

Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models

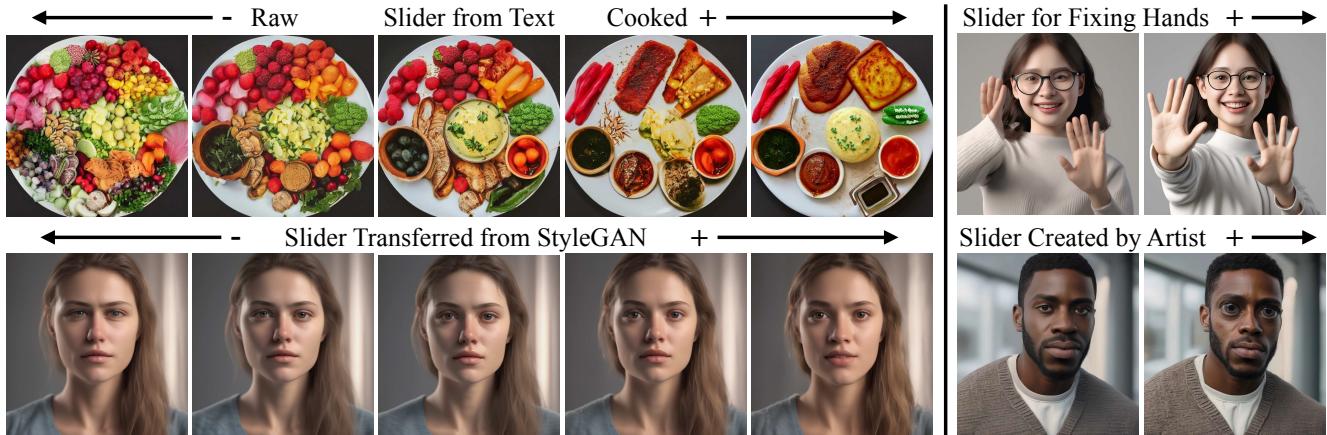
Rohit Gandikota¹Joanna Materzyńska²Tingrui Zhou³Antonio Torralba²David Bau¹¹Northeastern University ²Massachusetts Institute of Technology ³ Independent Researcher

Figure 1. Given a small set of text prompts or paired image data, our method identifies low-rank directions in diffusion parameter space for targeted concept control with minimal interference to other attributes. These directions can be derived from pairs of opposing textual concepts or artist-created images, and they are composable for complex multi-attribute control. We demonstrate the effectiveness of our method by fixing distorted hands in Stable Diffusion outputs and transferring disentangled StyleGAN latents into diffusion models.

Abstract

We present a method to create interpretable concept sliders that enable precise control over attributes in image generations from diffusion models. Our approach identifies a low-rank parameter direction corresponding to one concept while minimizing interference with other attributes. A slider is created using a small set of prompts or sample images; thus slider directions can be created for either textual or visual concepts. Concept Sliders are plug-and-play: they can be composed efficiently and continuously modulated, enabling precise control over image generation. In quantitative experiments comparing to previous editing techniques, our sliders exhibit stronger targeted edits with lower interference. We showcase sliders for weather, age, styles, and expressions, as well as slider compositions. We show how sliders can transfer latents from StyleGAN for intuitive editing of visual concepts for which textual description is difficult. We also find that our method can help address persistent quality issues in Stable Diffusion XL including repair of object deformations and fixing distorted hands. Our code, data, and trained sliders are available at sliders.baulab.info

1. Introduction

Artistic users of text-to-image diffusion models [4, 9, 19, 36, 37] often need finer control over the visual attributes and concepts expressed in a generated image than currently possible. Using only text prompts, it can be challenging to precisely modulate continuous attributes such as a person’s age or the intensity of the weather, and this limitation hinders creators’ ability to adjust images to match their vision [43]. In this paper, we address these needs by introducing interpretable *Concept Sliders* that allow nuanced editing of concepts within diffusion models. Our method empowers creators with high-fidelity control over the generative process as well as image editing. Our code and trained sliders will be open sourced.

Concept Sliders solve several problems that are not well-addressed by previous methods. Direct prompt modification can control many image attributes, but changing the prompt often drastically alters overall image structure due to the sensitivity of outputs to the prompt-seed combina-

¹[gandikota.ro, davidbau]@northeastern.edu

²[jomat, torralba]@mit.edu

³shu_teiei@outlook.jp

tion [22, 38, 44]. Post-hoc techniques such PromptTo-Prompt [13] and Pix2Video [3] enable editing visual concepts in an image by inverting the diffusion process and modifying cross-attentions. However, those methods require separate inference passes for each new concept and can support only a limited set of simultaneous edits. They require engineering a prompt suitable for an individual image rather than learning a simple generalizable control, and if not carefully prompted, they can introduce entanglement between concepts, such as altering race when modifying age (see Appendix). In contrast, Concept Sliders provide lightweight plug-and-play adaptors applied to pre-trained models that enable precise, continuous control over desired concepts in a single inference pass, with efficient composition (Figure 6) and minimal entanglement (Figure 11).

Each Concept Slider is a low-rank modification of the diffusion model. We find that the low-rank constraint is a vital aspect of precision control over concepts: while finetuning without low-rank regularization reduces precision and generative image quality, low-rank training identifies the minimal concept subspace and results in controlled, high-quality, disentangled editing (Figure 11). Post-hoc image editing methods that act on single images rather than model parameters cannot benefit from this low-rank framework.

Concept Sliders also allow editing of visual concepts that cannot be captured by textual descriptions; this distinguishes it from prior concept editing methods that rely on text [7, 8]. While image-based model customization methods [6, 25, 38] can add new tokens for new image-based concepts, those are difficult to use for image editing. In contrast, Concept Sliders allow an artist to provide a handful of paired images to define a desired concept, and then a Concept Slider will then generalize the visual concept and apply it to other images, even in cases where it would be infeasible to describe the transformation in words.

Other generative image models, such as GANs, have previously exhibited latent spaces that provide highly disentangled control over generated outputs. In particular, it has been observed that StyleGAN [20] stylespace neurons offer detailed control over many meaningful aspects of images that would be difficult to describe in words [45]. To further demonstrate the capabilities of our approach, we show that it is possible to create Concept Sliders that transfer latent directions from StyleGAN’s style space trained on FFHQ face images [20] into diffusion models. Notably, despite originating from a face dataset, our method successfully adapts these latents to enable nuanced style control over diverse image generation. This showcases how diffusion models can capture the complex visual concepts represented in GAN latents, even those that may not correspond to any textual description.

We demonstrate that the expressiveness of Concept Sliders is powerful enough to address two particularly practical

applications—enhancing realism and fixing hand distortions. While generative models have made significant progress in realistic image synthesis, the latest generation of diffusion models such as Stable Diffusion XL [36] are still prone to synthesizing distorted hands with anatomically implausible extra or missing fingers [31], as well as warped faces, floating objects, and distorted perspectives. Through a perceptual user study, we validate that a Concept Slider for “realistic image” as well as another for “fixed hands” both create a statistically significant improvement in perceived realism without altering image content.

Concept Sliders are modular and composable. We find that over 50 unique sliders can be composed without degrading output quality. This versatility gives artists a new universe of nuanced image control that allows them to blend countless textual, visual, and GAN-defined Concept Sliders. Because our method bypasses standard prompt token limits, it empowers more complex editing than achievable through text alone.

2. Related Works

Image Editing Recent methods propose different approaches for single image editing in text-to-image diffusion models. They mainly focus on manipulation of cross-attentions of a source image and a target prompt [13, 22, 35], or use a conditional input to guide the image structure [30]. Unlike those methods that are applied to a single image, our model creates a semantic change defined by a small set of text pairs or image pairs, applied to the entire model. Analyzing diffusion models through Riemannian geometry, Park et al. [33] discovered local latent bases that enable semantic editing by traversing the latent space. Their analysis also revealed the evolving geometric structure over timesteps across prompts, requiring per-image latent basis optimization. In contrast, we identify generalizable parameter directions, without needing custom optimization for each image. Instruct-pix2pix [1] finetunes a diffusion model to condition image generation on both an input image and text prompt. This enables a wide range of text-guided editing, but lacks fine-grained control over edit strength or visual concepts not easily described textually.

Guidance Based Methods Ho et al. [14] introduce classifier free guidance that showed improvement in image quality and text-image alignment when the data distribution is driven towards the prompt and away from unconditional output. Liu et al. [28] present an inference-time guidance formulation to enhance concept composition and negation in diffusion models. By adding guidance terms during inference, their method improves on the limited inherent compositionality of diffusion models. SLD [40] proposes using guidance to moderate unsafe concepts in diffusion models. They propose a safe prompt which is used to guide the output away from unsafe content during inference.

