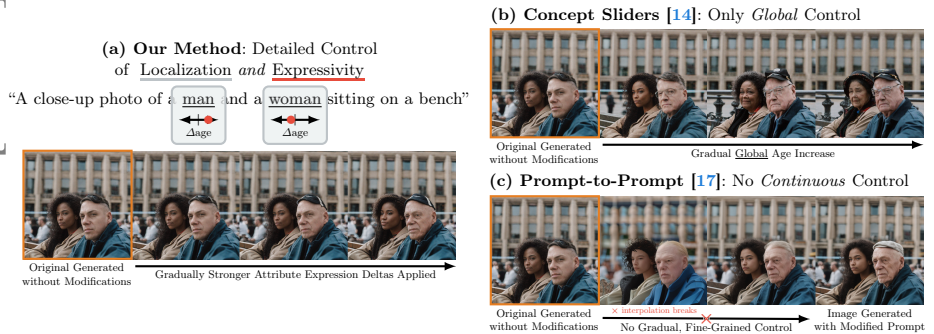


# Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions

Stefan Andreas Baumann<sup>1</sup>, Felix Krause<sup>1</sup>, Michael Neumayr<sup>2</sup>, Nick Stracke<sup>1</sup>, Vincent Tao Hu<sup>1</sup>, and Björn Ommer<sup>1</sup>

<sup>1</sup> CompVis @ LMU Munich, <sup>2</sup> MCML <sup>2</sup> TU Munich  
 {stefan.baumann,b.ommer}@lmu.de



**Fig. 1:** (a) Our method augments the prompt input of image generation models with *fine-grained control* of attribute expression in generated images (unmodified images are marked in **orange**) in a *subject-specific* manner *without additional cost* during generation. (b, c) Previous methods only allow *either* fine-grained expression control or fine-grained localization when starting from the image generated from a basic prompt.

**Abstract.** In recent years, advances in text-to-image (T2I) diffusion models have substantially elevated the quality of their generated images. However, achieving fine-grained control over attributes remains a challenge due to the limitations of natural language prompts (such as no continuous set of intermediate descriptions existing between “person” and “old person”). Even though many methods were introduced that augment the model or generation process to enable such control, methods that do not require a fixed reference image are limited to either enabling global fine-grained attribute expression control or coarse attribute expression control localized to specific subjects, not both simultaneously. We show that there exist directions in the commonly used token-level CLIP text embeddings that enable fine-grained subject-specific control of high-level attributes in text-to-image models. Based on this observation, we introduce one efficient optimization-free and one robust optimization-based method to identify these directions for specific attributes from contrastive text prompts. We demonstrate that these directions can be used to augment the prompt text input with fine-grained control over attributes of specific subjects in a compositional manner (control over multiple attributes of a single subject) without having to adapt the diffusion model.

Project page: [compvis.github.io/attribute-control](https://compvis.github.io/attribute-control).

Code is available at [github.com/CompVis/attribute-control](https://github.com/CompVis/attribute-control).

# 1 Introduction

Text-to-image (T2I) models have recently seen considerable progress in their capabilities and the quality of their generated images. However, a persistent challenge lies in achieving fine-grained control over the generated images, particularly concerning attribute expression. This can be attributed to the limitations of natural language, which does not allow for the very fine-grained description of attribute expression in a way that diffusion models reliably understand. Unlike image editing use cases, where a base image is given, and masks can be provided to adapt a global method to only affect a target subject, we target the pure generation use case, where no reference image is given. This means that subject instances can only be identified via the prompt. Even though many methods (see Tab. 1) were introduced that augment the model or generation process to enable such control without per-image optimization (in contrast to Imagic [22]), they are limited to either enabling global fine-grained attribute control [14, 23, 25, 27, 28] or coarse attribute control localized to specific subjects [17, 49], not both simultaneously.

Starting from a simple prompt for a T2I diffusion model, our goal is to influence the generation process in a fine-grained manner, in both localization and magnitude. More specifically, in a prompt such as “a photo of a woman standing next to a man in front of their car”, we want to influence attributes of each subject – “woman”, “man”, and “car” – separately and with fine-grained control over their individual attribute expression. Typically, one would add these attributes to the subjects in the prompt (“old person”), but this approach yields only very coarse control of these attributes. Additionally, when composing multiple of such attributes in the prompt, diffusion models often ignore a subset of them [12, 28].

