR) [10] is a recently proposed synthetic face dataset used for skin tone estimation. Specifically, we use images from the validation set containing 234 images and 702 facial crops. The validation set is released with ground-truth UV albedo maps. In order to obtain ground-truth skin tone types, we follow [10] to compute the Individual Typology Angle (ITA) score [5] of an albedo map to be the average of all pixel-wise ITA values with a pre-computed skin region area. For each image, ITA can be used to classify the skin tone type according to 6 categories, ranging from very light (*i.e.*, type 1) to dark (*i.e.*, type 6) [5, 8]. We randomly select 25 images per skin tone type as the reference images.

FairFace [16] contains face images with annotations for 2 perceived gender, 9 perceived age, and 7 race categories. As discussed in Section 4.1 of the main paper, although we value the contribution of the FairFace database to the community, we prefer using race labels and instead advocate for skin tone descriptions that recognize phenotypic diversity within broad racial categories [3]. Therefore, we only use their age annotations in our experiments.

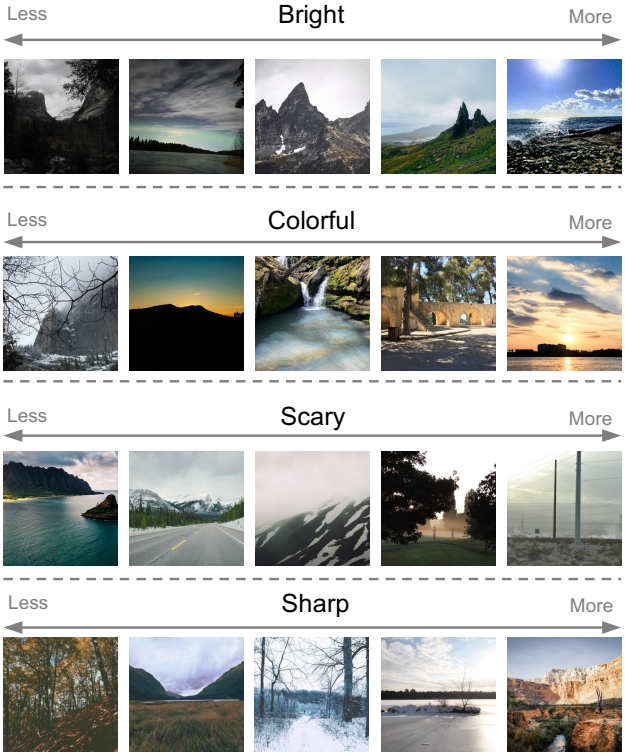


Figure A1. **Examples of reference images from LHQ [26].** We show randomly picked images for four attributes. Images within each category are classified into one of five groups.

Landscape (LHQ) [26] provides unlabeled natural scene images, allowing us to extend ITI-GEN to a different domain beyond human faces. With the annotation tool from [28], each image can be labeled with a score s ranging from 0 to 1, with a higher value indicating a closer match to the corresponding attribute. Using this score, we classify each image into one of the five degrees of the target attribute, resulting in a multi-category attribute. Figure A1 shows example reference images in the LHQ dataset. *Note that, the purpose of this experiment is not to justify LHQ as a perfect resource for learning tokens for perception attributes, but to investigate the capability of our* ITI-GEN *framework that can leverage the data from another domain as guidance.*

D. Evaluation Metrics.

Distribution Discrepancy ($\mathbb{D}_{\text{KL}}$). Following [6, 7], we use the CLIP model to predict the attributes in the images. For the attributes in which every category can be accurately specified by natural language, we input the original prompt combined with different names of categories into CLIP for obtaining the attribute label. For instance, if we

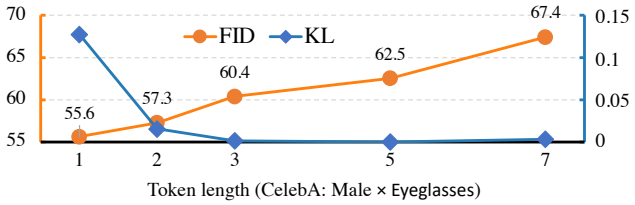


Figure A2. **Ablation study on tokens length.** Using fewer tokens is not sufficient enough to capture the concepts of attributes, leading to a relatively high distribution discrepancy (*i.e.*, KL). On the other hand, using more tokens may degrade the image quality due to language drifts (*i.e.*, relatively high FID scores).

want to evaluate the attribute "male" for the images generated from "a headshot of a person", we construct the input text of CLIP as ["a headshot of a man", "a headshot of a woman"]. For the attributes in which some of the categories can not be specified by natural languages, such as "eyeglasses" and "without eyeglasses" (due to the issue of negative prompt), we input the text ["a headshot of a person with eyeglasses", "a headshot of a person"]. For attributes that CLIP might be erroneous, we leverage pre-trained classifiers [16] combined with human evaluations. Specifically, for the skin tone, which is extremely difficult to obtain an accurate scale [1, 2, 14], we adopt the most commonly used Fitzpatrick skin type [5] combined with off-the-shelf models [10] for evaluation.

Fréchet Inception Distance (FID) [13]. We report the FID score to measure image quality. Specifically, we use the CleanFID library [22] to calculate the FID relates to statistics in FFHQ [17].

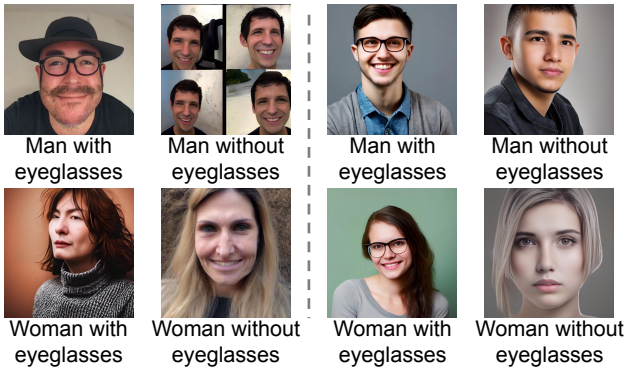
E. Additional Ablations and Analyses

E.1. Tokens Length

In our experiments, we set the length of inclusive tokens as 3 (q in Equation 3 of the main paper). Here, we provide further analyses on the size of q and show results in Figure A2. We see that fewer than 3 tokens may hurt the performance — cannot generate images with the desired attributes — potentially due to less representation capacity in capturing the concepts in the reference images. On the other hand, more tokens may result in adversarial effects or collapse. We hypothesize that prepending too many tokens after the original prompts leads to language drifts [19, 25]. This cannot be alleviated even with the semantic consistency loss (Equation 7 of the main paper) because simply forcing the two prompts with very different lengths to be close in the embedding space is ineffective.

E.2. Tokens Aggregation

As mentioned in Section 3.1 of the main paper, we use summation operation to aggregate the inclusive tokens of



(a) Concatenation **(b) Summation**

Figure A3. **Concatenation *vs.* summation on inclusive tokens aggregation.** We show an example of the combination of "Male" and "Eyeglasses" attributes. **(a)** Simply concatenating may reduce the image quality or fail to generate the images with corresponding attributes (*e.g.*, "Woman with eyeglasses") potentially because of the language drifts [19, 25]. **(b)** ITI-GEN provides better results with a conceptually simpler summation.