Model Editing Our method can be seen as a model editing approach, where by applying a low-rank adaptor, we single out a semantic attribute and allow for continuous control with respect to the attribute. To personalize the models for adding new concepts, customization methods based on finetuning exist [6, 25, 38]. Custom Diffusion [25] proposes a way to incorporate new visual concepts into pretrained diffusion models by finetuning only the cross-attention layers. On the other hand, Textual Inversion [6] introduces new textual concepts by optimizing an embedding vector to activate desired model capabilities. Previous works [7, 12, 23, 24, 46] proposed gradient based fine-tuning-based methods for the permanent erasure of a concept in a model. Ryu et al. [39] proposed adapting LoRA [16] for diffusion model customization. Recent works [47] developed low rank implementations of erasing concepts [7] allowing the ability to adjust the strength of erasure in an image. [17] implemented image based control of concepts by merging two overfitted LoRAs to capture an edit direction. Similarly, [8, 32] proposed closed-form formulation solutions for debiasing, redacting or moderating concepts within the model’s cross-attention weights. Our method does not modify the underlying text-to-image diffusion model and can be applied as a plug-and-play module easily stacked across different attributes.

Semantic Direction in Generative models In Generative Adversarial Networks (GANs), manipulation of semantic attributes has been widely studied. Latent space trajectories have been found in a self-supervised manner [18]. PCA has been used to identify semantic directions in the latent or feature spaces [11]. Latent subspaces corresponding to detailed face attributes have been analyzed [42]. For diffusion models, semantic latent spaces have been suggested to exist in the middle layers of the U-Net architecture [26, 34]. It has been shown that principal directions in diffusion model latent spaces (h-spaces) capture global semantics [10]. Our method directly trains low-rank subspaces corresponding to semantic attributes. By optimizing for specific global directions using text or image pairs as supervision, we obtain precise and localized editing directions. Recent works have [49] introduced the low-rank representation adapter, which employs a contrastive loss to fine-tune LoRA to achieve fine-grained control of concepts in language models.

3. Background

3.1. Diffusion Models

Diffusion models are a subclass of generative models that operationalize the concept of reversing a diffusion process to synthesize data. Initially, the forward diffusion process gradually adds noise to the data, transitioning it from an organized state x_0 to a complete Gaussian noise x_T . At any timestep t , the noised image is modelled as:

$$x_t \leftarrow \sqrt{1 - \beta_t} x_0 + \sqrt{\beta_t} \epsilon \quad (1)$$

Where ϵ is a randomly sampled gaussian noise with zero mean and unit variance. Diffusion models aim to reverse this diffusion process by sampling a random Gaussian noise X_T and gradually denoising the image to generate an image x_0 . In practice [15, 29], the objective of diffusion model is simplified to predicting the true noise ϵ from Eq. 1 when x_t is fed as input with additional inputs like the timestep t and conditioning c .

$$\nabla_{\theta} \|\epsilon - \epsilon_{\theta}(x_t, c, t)\|^2 \quad (2)$$

Where $\epsilon_{\theta}(x_t, c, t)$ is the noise predicted by the diffusion model conditioned on c at timestep t . In this work, we work with Stable Diffusion [37] and Stable Diffusion XL [36], which are latent diffusion models that improve efficiency by operating in a lower dimensional latent space z of a pre-trained variational autoencoder. They convert the images to a latent space and run the diffusion training as discussed above. Finally, they decode the latent z_0 through the VAE decoder to get the final image x_0

3.2. Low-Rank Adaptors

The Low-Rank Adaptation (LoRA) [16] method enables efficient adaptation of large pre-trained language models to downstream tasks by decomposing the weight update ΔW during fine-tuning. Given a pre-trained model layer with weights $W_0 \in \mathbb{R}^{d \times k}$, where d is the input dimension and k the output dimension, LoRA decomposes ΔW as

$$\Delta W = BA \quad (3)$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with $r \ll \min(d, k)$ being a small rank that constrains the update to a low dimensional subspace. By freezing W_0 and only optimizing the smaller matrices A and B , LoRA achieves massive reductions in trainable parameters. During inference, ΔW can be merged into W_0 with no overhead by a LoRA scaling factor α :

$$W = W_0 + \alpha \Delta W \quad (4)$$

4. Method

Concept Sliders are a method for fine-tuning LoRA adaptors on a diffusion model to enable concept-targeted image control as shown in Figure 2. Our method learns low-rank parameter directions that increase or decrease the expression of specific attributes when conditioned on a target concept. Given a target concept c_t and model θ , our goal is to obtain θ^* that modifies the likelihood of attributes c_+ and c_- in image X when conditioned on c_t - increase likelihood of attribute c_+ and decrease likelihood of attribute c_- .

$$P_{\theta^*}(X|c_t) \leftarrow P_{\theta}(X|c_t) \left(\frac{P_{\theta}(c_+|X)}{P_{\theta}(c_-|X)} \right)^{\eta} \quad (5)$$

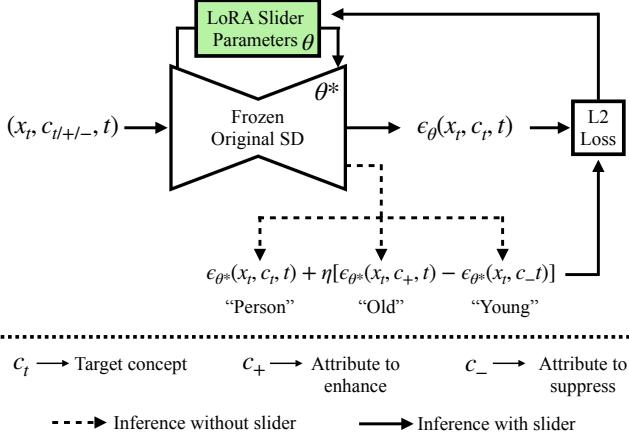


Figure 2. Concept Sliders are created by fine-tuning LoRA adaptors using a guided score that enhances attribute c_+ while suppressing attribute c_- from the target concept c_t . The slider model generates samples x_t by partially denoising Gaussian noise over time steps 1 to t , conditioned on the target concept c_t .

Where $P_\theta(X|c_t)$ represents the distribution generated by the original model when conditioned on c_t . Expanding $P(c_+|X) = \frac{P(X|c_+)P(c_+)}{P(X)}$, the gradient of the log probability $\nabla \log P_{\theta^*}(X|c_t)$ would be proportional to:

$$\nabla \log P_\theta(X|c_t) + \eta (\nabla \log P_\theta(X|c_+) - \nabla \log P_\theta(X|c_-)) \quad (6)$$

Based on Tweedie’s formula [5] and the reparametrization trick of [15], we can introduce a time-varying noising process and express each score (gradient of log probability) as a denoising prediction $\epsilon(X, c_t, t)$. Thus Eq. 6 becomes:

$$\begin{aligned} \epsilon_{\theta^*}(X, c_t, t) &\leftarrow \epsilon_\theta(X, c_t, t) + \\ &\eta (\epsilon_\theta(X, c_+, t) - \epsilon_\theta(X, c_-, t)) \end{aligned} \quad (7)$$

The proposed score function in Eq. 7 shifts the distribution of the target concept c_t to exhibit more attributes of c_+ and fewer attributes of c_- . In practice, we notice that a single prompt pair can sometimes identify a direction that is entangled with other undesired attributes. We therefore incorporate a set of preservation concepts $p \in \mathcal{P}$ (for example, race names while editing age) to constrain the optimization. Instead of simply increasing $P_\theta(c_+|X)$, we aim to increase, for every p , $P_\theta((c_+, p)|X)$, and reduce $P_\theta((c_-, p)|X)$. This leads to the disentanglement objective:

$$\begin{aligned} \epsilon_{\theta^*}(X, c_t, t) &\leftarrow \epsilon_\theta(X, c_t, t) + \\ &\eta \sum_{p \in \mathcal{P}} (\epsilon_\theta(X, (c_+, p), t) - \epsilon_\theta(X, (c_-, p), t)) \end{aligned} \quad (8)$$

The disentanglement objective in Equation 8 finetunes the Concept Slider modules while keeping pre-trained weights