We show that there exist directions in the commonly used token-level CLIP [39] text embeddings that enable fine-grained subject-specific control of high-level attributes in T2I models, as opposed to inversion methods that learn instance information [13, 32]. Based on this observation, we introduce methods to identify these directions for specific attributes and show how they can be used to augment the prompt text input with fine-grained continuous control over attributes of specific subjects in a compositional (stackable) manner without adapting the diffusion model or incurring additional costs during generation.

We summarize our main contributions as follows:

- We show that token-level edit directions that allow fine-grained control of subject-specific attributes exist in common CLIP text embeddings and that diffusion models are capable of interpreting them
- We show that T2I diffusion models are capable of backpropagating *high-level semantic* concepts to their text embedding input as adaptations of existing embeddings using just the reconstruction loss objective on a single image
- We introduce two approaches for identifying those directions for specific attributes or concepts from contrasting text prompts describing these concepts, one simple optimization-free method and one optimization-based one that identifies more robust directions

- We show that these token-level edit directions enable fine-grained, subject-specific, compositional control of attributes and concepts in common diffusion models

## 2 Related Work

This section provides an overview of work related to our method. We also compare our method with others in Tab. 1.

**Influencing the Generation Process of Diffusion Models** Since Diffusion Models do not provide the same ordered and interpretable latent space as Generative Adversarial Networks [8, 15, 40] do, a considerable research interest has formed to provide Diffusion Models with greater ability to control visual detail. A common approach for this is to directly use new or adapt existing neural features to guide the generation process towards semantic changes [5, 14, 17, 22, 49]. This allows for fine-grained and even non-text-based instructions to appropriately affect the sample, like interpolating between specific concepts [5, 14, 17, 22], subject-specific control [49] or entirely erasing concepts [26]. Our work encompasses the interpolating abilities as well as subject-specific control in complex samples, therefore combining two fundamental properties that have been mutually exclusive for generic methods so far.

*Image Editing.* One special case of influencing the generation process of diffusion models is the case where a reference image is given. Here, the expectation typically is that only a few aspects of the image should be altered. Numerous prior works [4, 6, 17, 31, 51] have shown difficulties properly disentangling multiple concepts in a sample to prevent global changes and enable a user to locally concentrate edits. To approach this problem, one can directly integrate specific changes into the image using the reverse diffusion process [31], by inversion [32, 35] or by marking regions using edit masks [30, 44] to enforce localization.

**Instance Personalization** Another related task is personalizing diffusion models to enable them to generate images that contain specific instances of subjects. Approaches based on finetuning include DreamBooth [43], which leverages a prior preservation loss to adapt the visual backbone, while approaches like Textual Inversion [13] restrict the adaptation of the model to an added embedding vector to be optimized to represent a specific instance or concept. Furthermore Kumari et al. [24] propose CustomDiffusion for efficient training of multiple concepts by finetuning only cross-attention layers. Moreover, gaining enhanced control in scenarios where samples exhibit complex compositional structures remains a significant challenge for text-to-image models [47]. A prominent strategy to achieve controllable image synthesis involves the use of energy-based models, as suggested by several previous works [27, 28, 33].

**Global Edit Directions in CLIP Space** A wide range of previous works [1–3, 36] investigated the use of the CLIP embedding space [39] to guide the image generation process of StyleGAN [21]. These studies identify directions within CLIP’s embedding space that correspond to global semantic changes and utilize these directions to steer the generation process.

**Table 1:** High-level comparison of our method with other methods that allow attribute control in T2I diffusion models.