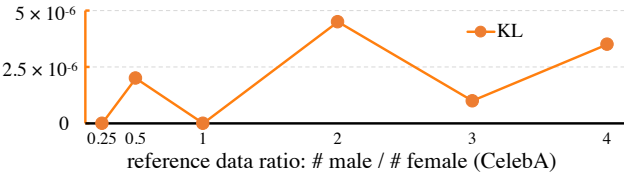


Figure A4. **Ablation study on the ratio of different categories in the reference set.** We study on the perceived gender attribute in CelebA by changing the ratio of images from the "male" and "female" categories. ITI-GEN is robust (*i.e.*, with very small distribution discrepancy, KL) to the ratio of different categories in the reference image set.

multiple attributes to achieve permutation invariance. Here, we provide another option — concatenation. Specifically, we ignore the positional encodings before feeding the inclusive tokens in the CLIP text encoder. Thus, the attention mechanism applied to prompt tokens is permutation invariant. Figure A3 shows comparison results. We notice that ITI-GEN (with token summation) not only achieves better results than concatenation but also offers a simpler and cleaner solution for token aggregation.

E.3. Imbalanced Reference Images

As mentioned in Section 4.1 of the main paper, we select 25 reference images per category in our experiments. We also mentioned that ITI-GEN is robust to imbalanced data distributions in Section 3.3. Here, we provide additional results as evidence. We change the ratio of "male" images *vs.* "female" images for the Perceived Gender attribute in CelebA and show the results in Figure A4. ITI-GEN can always generate images with nearly a balanced distribution.

Figure A5. Results of ITI-GEN with (a) mutually exclusive and (b) overlapped reference images for attributes: *gender*×*eyeglasses*.

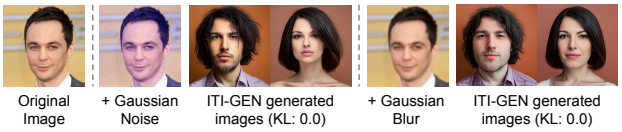


Figure A6. ITI-GEN with corrupted reference data. The attribute of interests is *gender*.

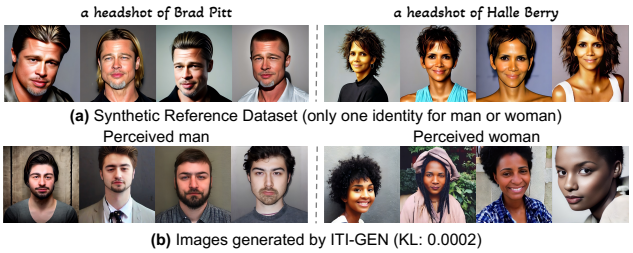


Figure A7. ITI-GEN can leverage less diverse reference images in (a) for inclusive generation for the *gender* attribute in (b).

E.4. Overlapped Reference Images

As mentioned in Section 3.2 of the main paper, we need a reference dataset for each attribute. However, this does not pose a practical issue (which seems like an all-too-exhaustive list to cover), because each reference dataset does not have to be mutually exclusive. An existing dataset (*e.g.*, CelebA or smaller) can be divided into overlapped sub-datasets, either manually or using a classifier. To demonstrate this, we compare two settings: (a) *Exclusive* — two datasets, each containing 50 images with equal gender and eyeglasses distribution, respectively; (b) *Overlapped* — a single dataset of 50 images with equal numbers between man and woman labels, as well as with and without eyeglasses. The results in Figure A5 show that using a smaller, *overlapped* dataset does not affect the performance.

E.5. Corrupted Reference Images

In this subsection, we further study whether the quality of the provided reference image strongly affects the generalization and the application of the ITI-GEN. We provide the results with noisy or blurred reference images in Figure A6. We also experiment with less diverse reference images (only using the images with one identity) and show results in Figure A7. Both demonstrate the robustness of ITI-GEN to the quality of reference data.

Figure A8. **Qualitative results of ITI-GEN when $K_m = 1$.** When only "female" images are provided as the reference images (left in (a)), ITI-GEN can leverage the synthetic data generated by the original prompt ("*a headshot of a person*", right in (a)), together with the real data, to construct the reference image set. By jointly using these two sources, ITI-GEN learns inclusive tokens representing the concept of "female", which can be used to synthesize images for the desired category, as shown in (b). Section E.6 illustrates details.

E.6. Single Category Attribute ($K_m = 1$)

In the main paper, we mainly studied the attributes that have more than one category (K_m is larger than 1 in Equation 3 of the main text). What if we only have the reference images from one category of the target attribute ($K_m = 1$)? In light of our pairwise direction loss (Equation 4), there are at least two different categories needed in the reference images. Here, we show that ITI-GEN can leverage the synthetic data generated by the original prompt (*e.g.*, "*a headshot of a person*") as an additional category to compute the directional loss for the case of $K_m = 1$.

We verify this idea by using only images of the "female" category from the perceived gender attribute. From Figure A8, we can observe that by leveraging the female real images and another set of synthetic images generated from "*a headshot of a person*", ITI-GEN is able to synthesize female images. Further quantitative evaluation for the images generated by ITI-GEN indicates that 100% perceived woman is obtained.

F. Additional Results

Due to space limitations, we only reported the results of several attributes mainly to cover the attributes relating to social factors and facial expressions in the main paper. In this section, we provide additional results and detailed comparisons to strong baseline methods.

Table A1. **FID (↓) comparison**. Reference images for ITI-GEN are from FAIR benchmark [10]. ITI-GEN produces lower FID than all the other baselines. **SD**: vanilla stable diffusion. **EI**: ethical intervention. **HPS**: hard prompt searching. **PD**: prompt debiasing. **CD**: custom diffusion.

SD [24]	EI [4]	HPS [9]	CD [18]	PD [7]	**ITI-GEN**
67.4	81.4	69.9	62.4	63.3	**51.8**

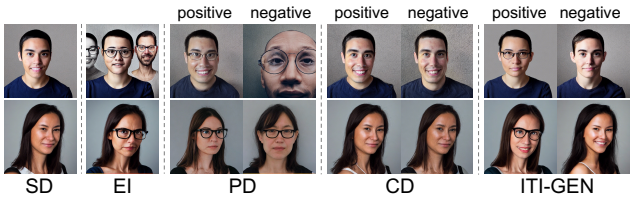


Figure A9. **Visualization of different methods.** The prompt is *"a headshot of a person"*. Attributes are *gender* × *eyeglasses*. Images across each line are sampled using the same random seed.

F.1. Qualitative and FID Results for Baselines

We only provide KL Divergence metric ($\mathbb{D}_{\text{KL}}$) in the main paper for different baselines. Here, we incorporate the comparisons of FID in Table A1 and visualizations in Figure A9 with other baselines.

F.2. Single Binary Results

We summarize the full results of single binary attributes with CelebA [21] in Table A2. We compare with the baseline Stable Diffusion model [24] and Hard Prompt Searching [9], which demonstrated strong performance in many attributes (cf. Table 1 of the main paper). From Table A2, we observe ITI-GEN achieves the best performance in nearly all 40 attributes except some subtle facial attributes (*e.g.*, "Wearing Necklace"). We use the prompt "*a headshot of a person*" in Table A2 and show qualitative results of other prompts (*e.g.*, other occupations such as politician and musician) in Figure A21. Furthermore, we list all the hard prompts used in our experiments in Figure A3.

As mentioned in Section 1 of the main paper, we reiterate that ITI-GEN is designed to handle several cases (attributes) that Hard Prompts may struggle with. First, attributes with fine-grained categories may be difficult to express in language. Second, linguistic ambiguity, as shown in Figure A10 (a). Third, model misrepresentation, as illustrated in Figure A10 (b). More importantly, we argue that ITI-GEN is *not* to replace Hard Prompts (especially for attributes that are already can be handled by language) but to support complex prompts with multiple attributes, as illustrated in Figure A11.

F.3. Multiple Attributes

We now consider multi-attribute cases and show additional results in Figure A12. To fully characterize the performance of ITI-GEN, we study three additional settings based on the attribute correlation matrix from the CelebA dataset [21] (see Figure 2 in [27] for the correlation matrix).