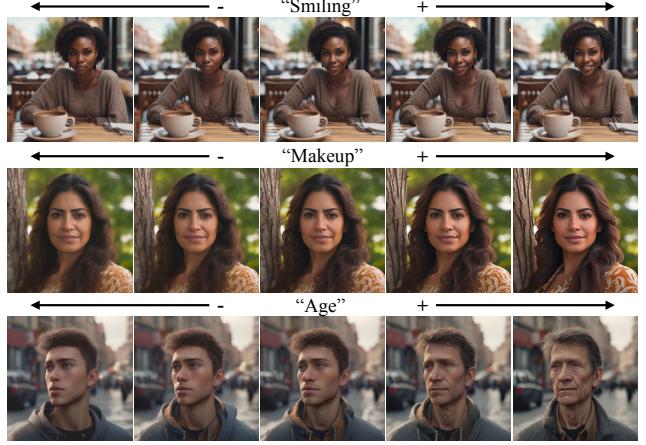


Figure 3. Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure. Traversing the sliders towards the negative direction produces an opposing effect on the attributes.

fixed. Crucially, the LoRA formulation in Equation 4 introduces a scaling factor α that can be modified at inference time. This scaling parameter α allows adjusting the strength of the edit, as shown in Figure 1. Increasing α makes the edit stronger without retraining the model. Previous model editing method [7], suggests a stronger edit by retraining with increased guidance η in Eq. 8. However, simply scaling α at inference time produces the same effect of strengthening the edit, without costly retraining.

4.1. Learning Visual Concepts from Image Pairs

We propose sliders to control nuanced visual concepts that are harder to specify using text prompts. We leverage small paired before/after image datasets to train sliders for these concepts. The sliders learn to capture the visual concept through the contrast between image pairs (x^A, x^B) .

Our training process optimizes the LORA applied in both the negative and positive directions. We shall write ϵ_{θ_+} for the application of positive LoRA and ϵ_{θ_-} for the negative case. Then we minimize the following loss:

$$\|\epsilon_{\theta_-}(x_t^A, \cdot, t) - \epsilon\|^2 + \|\epsilon_{\theta_+}(x_t^B, \cdot, t) - \epsilon\|^2 \quad (9)$$

This has the effect of causing the LORA to align to a direction that causes the visual effect of A in the negative direction and B in the positive direction. Defining directions visually in this way not only allows an artist to define a Concept Slider through custom artwork; it is also the same method we use to transfer latents from other generative models such as StyleGAN.

5. Experiments

We evaluate our approach primarily on Stable Diffusion XL [36], a high-resolution 1024-pixel model, and we conduct ad-

	Prompt2Prompt		Our Method		Composition	
	Δ CLIP	LPIPS	Δ CLIP	LPIPS	Δ CLIP	LPIPS
Age	1.10	0.15	3.93	0.06	3.14	0.13
Hair	3.45	0.15	5.59	0.10	5.14	0.15
Sky	0.43	0.15	1.56	0.13	1.55	0.14
Rusty	7.67	0.25	7.60	0.09	6.67	0.18

Table 1. Compared to Prompt2Prompt [13], our method achieves comparable efficacy in terms of Δ CLIP score while inducing finer edits as measured by LPIPS distance to the original image. The Δ CLIP metric measures the change in CLIP score between the original and edited images when evaluated on the text prompt describing the desired edit. Results are shown for a single positive scale of the trained slider.

ditional experiments on SD v1.4 [37]. All models are trained for 500 epochs. We demonstrate generalization by testing sliders on diverse prompts - for example, we evaluate our "person" slider on prompts like "doctor", "man", "woman", and "barista". For inference, we follow the SDEdit technique of Meng et al. [30]: to maintain structure and semantics, we use the original pre-trained model for the first t steps, setting the LoRA adaptor multipliers to 0 and retaining the pre-trained model priors. We then turn on the LoRA adaptor for the remaining steps.

5.1. Textual Concept Sliders

We validate the efficacy of our slider method on a diverse set of 30 text-based concepts, with full examples in the Appendix. Table 1 compares our method against two baselines: an approach we propose inspired by SDEdit [30] and Liu et al.[28] that uses a pretrained model with the standard prompt for t timesteps, then starts composing by adding prompts to steer the image, and prompt2prompt[13], which leverages cross-attention for image editing after generating reference images. While the former baseline is novel, all three enable finer control but differ in how edits are applied. Our method directly generates 2500 edited images per concept, like "image of a person", by setting the scale parameter at inference. In contrast, the baselines require additional inference passes for each new concept (e.g "old person"), adding computational overhead. Our method consistently achieves higher CLIP scores and lower LPIPS versus the original, indicating greater coherence while enabling precise control. The baselines are also more prone to entanglement between concepts. We provide further analysis and details about the baselines in the Appendix.

Figure 3 shows typical qualitative examples, which maintains good image structure while enabling fine grained editing of the specified concept.

5.2. Visual Concept Sliders

Some visual concepts like precise eyebrow shapes or eye sizes are challenging to control through text prompts alone.

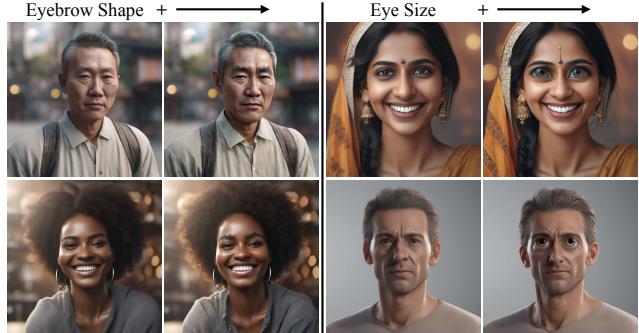


Figure 4. Controlling fine-grained attributes like eyebrow shape and eye size using image pair-driven concept sliders with optional text guidance. The eye size slider scales from small to large eyes using the Ostris dataset [2].

	Training Data	Custom Diffusion	Textual Inversion	Our Method
Δ_{eye}	1.84	0.97	0.81	1.75
LPIPS	0.03	0.23	0.21	0.06

Table 2. Our results demonstrate the effectiveness of our sliders for intuitive image editing based on visual concepts. The metric Δ_{eye} represents the ratio of change in eye size compared to the original image. Our method achieves targeted editing of eye size while maintaining similarity to the original image distribution, as measured by the LPIPS.

To enable sliders for these granular attributes, we leverage paired image datasets combined with optional text guidance. As shown in Figure 4, we create sliders for "eyebrow shape" and "eye size" using image pairs capturing the desired transformations. We can further refine the eyebrow slider by providing the text "eyebrows" so the direction focuses on that facial region. Using image pairs with different scales, like the eye sizes from Ostris [2], we can create sliders with stepwise control over the target attribute.

We quantitatively evaluate the eye size slider by detecting faces using FaceNet [41], cropping the area, and employing a face parser [48]to measure eye region across the slider range. Traversing the slider smoothly increases the average eye area 2.75x, enabling precise control as shown in Table 2. Compared to customization techniques like textual inversion [6] that learns a new token and custom diffusion [25] that fine-tunes cross attentions, our slider provides more targeted editing without unwanted changes. When model editing methods [6, 25] are used to incorporate new visual concepts, they memorize the training subjects rather than generalizing the contrast between pairs. We provide more details in the Appendix.

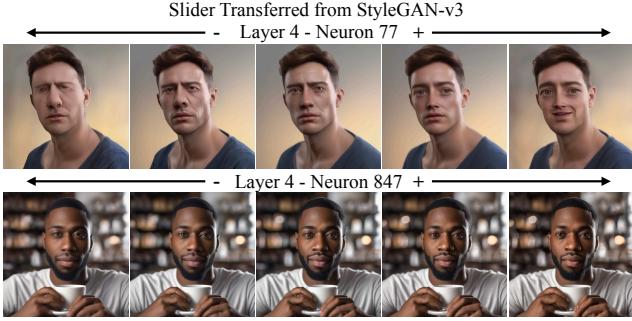


Figure 5. We demonstrate transferring StyleGAN style space latents to the diffusion latent space. We identify three neurons that edit facial structure: neuron 77 controls cheekbone structure, neuron 646 selectively adjusts the left side face width, and neuron 847 edits inter-ocular distance. We transfer these StyleGAN latents to the diffusion model to enable structured facial editing.

5.3. Sliders transferred from StyleGAN

Figure 5 demonstrates sliders transferred from the StyleGAN-v3 [21] style space that is trained on FFHQ [20] dataset. We use the method of [45] to explore the StyleGAN-v3 style space and identify neurons that control hard-to-describe facial features. By scaling these neurons, we collect images to train image-based sliders. We find that Stable Diffusion’s latent space can effectively learn these StyleGAN style neurons, enabling structured facial editing. This enables users to control nuanced concepts that are indescribable by words and styleGAN makes it easy to get generate the paired dataset.