Method	Continuous Fine-grained Attribute Control	Subject- Specific	Additional Computational Cost during Generation	Trained Components	Captures Correlations
<i>Editing Methods (require reference image)</i>					
DiffusionCLIP [23]	✓	✗	0	Full Model Finetune	✓
Null-text Inversion [32] + Prompt-to-Prompt [17]	≈ ✗ <sup>1</sup>	✓	Null-text Inversion & 2× Sampling	Unconditional Embeddings	≈ ✗
Imagic [22]	✓	✓	$\mathbf{e}_{gt}$ Optimization & Full Finetune	✗	≈ ✗
Dynamic Prompt Learning [49]	✗	✓	Token Optimization & Null-text Inversion	Dynamic Token Set	≈ ✓
iEdit [4]	✗	✗	0	Full Model Training	≈ ✓
<i>Generic Methods</i>					
Concept Sliders [14]	✓	✗	≈ 0	LoRA [20] Adaptors	✓
Prompt-to-Prompt [17]	≈ ✗ <sup>1</sup>	✓	2× Sampling	✗	✗
HMC sampling from [12]	✗	✓	Costly MCMC Sampling	✗	✓
Asryp [25]	✓	✗	≈ 0	Implicit Functions $f_t$	✓
<b>Our CLIP Difference Deltas</b>	✓	✓	0	✗	✓
<b>Our Learned Deltas</b>	✓	✓	0	Edit Deltas $\Delta\mathbf{e}$	✓
<b>Our Learned Deltas</b> + Delayed Application [14, 31]	✓	✓	0	Edit Deltas $\Delta\mathbf{e}$	≈ ✗
<b>Our Learned Deltas</b> + Prompt-to-Prompt [17]	✓	✓	2× Sampling	Edit Deltas $\Delta\mathbf{e}$	✗

### 3 Method

T2I Diffusion models [18, 42] model a reverse diffusion process  $p_\theta(\mathbf{x}_{0:T}|\mathbf{c})$  that enables sampling from the distribution of images  $p_\theta(\mathbf{x}_0|\mathbf{c})$  given a text prompt condition  $\mathbf{c}$  and a Gaussian noise sample  $\mathbf{x}_T$ . They iteratively denoise  $\mathbf{x}_T$  using a diffusion model  $\hat{\mathbf{x}}_{0,\theta}(\mathbf{x}_t|\mathbf{c}, t)$  which (implicitly) predicts the clean sample given a noised image. This conditioning  $\mathbf{c}$  is typically obtained using a CLIP [39] text encoder  $\mathcal{E}_{\text{CLIP}}$  as a tokenwise embedding  $\mathbf{e} = \mathcal{E}_{\text{CLIP}}(\text{prompt})$ .

#### Exerting Control over the Generation Process in Diffusion Models

We aim to influence the generated samples  $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{e})$ , more specifically the expression  $\text{expr}(A_i)$  of specific attributes  $A_i \in \mathcal{A}$  of a specific subject  $S_j \in \mathcal{S}$ . This subject is described in the prompt at  $\text{prompt}_{[S_j]}$  (with the corresponding embedding  $\mathbf{e}_{[S_j]}$ ) and is located in the generated image in  $\mathbf{x}_{0,[S_j]}$ . The location of the subject in the generated image  $\mathbf{x}_{0,[S_j]}$  is generally not known ahead of time, as it is dynamically determined during the generation process. Specifically, we aim for a change of the expression of  $A_i$  of subject  $S_j$ .

To control the expression of an attribute  $A_i$ , the nature of this image generation approach enables a limited number of aspects to target for influencing the generation process. To control the global expression  $\text{expr}_{\text{global}}(A_i)$ , one option is to modify the diffusion model by re-defining the expression of attributes on

<sup>1</sup> Prompt-to-Prompt does support adding adjectives such as “old” in front of a target subject, enabling coarse attribute modulation. Starting from this point, the weight of the new adjective can be modulated, but this does not suffice to reliably achieve continuous control from the starting image (see Fig. 7c).

a global level [14, 23, 25]. Another possibility is to directly modify the starting noise latent  $\mathbf{x}_T$  [25] or the intermediate noise latents  $\mathbf{x}_t$ , either using dynamically predicted [11, 12, 19, 27, 28, 33] or fixed [5] auxiliary directions. Alternatively, one can also influence how the diffusion model interprets the prompt embedding  $\mathbf{e}$  [17]. This allows for instance specificity through the prompt but does not work well for continuous attribute expression control (see Fig. 7c). Finally, one can directly modify the prompt embedding  $\mathbf{e}$ , influencing the text encoder  $\mathcal{E}_{\text{CLIP}}$ , or modifying the original tokens. Methods like [13, 16, 32, 48] previously investigated this approach for inserting instance appearance information, but not for fine-grained attribute control.