Table A2. **A full comparison with baseline methods with the 40 single attribute setting** ($\mathbb{D}_{\text{KL}}$ ↓). Reference images are from CelebA [21]. Following [6, 7], we use CLIP [23] as the attribute classifier. **SD**: vanilla stable diffusion [24]. **HPS**: hard prompt searching [9]. Given the strong capability of the existing text-to-image generative models, one can express the (most but not all) desired attributes directly using *Hard Prompts*. However, it faces challenges in certain attributes and ITI-GEN addresses most of these drawbacks. Please see Figure A10 for a side-by-side qualitative comparison between HPS and ITI-GEN. Please see Figure A11 for how ITI-GEN can be compatibly used with Hard Prompts.

Attribute	SD [24]	HPS [9]	ITI-GEN
5'o Clock Shadow	0.02957	**0.00847**	0.06882
Arched Eyebrows	0.32972	0.04570	**0.00892**
Attractive	0.11264	0.07405	**0.00000**
Bags Under Eyes	0.33325	0.10498	**0.01395**
Bald	0.51578	0.22175	**0.00892**
Bangs	0.33886	0.19975	**0.00000**
Big Lips	0.20984	0.02908	**0.00892**
Big Nose	0.32423	0.01629	**0.00056**
Black Hair	0.35189	0.12539	**0.00000**
Blond Hair	0.60804	0.00501	**0.00222**
Blurry	**0.01077**	0.25348	0.09707
Brown Hair	0.41683	0.14207	**0.05663**
Bushy Eyebrows	0.07108	0.29737	**0.02747**
Chubby	0.14293	0.40233	**0.00000**
Double Chin	0.28637	0.48016	**0.19274**
Eyeglasses	0.38773	0.32622	**0.00056**
Goatee	0.25933	0.04266	**0.00000**
Gray Hair	0.65905	0.27921	**0.17049**
Heavy Makeup	0.39293	0.10989	**0.04570**
High Cheekbones	0.47875	**0.00020**	0.03599
Male	0.01033	0.00005	**0.00000**
Mouth Slightly Open	0.07030	0.14207	**0.04570**
Mustache	0.02013	0.12009	**0.00000**
Narrow Eyes	0.14968	**0.00847**	0.08228
No Beard	0.22442	0.49463	**0.00222**
Oval Face	0.39526	0.03158	**0.02014**
Pale Skin	0.17394	0.00045	**0.00000**
Pointy Nose	0.48951	0.02221	**0.00000**
Receding Hairline	0.31784	0.61526	**0.02014**
Rosy Cheeks	0.46275	**0.03691**	0.14987
Sideburns	0.55409	0.04570	**0.02013**
Smiling	0.25059	0.02075	**0.00000**
Straight Hair	**0.08506**	0.61526	0.19274
Wavy Hair	0.47663	0.36806	**0.03599**
Wearing Earrings	0.32029	0.15998	**0.09707**
Wearing Hat	0.65144	0.12539	**0.01395**
Wearing Lipstick	0.50658	0.12539	**0.11323**
Wearing Necklace	0.63600	**0.05897**	0.49463
Wearing Necktie	0.46687	0.69315	**0.06882**
Young	0.65647	0.00056	**0.00000**

Table A3. **Hard Prompts used in our experiments.** Different attributes may not follow the same template and we carefully specify or express the attribute in the input prompt. The human-written hard prompts are used to generate images. Results are shown in Table A2.

Attribute	Positive Prompt	Negative Prompt
5'o Clock Shadow	a headshot of a person {with 5'o clock shadow}.	a headshot of a person {with no/without 5'o clock shadow}.
Arched Eyebrows	a headshot of a person {with arched eyebrows}.	a headshot of a person {with no/without arched eyebrows}.
Attractive	a headshot of {an attractive} person.	a headshot of {an unattractive} person.
Bags Under Eyes	a headshot of a person {with bags under eyes}.	a headshot of a person {with no/without bags under eyes}.
Bald	a headshot of a person {in bald}.	a headshot of a person {in no bald}.
Bangs	a headshot of a person {with bangs}.	a headshot of a person {with no/without bangs}.
Big Lips	a headshot of a person {with big lips}.	a headshot of a person {with small lips}.
Big Nose	a headshot of a person {with big nose}.	a headshot of a person {with small nose}.
Black Hair	a headshot of a person {with black hair}.	a headshot of a person {with no/without black hair}.
Blond Hair	a headshot of a person {with blond hair}.	a headshot of a person {with no/without blond hair}.
Blurry	a headshot of a person {in blurry}.	a headshot of a person {in no/without blurry}.
Brown Hair	a headshot of a person {with brown hair}.	a headshot of a person {with no/without brown hair}.
Bushy Eyebrows	a headshot of a person {with bushy eyebrows}.	a headshot of a person {with no/without bushy eyebrows}.
Chubby	a headshot of a {chubby} person.	a headshot of a {no chubby} person.
Double Chin	a headshot of a person {with double chin}.	a headshot of a person {with no/without double chin}.
Eyeglasses	a headshot of a person {with eyeglasses}.	a headshot of a person {with no/without eyeglasses}.
Goatee	a headshot of a person {with goatee}.	a headshot of a person {with no/without goatee}.
Gray Hair	a headshot of a person {with gray hair}.	a headshot of a person {with no/without gray hair}.
Heavy Makeup	a headshot of a person {with heavy makeup}.	a headshot of a person {with no/without heavy makeup}.
High Cheekbones	a headshot of a person {with high cheekbones}.	a headshot of a person {with low cheekbones}.
Male	a headshot of a {man}.	a headshot of a {woman}.
Mouth Slightly Open	a headshot of a person {with mouth slightly open}.	a headshot of a person {with mouth closed}.
Mustache	a headshot of a person {with mustache}.	a headshot of a person {with no/without mustache}.
Narrow Eyes	a headshot of a person {with narrow eyes}.	a headshot of a person {with no/without narrow eyes}.
No Beard	a headshot of a person {with no/without beard}.	a headshot of a person {with beard}.
Oval Face	a headshot of a person {with oval face}.	a headshot of a person {with no/without oval face}.
Pale Skin	a headshot of a person {with pale skin}.	a headshot of a person {with dark skin}.
Pointy Nose	a headshot of a person {with pointy nose}.	a headshot of a person {with no/without pointy nose}.
Receding Hairline	a headshot of a person {with receding hairline}.	a headshot of a person {with no/without receding hairline}.
Rosy Cheeks	a headshot of a person {with rosy cheeks}.	a headshot of a person {with no/without rosy cheeks}.
Sideburns	a headshot of a person {with sideburns}.	a headshot of a person {with no/without sideburns}.
Smiling	a headshot of a person {with smiling}.	a headshot of a person {with no/without smiling}.
Straight Hair	a headshot of a person {with straight hair}.	a headshot of a person {with no/without straight hair}.
Wavy Hair	a headshot of a person {with wavy hair}.	a headshot of a person {with no/without wavy hair}.
Wearing Earrings	a headshot of a person {wearing earrings}.	a headshot of a person {without wearing earrings}.
Wearing Hat	a headshot of a person {wearing hat}.	a headshot of a person {without wearing hat}.
Wearing Lipstick	a headshot of a person {wearing lipstick}.	a headshot of a person {without wearing lipstick}.
Wearing Necklace	a headshot of a person {wearing necklace}.	a headshot of a person {without wearing necklace}.
Wearing Necktie	a headshot of a person {wearing necktie}.	a headshot of a person {without wearing necktie}.
Young	a headshot of a {young} person.	a headshot of {an old} person.

Specifically, we select three attribute combinations with different levels of attribute entanglement (*i.e.*, co-occurrence frequency) — a higher co-occurrence value means the attribute combination is more common in daily life while a lower co-occurrence value indicates a rare case in the original CelebA dataset. Admittedly, there are several cases ITI-GEN does not always generate images with a balanced distribution or faithfully generates images with specific attributes. Please see Figure A12 for details.