5.4. Composing Sliders

A key advantage of our low-rank slider directions is composability - users can combine multiple sliders for nuanced control rather than being limited to one concept at a time. For example, in Figure 6 we show blending "cooked" and "fine dining" food sliders to traverse this 2D concept space. Since our sliders are lightweight LoRA adaptors, they are easy to share and overlay on diffusion models. By downloading interesting slider sets, users can adjust multiple knobs simultaneously to steer complex generations. In Figure 7 we qualitatively show the effects of composing multiple sliders progressively up to 50 sliders at a time. We use far greater than 77 tokens (the current context limit of SDXL [36]) to create these 50 sliders. This showcases the power of our method that allows control beyond what is possible through prompt-based methods alone. We further validate multi-slider composition in the appendix.

6. Concept Sliders to Improve Image Quality

One of the most interesting aspects of a large-scale generative model such as Stable Diffusion XL is that, although

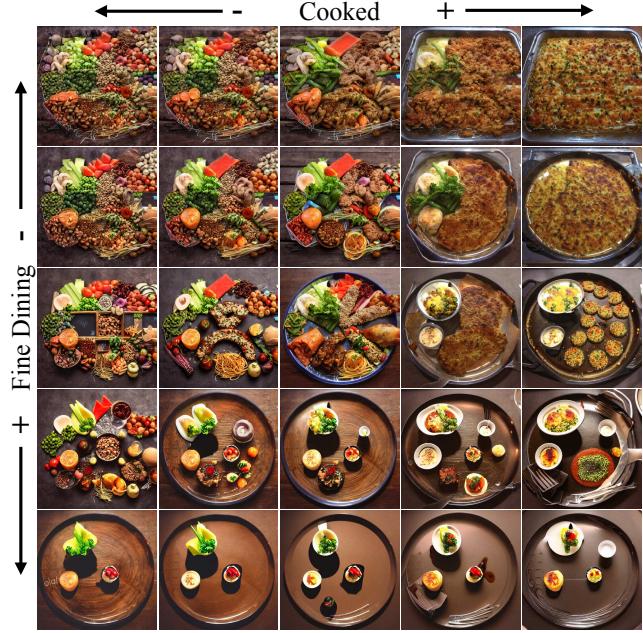


Figure 6. Composing two text-based sliders results in a complex control over food images. We show the effect of applying both the "cooked" slider and "fine-dining" slider to a generated image. These sliders can be used in both positive and negative directions.



Figure 7. We show composition capabilities of concept sliders. We progressively compose multiple sliders in each row from left to right, enabling nuanced traversal of high-dimensional concept spaces. We demonstrate composing sliders trained from text prompts, image datasets, and transferred from GANs.

their image output can often suffer from distortions such as warped or blurry objects, the parameters of the model contains a latent capability to generate higher-quality output with fewer distortions than produced by default. Concept Sliders can unlock these abilities by identifying low-rank parameter directions that repair common distortions.

Fixing Hands Generating realistic-looking hands is a persistent challenge for diffusion models: for example, hands are typically generated with missing, extra, or misplaced fingers. Yet the tendency to distort hands can be directly controlled by a Concept Slider: Figure 9 shows the effect



Figure 8. The repair slider enables the model to generate images that are more realistic and undistorted. The parameters under the control of this slider help the model correct some of the flaws in their generated outputs like distorted humans and pets in (a, b), unnatural objects in (b, c, d), and blurry natural images in (b,c)



Figure 9. We demonstrate a slider for fixing hands in stable diffusion. We find a direction to steer hands to be more realistic and away from "poorly drawn hands".

of a "fix hands" Concept Slider that lets users smoothly adjust images to have more realistic, properly proportioned hands. This parameter direction is found using a complex prompt pair boosting "realistic hands, five fingers, 8k hyper-realistic hands" and suppressing "poorly drawn hands, distorted hands, misplaced fingers". This slider allows hand quality to be improved with a simple tweak rather manual prompt engineering.

To measure the "fix hands" slider, we conduct a user study on Amazon Mechanical Turk. We present 300 random images with hands to raters—half generated by Stable Diffusion XL and half by XL with our slider applied (same seeds and prompts). Raters are asked to assess if the hands appear distorted or not. Across 150 SDXL images, raters find 62% have distorted hands, confirming it as a prevalent problem. In contrast, only 22% of the 150 slider images are rated as having distorted hands.

Repair Slider In addition to controlling specific concepts like hands, we also demonstrate the use of Concept Sliders to guide generations towards overall greater realism. We identify single low-rank parameter direction that shifts images away from common quality issues like distorted subjects, un-



Figure 10. We demonstrate the effect of our "repair" slider on fine details: it improves the rendering of densely arranged objects, it straightens architectural lines, and it avoids blurring and distortions at the edges of complex shapes.

natural object placement, and inconsistent shapes. As shown in Figures 8 and 10, traversing this "repair" slider noticeably fixes many errors and imperfections.

Through a perceptual study, we evaluate the realism of 250 pairs of slider-adjusted and original SD images. A majority of participants rate the slider images as more realistic in 80.39% of pairs, indicating our method enhances realism. However, FID scores do not align with this human

assessment, echoing prior work on perceptual judgment gaps [27]. Instead, distorting images along the opposite slider direction improves FID, though users still prefer the realism-enhancing direction. We provide more details about the user studies in the appendix.

7. Ablations

We analyze the two key components of our method to verify that they are both necessary: (1) the disentanglement formulation and (2) low-rank adaptation. Table 3 shows quantitative measures on 2500 images, and Figure 11 shows qualitative differences. In both quantitative and quantitative measures, we find that the disentanglement objective from Eq.8 success in isolating the edit from unwanted attributes (Fig.11.c); for example without this objective we see undesired changes in gender when asking for age as seen in Table 3, Interference metric which measures the percentage of samples with changed race/gender when making the edit. The low-rank constraint is also helpful: it has the effect of precisely capturing the edit direction with better generalization (Fig.11.d); for example, note how the background and the clothing are better preserved in Fig.11.b. Since LORA is parameter-efficient, it also has the advantage that it enables lightweight modularity. We also note that the SDEdit-inspired inference technique allows us to use a wider range of alpha values, increasing the editing capacity, without losing image structure. We find that SDEdit’s inference technique expands the usable range of alpha before coherence declines relative to the original image. We provide more details in the Appendix.

	w/o Ours	w/o Disentanglement	w/o Low Rank
Δ_{CLIP}	3.93	3.39	3.18
LPIPS	0.06	0.17	0.23
Interference	0.10	0.36	0.19

Table 3. The disentanglement formulation enables precise control over the age direction, as shown by the significant reduction in the Interference metric which measures the percentage of samples with gender/race change, compared to the original images. By using LoRA adaptors, sliders achieve finer editing in terms of both structure and edit direction, as evidenced by improvements in LPIPS and Interference. Concept strength is maintained, with similar Δ_{CLIP} scores across ablations.

8. Limitations

While the disentanglement formulation reduces unwanted interference between edits, we still observe some residual effects as shown in Table 3 for our sliders. This highlights the need for more careful selection of the latent directions to preserve, preferably an automated method, in order to

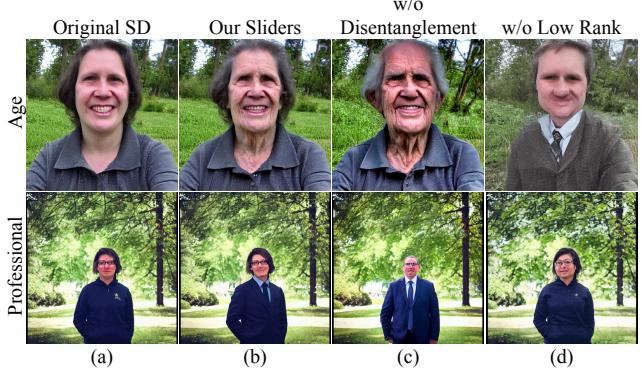


Figure 11. The disentanglement objective (Eq. 8) helps avoid undesired attribute changes like change in race or gender when editing age. The low-rank constraint enables a precise edit.

further reduce edit interference. Further study is required to determine the optimal set of directions that minimizes interference while retaining edit fidelity. We also observe that while the inference SDEdit technique helps preserve image structure, it can reduce edit intensity compared to the inference-time method, as shown in Table 1. The SDEdit approach appears to trade off edit strength for improved structural coherence. Further work is needed to determine if the edit strength can be improved while maintaining high fidelity to the original image.

9. Conclusion

Concept Sliders are a simple and scalable new paradigm for interpretable control of diffusion models. By learning precise semantic directions in latent space, sliders enable intuitive and generalized control over image concepts. The approach provides a new level of flexibility beyond text-driven, image-specific diffusion model editing methods, because Concept Sliders allow continuous, single-pass adjustments without extra inference. Their modular design further enables overlaying many sliders simultaneously, unlocking complex multi-concept image manipulation.

We have demonstrated the versatility of Concept Sliders by measuring their performance on Stable Diffusion XL and Stable Diffusion 1.4. We have found that sliders can be created from textual descriptions alone to control abstract concepts with minimal interference with unrelated concepts, outperforming previous methods. We have demonstrated and measured the efficacy of sliders for nuanced visual concepts that are difficult to describe by text, derived from small artist-created image datasets. We have shown that Concept Sliders can be used to transfer StyleGAN latents into diffusion models. Finally, we have conducted a human study that verifies the high quality of Concept Sliders that enhance and correct hand distortions. Our code and data will be made publicly available.

Acknowledgments

We thank Jaret Burkett (aka Ostris) for the continued discussion on the image slider method and for sharing their eye size dataset. RG and DB are supported by Open Philanthropy.

Code

Our methods are available as open-source code. Source code, trained sliders, and data sets for reproducing our results can be found at sliders.baulab.info and at <https://github.com/rohitgandikota/sliders>.

References

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. [2](#)
- [2] Jarret Burkett. Ostris/ai-toolkit: Various ai scripts. mostly stable diffusion stuff., 2023. [5](#)
- [3] Duygu Ceylan, Chun-Hao Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *International Conference on Computer Vision (ICCV)*, 2023. [2](#)
- [4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. [1](#)
- [5] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. [4](#)
- [6] 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, 3, 5](#)
- [7] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. [2, 3, 4](#)
- [8] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. [2, 3](#)
- [9] Google. Imagen, unprecedented photorealism x deep level of language understanding, 2022. [1](#)
- [10] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. *arXiv preprint arXiv:2303.11073*, 2023. [3](#)
- [11] Erik Häkkinen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. [3](#)
- [12] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023. [3](#)
- [13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. [2, 5](#)
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [2](#)
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. [3, 4](#)
- [16] 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. [3](#)
- [17] Norm Inui. Sd/sdxl tricks beneath the papers and codes, 2023. [3](#)
- [18] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. [3](#)
- [19] et al James Betker. Improving image generation with better captions. *OpenAI Reports*, 2023. [1](#)
- [20] 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. [2, 6](#)
- [21] Tero Karras, Miika Aittala, Samuli Laine, Erik Häkkinen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. [6](#)
- [22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagine: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. [2](#)
- [23] Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023. [3](#)
- [24] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *International Conference on Computer Vision (ICCV)*, 2023. [3](#)
- [25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [2, 3, 5](#)
- [26] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. [3](#)
- [27] Tuomas Kynkänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. [8](#)
- [28] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. [2, 5](#)
- [29] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. [3](#)

- [30] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 5
- [31] Mothrider. “can an ai draw hands?”, 2022. 2
- [32] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 3
- [33] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *arXiv preprint arXiv:2307.12868*, 2023. 2
- [34] Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023. 3
- [35] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2
- [36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3, 4, 6
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 5
- [38] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3
- [39] Simo Ryu. Cloneofsimo/lora: Using low-rank adaptation to quickly fine-tune diffusion models.s, 2023. 3
- [40] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *arXiv preprint arXiv:2211.05105*, 2022. 2
- [41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5
- [42] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [43] Staffell. The sheer number of options and sliders using stable diffusion is overwhelming., 2023. 1
- [44] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2023. 2
- [45] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 2, 6
- [46] Eric Zhang, Kai Wang, Xinqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 3
- [47] Tingrui Zhou. Github - p1atdev/leco: Low-rank adaptation for erasing concepts from diffusion models., 2023. 3
- [48] Zllrunning. Using modified bisenet for face parsing in pytorch, 2019. 5
- [49] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. 3

Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models

Supplementary Material

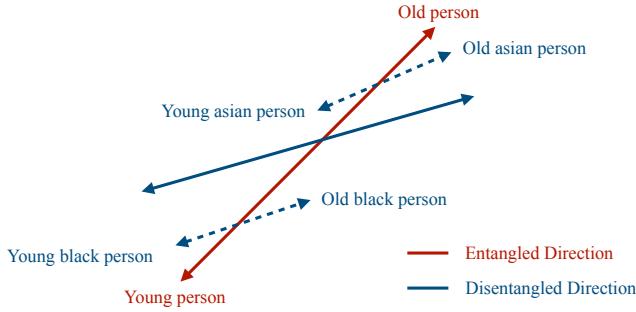


Figure 12. In this schematic we illustrate how multiple preservation concepts are used to disentangle a direction. For the sake of clarity in figure, we show examples for just two races. In practice, we preserve a diversity of several protected attribute directions.

10. Disentanglement Formulation

We visualize the rationale behind our disentangled formulation for sliders. When training sliders on single pair of prompts, sometimes the directions are entangled with unintended directions. For example, as we show in Figure 11, controlling age can interfere with gender or race. We therefore propose using multiple paired prompts for finding a disentangled direction. As shown in Figure 12, we explicitly define the preservation directions (dotted blue lines) to find a new edit direction (solid blue line) invariant to the preserve features.

11. SDEdit Analysis

We ablate SDEdit’s contribution by fixing slider scale while varying SDEdit timesteps over 2,500 images. Figure 13 shows inverse trends between LPIPS and CLIP distances as SDEdit time increases. Using more SDEdit maintains structure, evidenced by lower LPIPS score, while maintaining lower CLIP score. This enables larger slider scales before risking structural changes. We notice that on average, timestep 750 - 850 has the best of both worlds with spatial structure preservation and increased efficacy.

12. Textual Concepts Sliders

We quantify slider efficacy and control via CLIP score change and LPIPS distance over 15 sliders at 12 scales in Figure 14. CLIP score change validates concept modification strength. Tighter LPIPS distributions demonstrate precise spatial manipulation without distortion across scales. We show additional qualitative examples for textual concept sliders in Figures 27-32.

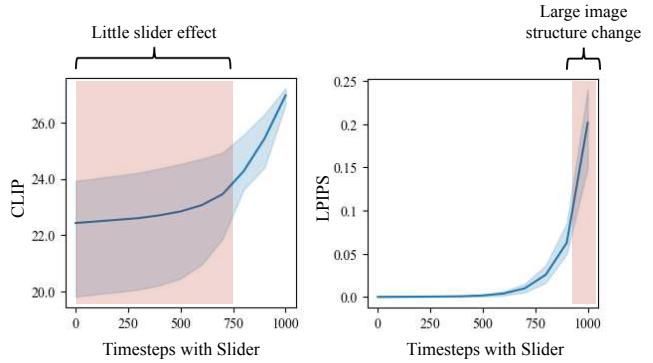


Figure 13. The plot examines CLIP score change and LPIPS distance when applying the same slider scale but with increasing SDEdit times. Higher timesteps enhance concept attributes considerably per CLIP while increased LPIPS demonstrates change in spatial stability. On the x-axis, 0 corresponds to no slider application while 1000 represents switching from start.

12.1. Baseline Details

We compare our method against Prompt-to-prompt and a novel inference-time prompt composition method. For Prompt-to-prompt we use the official implementation code. We use the Refinement strategy they propose, where new token is added to the existing prompt for image editing. For example, for the images in Figure 15, we add the token “old” for the original prompt “picture of person” to make it “picture of old person”. For the composition method, we use the principles from Liu et al . Specifically, we compose the score functions coming from both “picture of person” and “old person” through additive guidance. We also utilize the SDEdit technique for this method to allow finer image editing.