F.4. Multi-Category Attributes

In Figure 6 and Figure 7 of the main paper, we investigated the combinations of multi-category attributes. Here, we further study another challenging setup: Perceived Gender (CelebA) $\times$ Skin Tone (FAIR) $\times$ Age (FairFace) (108 different combinations of categories in total). Qualitative results are shown in Figure A13 and in Figure A14. As expected, ITI-GEN is capable of handling multiple fine-grained attribute categories to achieve inclusiveness.

F.5. Other Domains

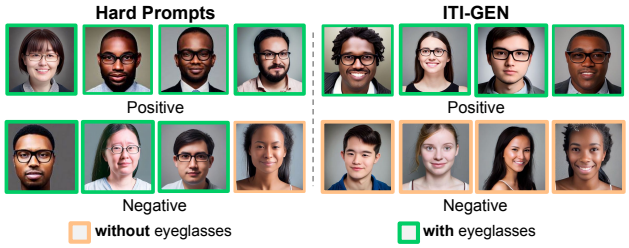
As shown in Figure 8 of the main paper, ITI-GEN can generalize to a different domain for perception attributes on scene images. In this subsection, we demonstrate more results of other attributes in Figure A15 for "colorfulness", Figure A16 for "sharpness", Figure A17 for "scary", Figure A18 for "contrast", Figure A19 for "brightness", and Figure A20 for "'brightness". As we observe, ITI-GEN generates more diverse results than the baseline model even with very complex input prompts.

F.6. Train-once-for-all Generalization

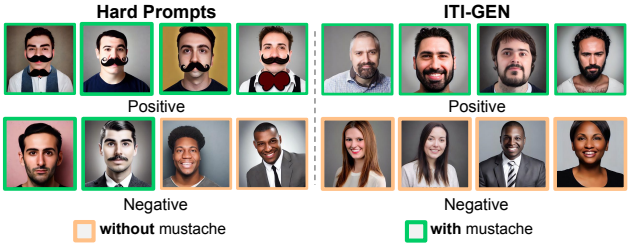
We provide additional qualitative results with different occupation prompts in Figure A21, Figure A22, Figure A23, Figure A24, and Figure A25.

F.7. Compatibility with ControlNet

We provide additional examples of compatibility with ControlNet in Figure A26.



(a) Target Attribute: eyeglasses



(b) Target Attribute: mustache

Figure A10. **Challenges of (a) linguistic ambiguity and (b) model misrepresentation.** While Hard Prompts demonstrated strong capabilities in generating images with desired attributes, they cannot handle some situations. **(a)** Vanilla text-to-image models can hardly understand *negative* prompts (*e.g.*, "not", "without") possibly due to *linguistic ambiguity*. **(b)** For some attributes (*e.g.*, mustache), directly using hand prompts results in misrepresented results caused by the model bias.

Figure A11. **Compatibility of ITI-GEN to hard prompts.** As mentioned in Section F.2 and Figure A10, Hard Prompts show accurate results with some attributes (*e.g.*, "young" and "perceived man" in the top row) but may result in misrepresented results for other attributes (*e.g.*, "mustache" in the middle row). ITI-GEN demonstrates strong compatibility with Hard Prompts to benefit a broad spectrum of attributes (bottom row).

G. Future Work

To establish the new direction and demonstrate its feasibility so that future works can easily build upon, we intentionally avoid sophisticated techniques to improve ITI-GEN in favor of simplicity and believe that additional modifications can further enhance the inclusive generative models.

Lifelong ITI-GEN. In this study, we assume all the attributes are accessible at the same time. In practice, we hope to show that ITI-GEN is capable of the continue learning setup. That is, adding new attributes while without forgetting or re-training the previous inclusive tokens.

Other Attributes. There are other attributes ITI-GEN might be able to control via appropriately prepared reference images. For example, the 3D geometry attributes such as head poses and materials such as normal and lighting.

References

[1] Gender shades. http://gendershades.org/. 3

[2] Google skin tone research. https://skintone.google/. 3

[3] Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. *arXiv preprint arXiv:2302.03629*, 2023. 2

[4] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *EMNLP*, 2022. 5

[5] Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13(4):191–208, 1991. 2, 3

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *preprint arXiv:2202.04053*, 2022. 2, 5

[7] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *preprint arXiv:2302.00070*, 2023. 1, 2, 5

[8] Sandra Del Bino and FJBJoD Bernerd. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology*, 169(s3):33–40, 2013. 2

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021. 5

[10] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *ECCV*, 2022. 2, 3, 5

[11] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 1

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An

(a) Male × Eyeglasses

(b) Male × Young

(c) Male × Heavy Makeup

Figure A12. **Additional results on multiple attributes.** We consider three settings based on the attribute co-occurrence matrix in the CelebA dataset (see Section F.3). The attribute combinations in (a) and (b) are relatively less entangled between the sub-categories whereas in (c) — a *failure* case of ITI-GEN— the category "with heavy makeup" is heavily entangled with the category "female" in CelebA, which indicates that other category combinations (*e.g.*, "man with heavy makeup") can rarely happen in our daily life. Therefore, the text-to-image model can hardly synthesize images with this underrepresented attribute combination.

image is worth one word: Personalizing text-to-image generation using textual inversion. *preprint arXiv:2208.01618*, 2022. 1

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3

[14] John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):550–560, 2021. 3

[15] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1

[16] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. In *WACV*, 2021. 2, 3, 22

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In

Figure A13. **Results of ITI-GEN on multi-category attributes** for Perceived Gender (2) × Skin Tone (6) × Age (9). Examples are randomly picked with "*a headshot of a person*" for **Perceived Man** × Skin Tone (6) × Age (9). Please see Figure A14 for more results on Perceived Woman × Skin Tone (6) × Age (9).

CVPR, 2019. 2, 3

[18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *preprint arXiv:2212.04488*, 2022. 1, 2, 5

[19] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *preprint arXiv:1909.04499*, 2019. 3

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 1

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 5

[22] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 3

[23] 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 *ICML*, 2021. 5

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5

[25] 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 *CVPR*, 2022. 1, 3

[26] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *ICCV*, 2021. 2, 11, 12, 13, 14, 15, 16, 22

[27] Robert Torfason, Eirikur Agustsson, Rasmus Rothe, and Radu Timofte. From face images and attributes to attributes.

Figure A14. **Results of ITI-GEN on multi-category attributes** for Perceived Gender (2) × Skin Tone (6) × Age (9). Examples are randomly picked with "*a headshot of a person*" for **Perceived Woman** × Skin Tone (6) × Age (9). Please see Figure A13 for more results on Perceived Man × Skin Tone (6) × Age (9).