12.2. Entanglement

The baselines are sometimes prone to interference with concepts when editing a particular concept. Table 4 shows quantitative analysis on interference while Figure 15 shows some qualitative examples. We find that Prompt-to-prompt and inference composition can sometimes change the race/gender when editing age. Our sliders with disentanglement object 8, show minimal interference as seen by *Interference* metric, which shows the percentage samples with race or gender changed out of 2500 images we tested. We also found through LPIPS metric that our method shows finer editing

<https://github.com/google/prompt-to-prompt/>

<https://energy-based-model.github.io/Compositional-Visual-Generation-with-Composable-Diffusion-Models/>

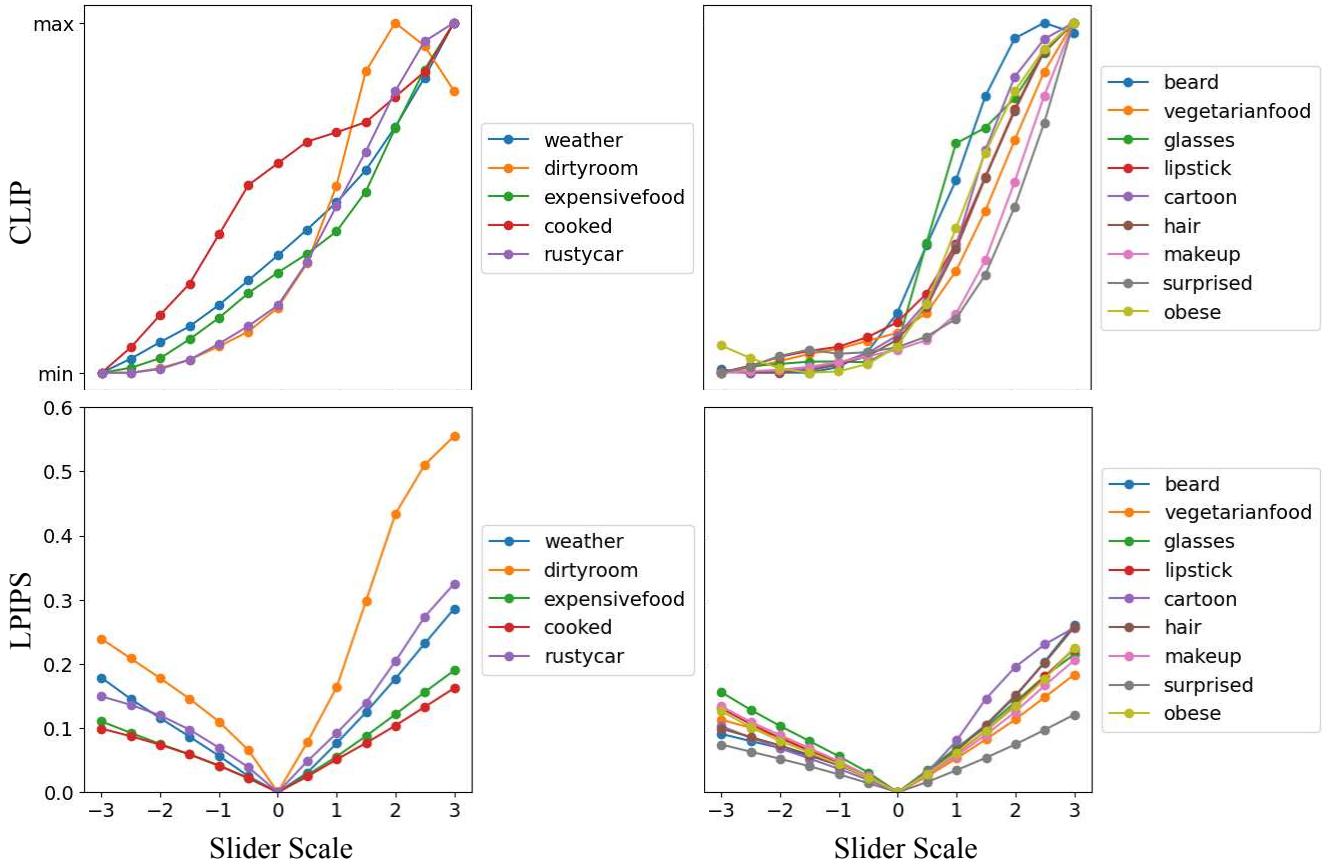


Figure 14. Analyzing attribute isolation efficacy vs stylistic variation for 15 slider types across 12 scales. We divide our figure into two columns. The left column contains concepts that have words for antonyms (*e.g.* expensive - cheap) showing symmetric CLIP score deltas up/down. The right column shows harder to negate sliders (*e.g.* no glasses) causing clipped negative range. We also note that certain sliders have higher lrips, such as “cluttered” room slider, which intuitively makes sense.

capabilities. We find similar conclusions through quantitative samples from Figure 15, that P2P and composition can alter gender, race or both when controlling age.

	P2P	Composition	Ours
Δ_{CLIP}	1.10	3.14	3.93
LPIPS	0.15	0.13	0.06
Interference	0.33	0.38	0.10

Table 4. The disentanglement formulation enables precise control over the age direction, as shown by the significant reduction in the Interference metric which measures the percentage of samples with gender/race change, compared to the original images. By using LoRA adaptors, sliders achieve finer editing in terms of both structure and edit direction, as evidenced by improvements in LPIPS and Interference. Concept strength is maintained, with similar Δ_{CLIP} scores across ablations.

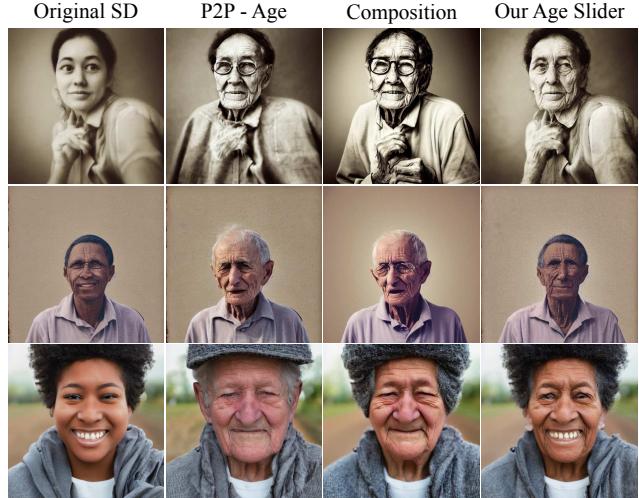


Figure 15. Concept Sliders demonstrate minimal entanglement when controlling a concept. Prompt-to-prompt and inference-time textual composition sometimes tend to alter race/gender when editing age.



Figure 16. Concept Sliders demonstrate more diverse outputs while also being effective at learning the new concepts. Customization methods can sometimes tend to learn unintended concepts like hair and eye colors.

13. Visual Concept

13.1. Baseline Details

We compare our method to two image customization baselines: custom diffusion and textual inversion . For fair comparison, we use the official implementations of both, modifying textual inversion to support SDXL. These baselines learn concepts from concept-labeled image sets. However, this approach risks entangling concepts with irrelevant attributes (e.g. hair, skin tone) that correlate spuriously in the dataset, limiting diversity.

13.2. Precise Concept Capturing

Figure 16 shows non-cherry-picked customization samples from all methods trained on the large-eyes Ostris dataset . While exhibiting some diversity, samples frequently include irrelevant attributes correlated with large eyes in the dataset, e.g. blonde hair in custom diffusion, blue eyes in textual inversion. In contrast, our paired image training isolates concepts by exposing only local attribute changes, avoiding spurious correlation learning.

14. Composing Sliders

We show a 2 dimensional slider by composing “cooked” and “fine dining” food sliders in Figure 17. Next, we show progressive composition of sliders one by one in Figures 18,19. From top left image (original SDXL), we progressively generate images by composing a slider at each step. We show how our sliders provide a semantic control over images.

<https://github.com/adobe-research/custom-diffusion>
https://github.com/rinongal/textual_inversion
<https://github.com/ostris/ai-toolkit>



Figure 17. Composing two text-based sliders results in a complex control over thanksgiving food options. We show the effect of applying both the “cooked” slider and “fine-dining” slider to a generated image of thanksgiving dinner. These sliders can be used in both positive and negative directions.



Figure 18. Concept Sliders can be composed for a more nuanced and complex control over attributes in an image. From stable diffusion XL image on the top left, we progressively compose a slider on top of the previously added stack of sliders. By the end, bottom right, we show the image by composing all 10 sliders.

15. Editing Real Images

Concept sliders can also be used to edit real images. Manually engineering a prompt to generate an image similar to the real image is very difficult. We use null inversion which finetunes the unconditional text embedding in the classifier free guidance during inference. This allows us to find the right setup to turn the real image as a diffusion model generated image. Figure 20 shows Concept Sliders used on real images to precisely control attributes in them.

<https://null-text-inversion.github.io>



Figure 19. Concept Sliders can be composed for a more nuanced and complex control over attributes in an image. From stable diffusion XL image on the top left, we progressively compose a slider on top of the previously added stack of sliders. By the end, bottom right, we show the image by composing all 10 sliders.

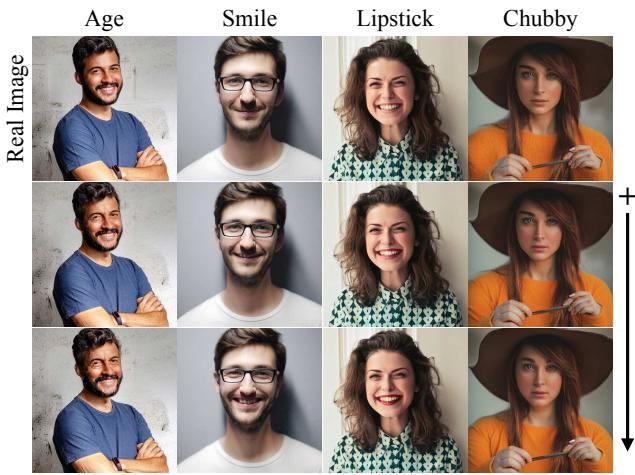


Figure 20. Concept Sliders can be used to edit real images. We use null inversion method to convert real image as a diffusion model generated image. We then run our Concept Sliders on that generation to enable precise control of concepts.

16. Sliders to Improve Image Quality

We provide more qualitative examples for "fix hands" slider in Figure 21. We also show additional examples for the "repair" slider in Figure 22-24

16.1. Details about User Studies

We conduct two human evaluations analyzing our "repair" and "fix hands" sliders. For "fix hands", we generate 150 images each from SDXL and our slider using matched seeds and prompts. We randomly show each image to an odd number users and have them select issues with the hands: 1) misplaced/distorted fingers, 2) incorrect number of fingers, 3) none. as shown in Figure 25 62% of the 150 SDXL images have hand issues as rated by a majority of users. In contrast, only 22% of our method's images have hand issues, validating effectiveness of our fine-grained control.



Figure 21. Concept Sliders can be used to fix common distortions in diffusion model generated images. We demonstrate "Fix Hands" slider that can fix distorted hands.

We conduct an A/B test to evaluate the efficacy of our proposed "repair" slider. The test set consists of 300 image pairs (Fig. 26), where each pair contains an original image alongside the output of our method when applied to that image with the same random seed. The left/right placement of these two images is randomized. Through an online user study, we task raters to select the image in each pair that exhibits fewer flaws or distortions, and to describe the reasoning behind their choice as a sanity check. For example, one rater selected the original image in Fig. 22.a, commenting that "*The left side image is not realistic because the chair is distorted.*". Similarly a user commented "*Giraffes heads are separate unlikely in other image*" for Fig. 23.c. Across all 300 pairs, our "repair" slider output is preferred as having fewer artifacts by 80.39% of raters. This demonstrates that the slider effectively reduces defects relative to the original. We manually filter out responses with generic comments (e.g., "more realistic"), as the sanity check prompts raters for specific reasons. After this filtering, 250 pairs remain for analysis.



Figure 22. Concept Sliders can be used to fix common distortions in diffusion model generated images. The repair slider enables the model to generate images that are more realistic and undistorted.



Figure 23. Concept Sliders can be used to fix common distortions in diffusion model generated images. The repair slider enables the model to generate images that are more realistic and undistorted.

Repair Slider



Figure 24. Concept Sliders can be used to fix common distortions in diffusion model generated images. The repair slider enables the model to generate images that are more realistic and undistorted.

Issues with hands in the Images

Requester: Rohit Gandikota Reward: \$0.03 per task Tasks available: 0 Duration: 3 Minutes

Qualifications Required: Masters has been granted

Instructions Shortcuts Are there too many or too few finger? Are the Fingers in wrong place?



Select an option

Too many/ Too few Fingers	1
Fingers in wrong place	2
None of the Above	3

Figure 25. User study interface on Amazon Mechanical Turk. Users are shown images randomly sampled from either SDXL or our “fix hands” slider method, and asked to identify hand issues or mark the image as free of errors. Aggregate ratings validate localization capability of our finger control sliders. For the example shown above, users chose the option “Fingers in wrong place”

Which Image is more Realistic?	Reward: \$0.03 per task	Tasks available: 0	Duration: 3 Minutes
Requester: Rohit Gandikota			
Qualifications Required: Masters has been granted			

Instructions: Given two images, choose the image that is more realistic and describe in a short sentence why you think it is. Remember it is a comparative study (both images could be unrealistic, but choose the better one)

Example Answers : "the wall in the background has some distortion", "the windows in the image do not look natural"



Which Image is More Realistic? (Relative to each other) **Please take your time and carefully analyse the image**

- Left-side Image is Realistic Right-side Image is Realistic

Tell us in few words why the other image is not realistic ...

Submit

Figure 26. Interface for our "realistic" slider user study. Users are shown an original SDXL image and the corresponding output from our slider, with left/right placement randomized. Users select the image they find more photorealistic and describe their rationale as a sanity check. For example, one user selected the slider image as more realistic in the shown example, commenting "*The black-haired boy's face, right arm and left foot are distorted in right image.*" Another user also chose the slider output, noting "*The right side image has a floating head*". Asking raters to give reasons aims to reduce random selections.



Figure 27. We demonstrate the effects of modifying an image with different sliders like “curly hair”, “surprised”, “chubby”. Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure.

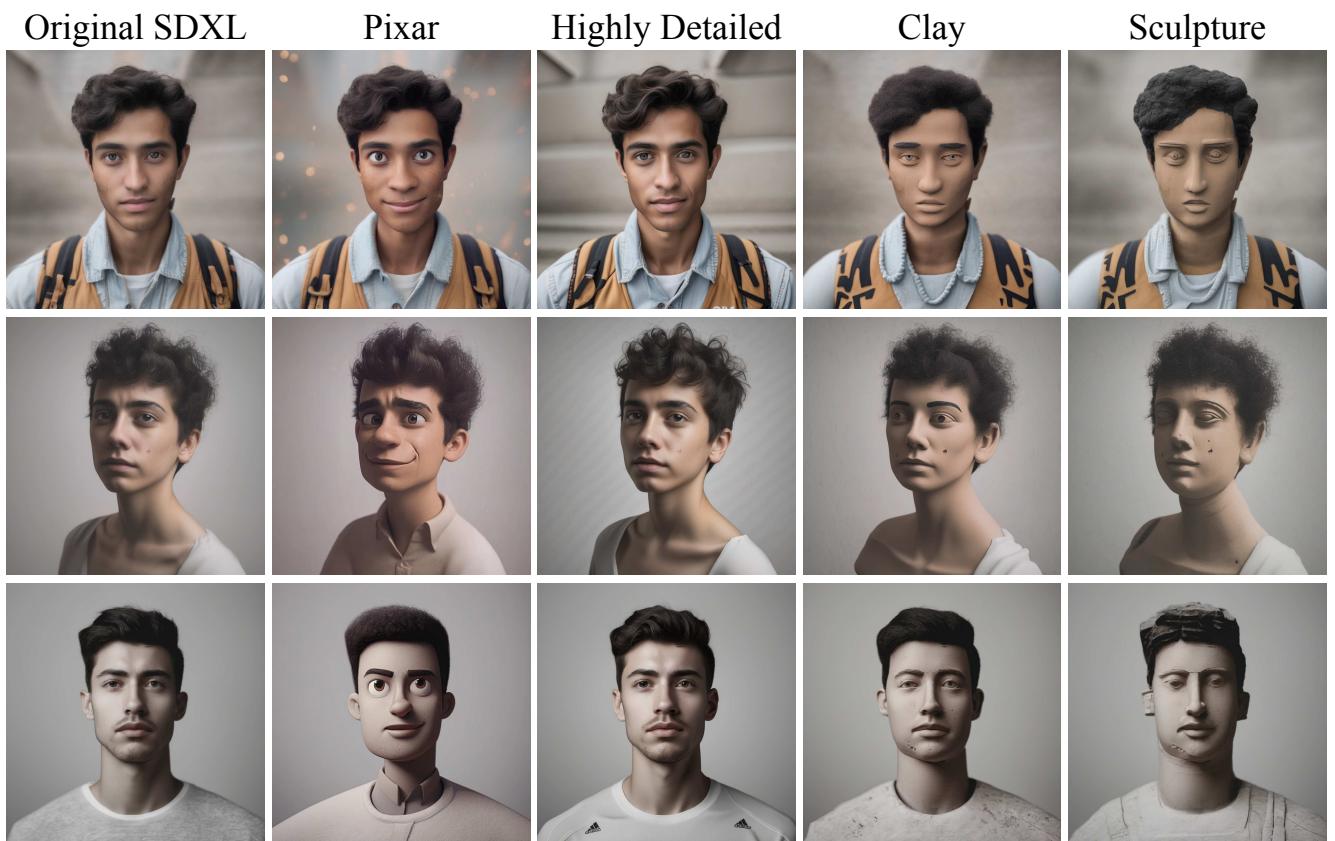


Figure 28. We demonstrate style sliders for "pixar", "realistic details", "clay", and "sculpture". Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure.