In *ACCV*, 2017. 5

[28] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2

[29] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *preprint arXiv:2302.05543*, 2023. 22

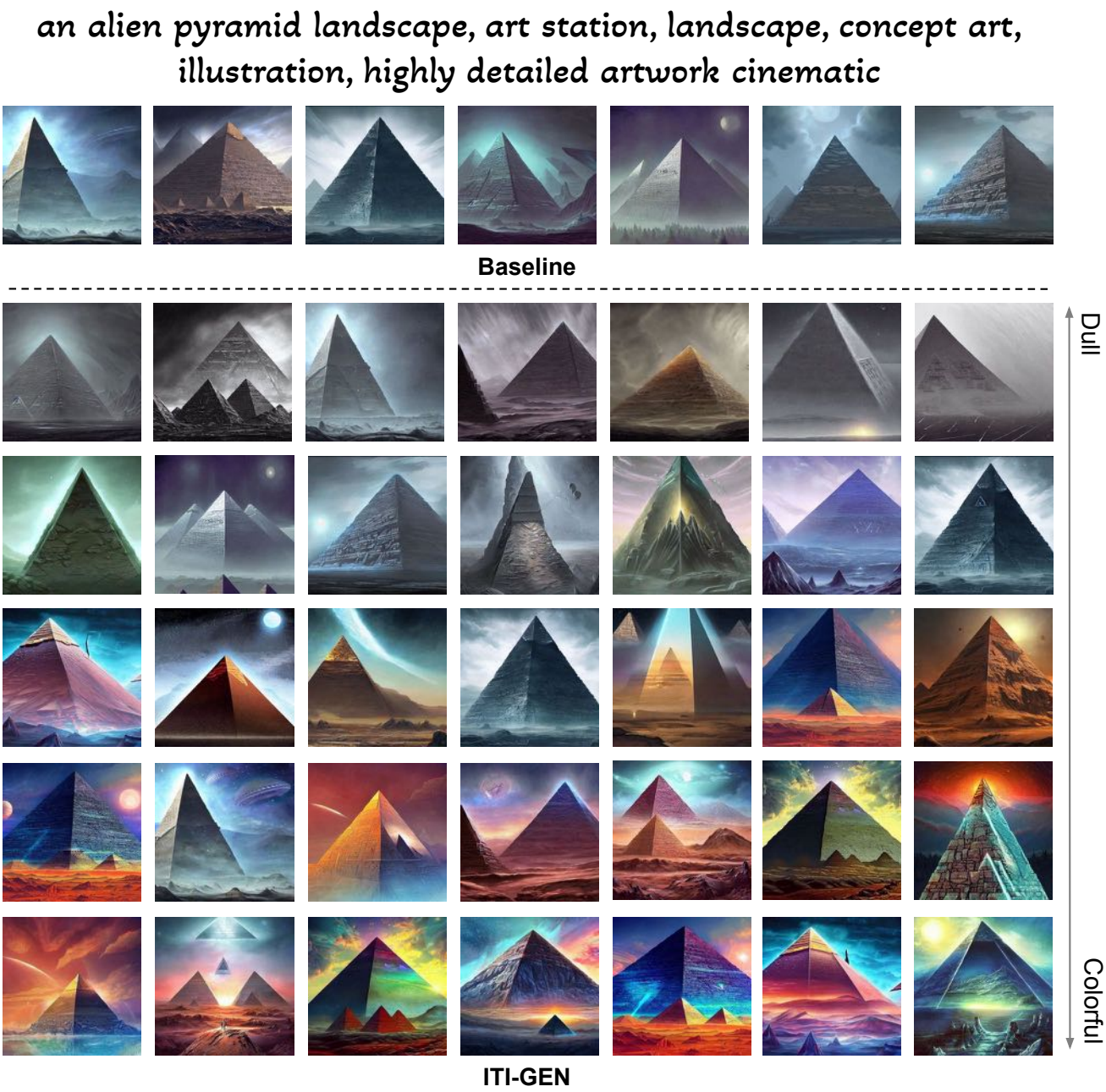


Figure A15. **ITI-GEN with perception attributes ("Colorfulness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of colorfulness. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

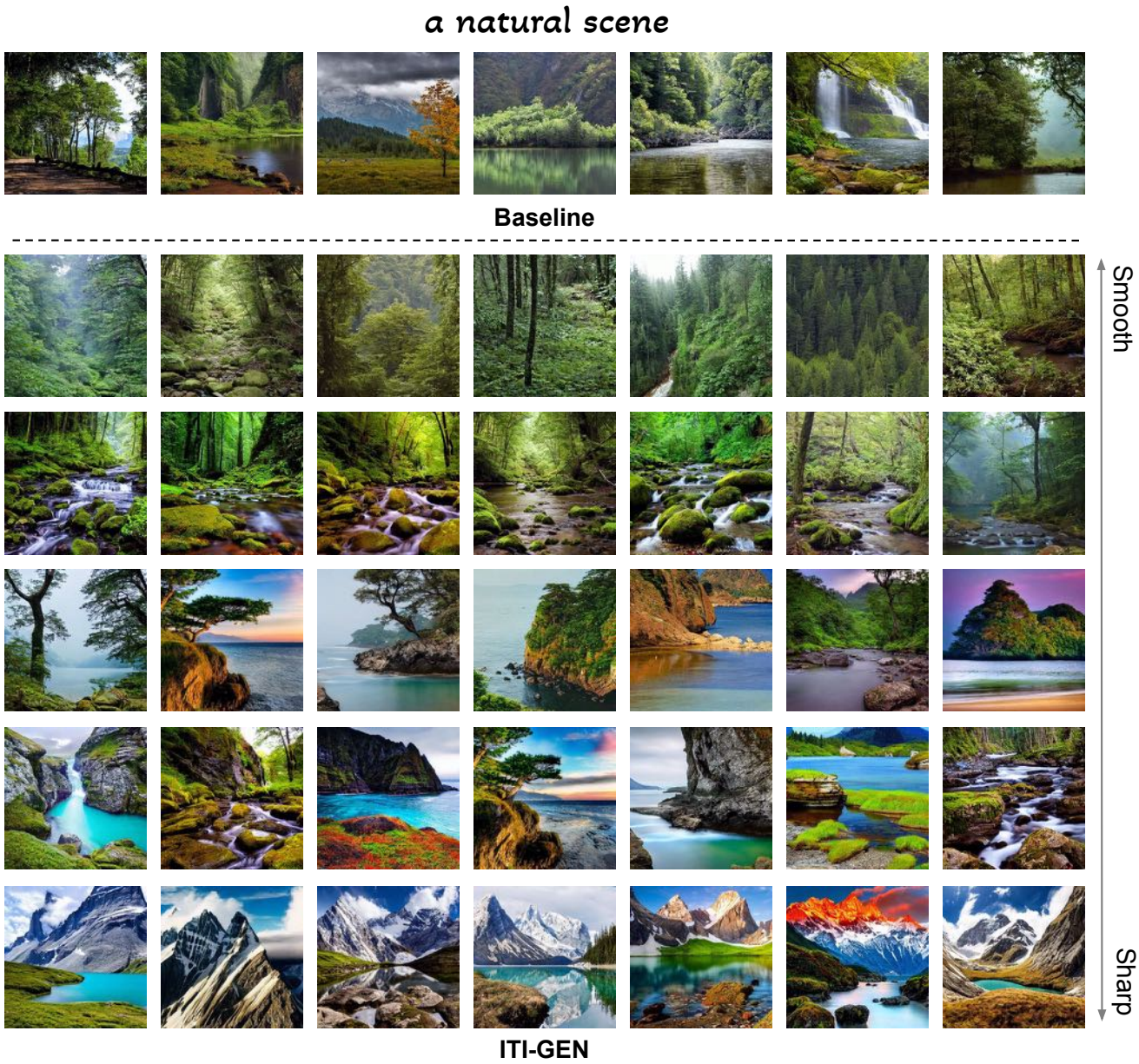


Figure A16. **ITI-GEN with perception attributes ("Sharpness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of sharpness. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