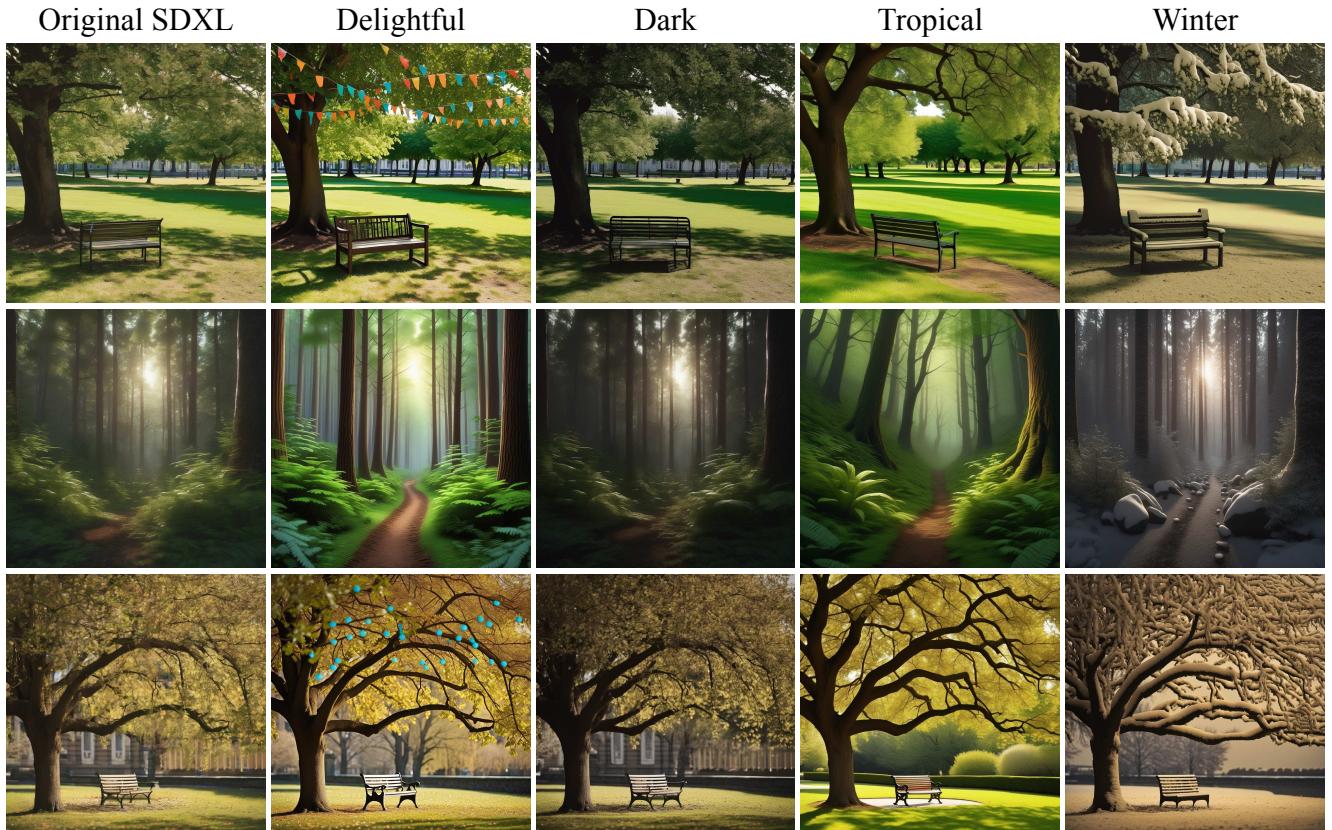


Figure 29. We demonstrate weather sliders for "delightful", "dark", "tropical", and "winter". For delightful, we notice that the model sometimes make the weather bright or adds festive decorations. For tropical, it adds tropical plants and trees. Finally, for winter, it adds snow.

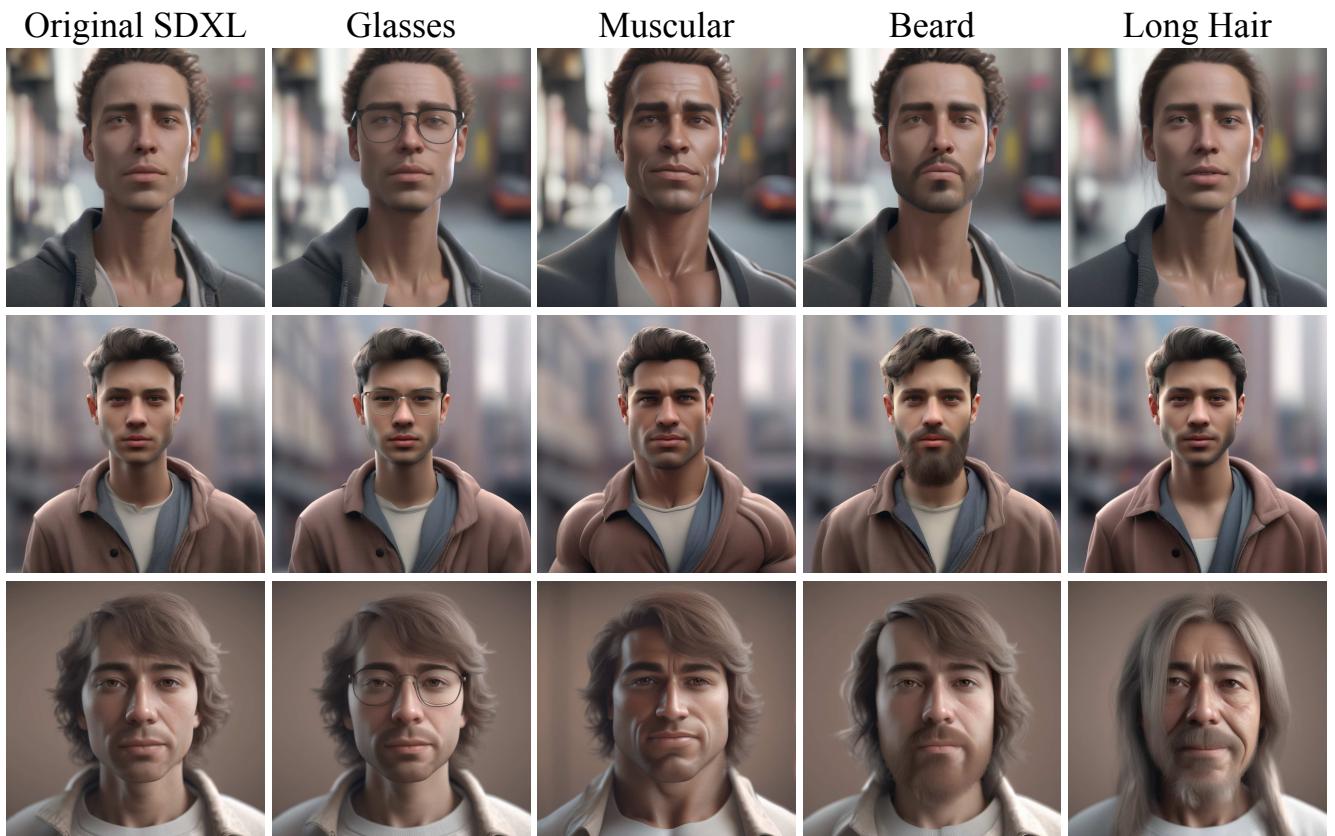


Figure 30. We demonstrate sliders to add attributes to people like "glasses", "muscles", "beard", and "long hair". Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure.



Figure 31. We demonstrate sliders to control attributes of vehicles like “rusty”, “futuristic”, “damaged”. Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure.

Original SDXL



Royal



Modern



Good Interior



Figure 32. Our sliders can also be used to control styles of furniture like “royal”, “Modern”. Our text-based sliders allow precise editing of desired attributes during image generation while maintaining the overall structure.