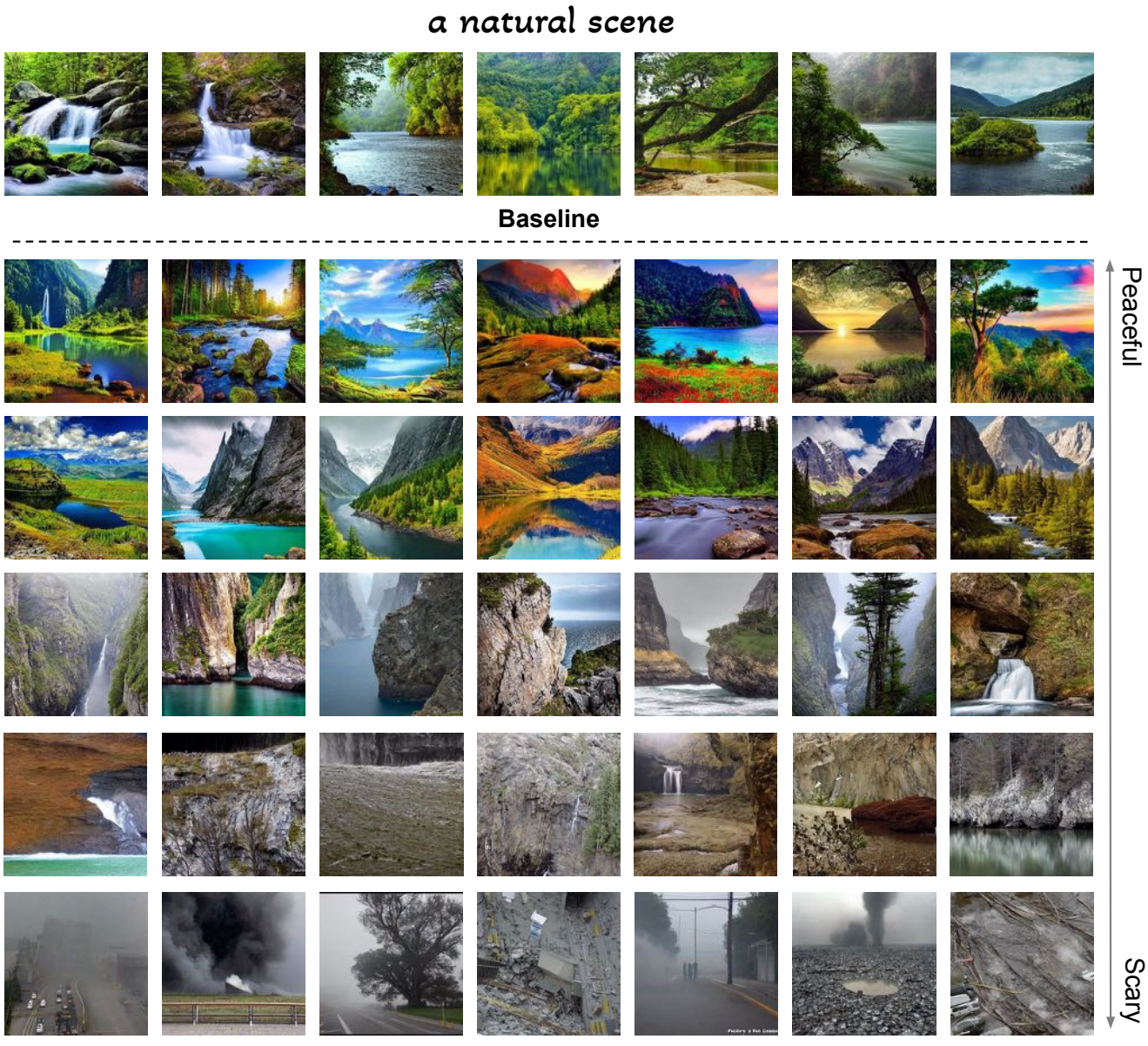


Figure A17. **ITI-GEN with perception attributes ("Scary") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of scary. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

Figure A18. **ITI-GEN with perception attributes ("Contrast") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of contrast. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

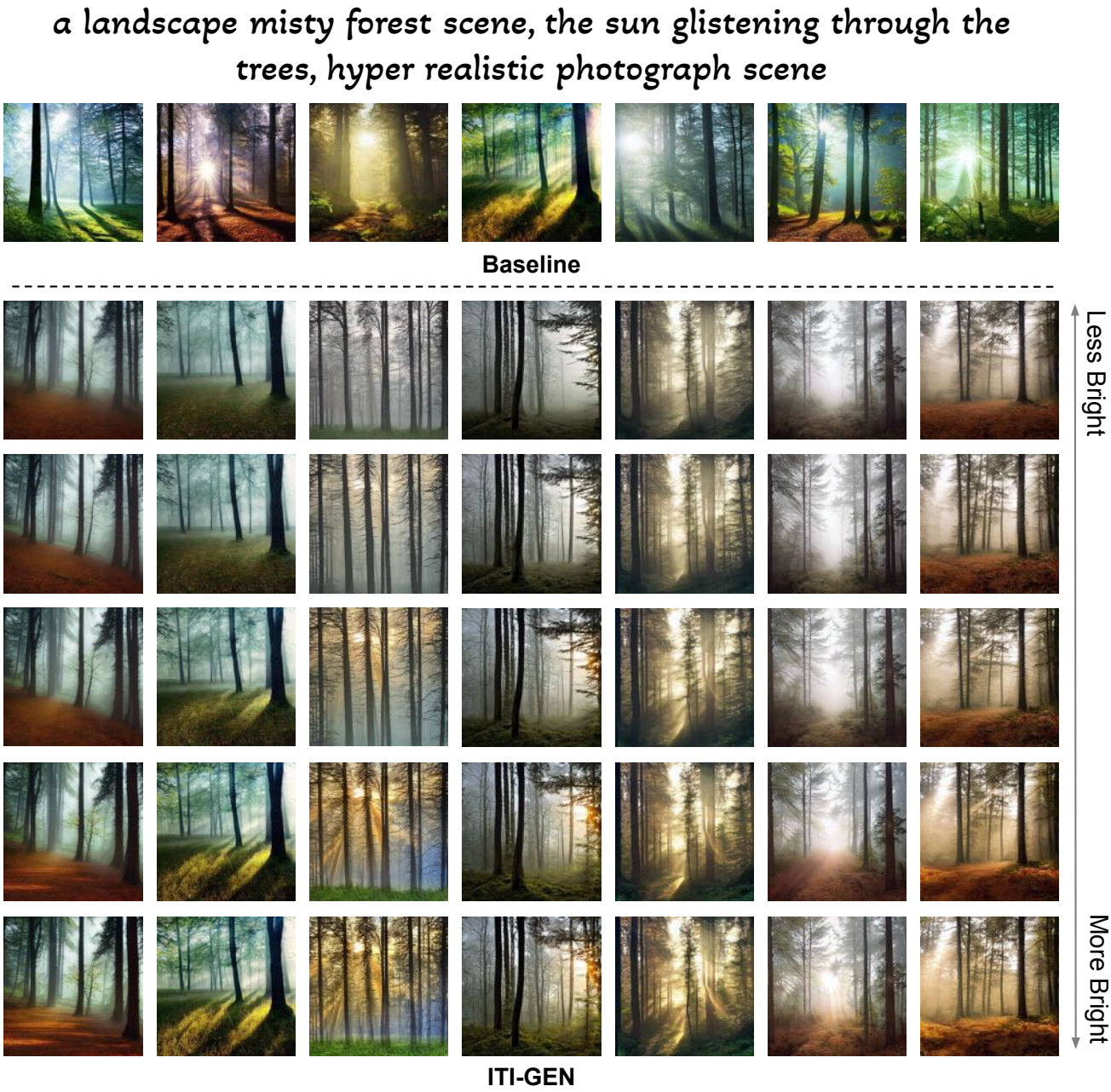


Figure A19. **ITI-GEN with perception attributes ("Brightness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of brightness. In this example, we intentionally pick images using the same random seed in each column for ITI-GEN. Please compare the first and last examples in each column for a clear change in brightness. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

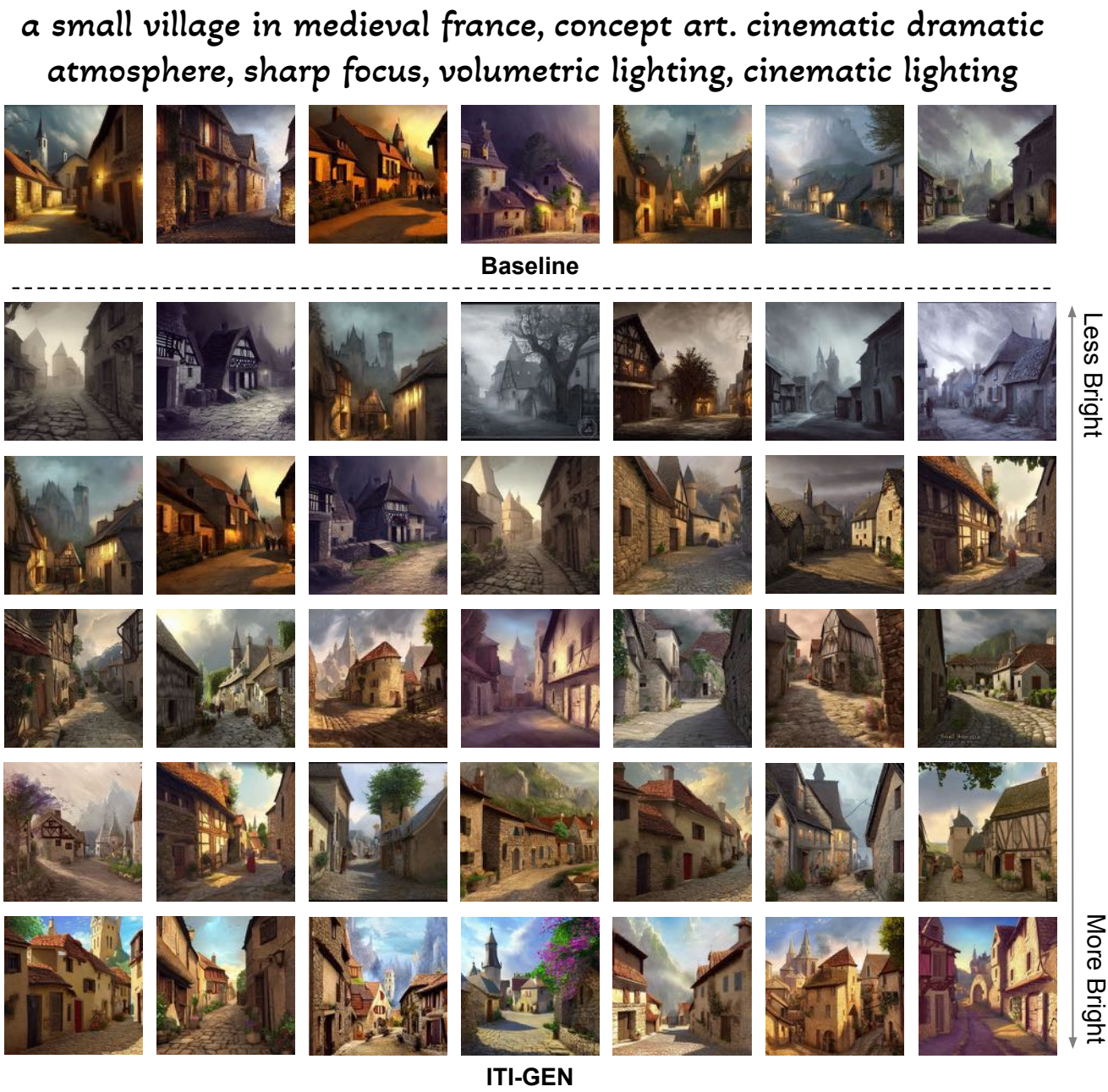


Figure A20. **ITI-GEN with perception attributes ("Brightness") on scene images.** ITI-GEN (bottom) enables the baseline Stable Diffusion (top) to generate images with different levels of brightness. See Section C for details and Figure A1 for reference image examples from LHQ [26]. Better viewed in color.

Figure A21. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

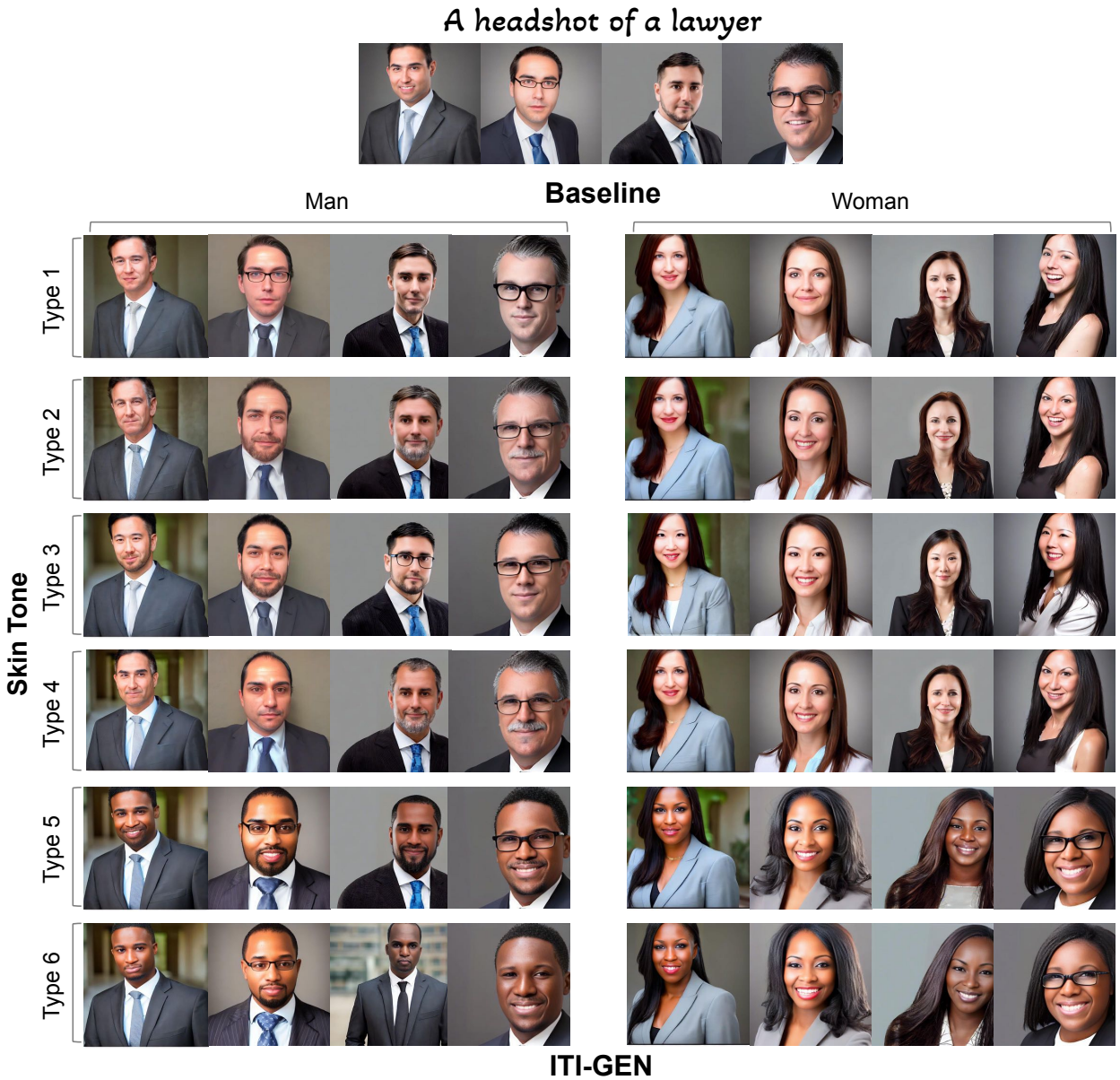


Figure A22. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

Figure A23. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

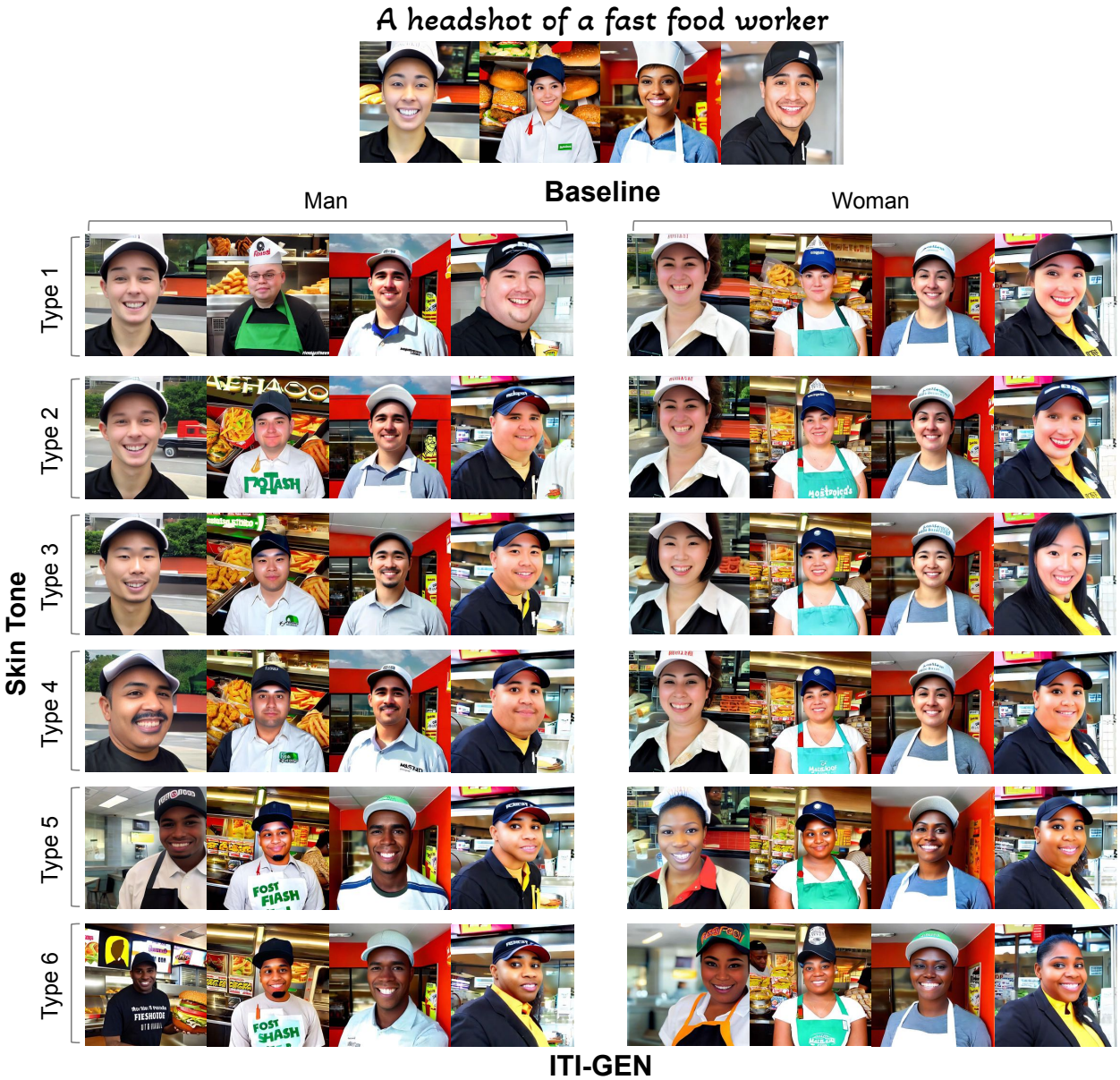
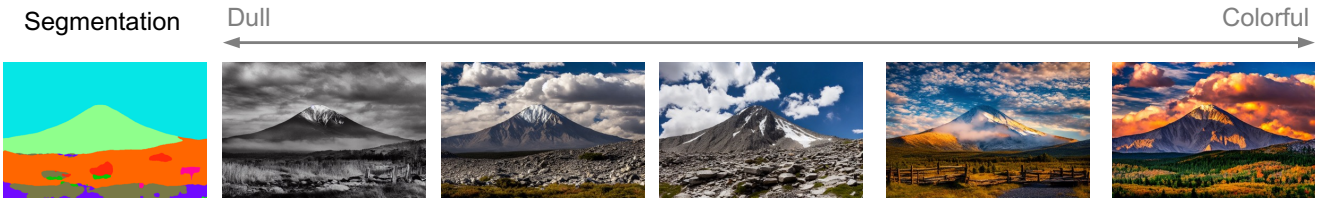
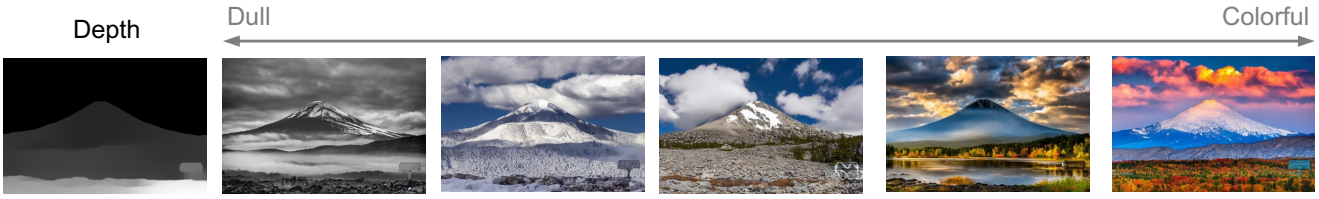


Figure A24. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

Figure A25. **Additional results on train-once-for-all generalization.** Inclusive tokens of ITI-GEN trained with a neutral prompt ("*a headshot of a person*") can be applied to out-of-domain prompts in these three examples to alleviate stereotypes.

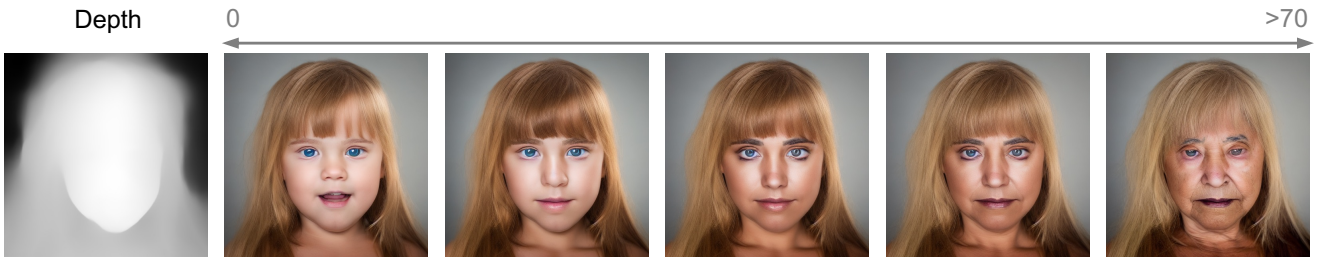


(a) Condition: segmentation map. Attribute: colorfulness.



(b) Condition: depth map. Attribute: colorfulness.

(c) Condition: canny edge map. Attribute: brightness.



(d) Condition: depth map. Attribute: age.

Figure A26. **Additional results on the compatibility with ControlNet [29].** All examples are based on *train-once-for-all* generation (Section 3.3 of the main paper). For scene images in (a), (b), and (c), the inclusive tokens are trained with "*a natural scene*" using LHQ images [26]. For human faces in (d), the tokens for age attribute are trained with "*a headshot of a person*" using FairFace images [16]. As discussed in Section B, our method is designed for improving inclusiveness but not for image editing.