As it already directly enables instance-specific coarse attribute control via text, the tokenwise prompt embedding  $\mathbf{e}$  is a natural choice as the basis for a method for fine-grained attribute control, both in expression and instance-specificity. For this to be practical, multiple requirements have to be fulfilled:

- i) The diffusion model has to be capable of interpreting modified prompt embeddings  $\mathbf{e}'$  that do not directly correspond to a possible text prompt.
- ii) Fine-grained changes of the token-wise prompt embedding have to be localized in what they affect (specifically relating to a subject  $S_j$  in both prompt embedding  $\mathbf{e}$  and image  $\mathbf{x}_0$ ). This localization has to be discoverable and interpretable.
- iii) The tokenwise CLIP embedding space, as interpreted by the diffusion model, has to locally (around anchor points from real words) approximate a Euclidean manifold that behaves similarly across similar anchor points to enable composable smooth local edits along fixed category-specific directions  $\Delta\mathbf{e}$ .
- iv) These directions  $\Delta\mathbf{e}$  have to be practically discoverable.

We investigate requirements i-iii) in Sec. 3.1 and iv) in Sec. 3.2 and Sec. 3.3.

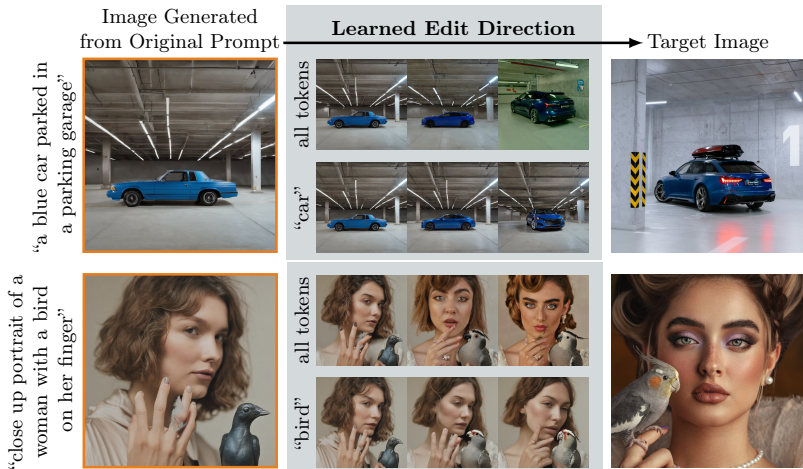
### 3.1 Learning Semantic Edits from Text/Image Pairs

Let us now investigate whether the previously mentioned conditions i-iii) are fulfilled, and semantic directions that can be applied on top of embeddings of actual captions exist in the tokenwise CLIP embedding space.

A wide range of previous works [13, 16, 32, 48] found that the reconstruction loss of a pre-trained T2I diffusion model can be used to backpropagate instance appearance information to the prompt embedding. These instance text embeddings can then enable various personalization and image editing use cases in the generation process. This implies that the first condition – the diffusion model can interpret points that do not exactly lie in the text embedding space that the CLIP model provides – is given. Otherwise, simple gradient descent-based learning of very specific instance appearance information would not be plausible.

Although the objective is only based on pixel-wise reconstruction information, we find that this general approach can also learn semantic information directly. Using a single image/caption pair ( $\mathbf{x}_0$ , prompt), we apply a random amount of noise to the image and backpropagate the reconstruction loss

$$\mathcal{L}(\mathbf{x}_0, \mathbf{e} + \Delta\mathbf{e}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I}), t \sim \mathcal{U}(0, T)} \left[ w(t) \|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon | \mathbf{e} + \Delta\mathbf{e}, t)\|_2^2 \right] \quad (1)$$



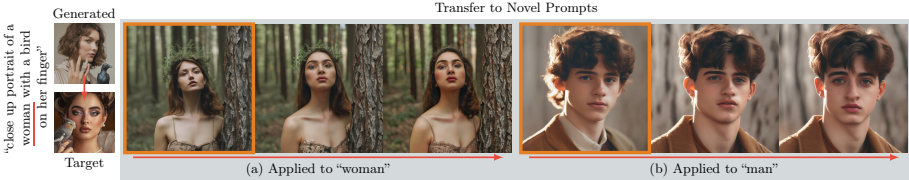
**Fig. 2:** We observe that starting from a target image (rightmost column) and a corresponding caption (leftmost column), edits  $\Delta\mathbf{e}$  to the tokenwise prompt embedding  $\mathbf{e}$  can be learned. Starting from the unmodified  $\mathbf{e}$ , these edits go from the generated images (unmodified images are marked in **orange**) towards the target image. Masking this learned edit interpolation to only apply to one subject  $S_j$  (“car”, “bird”) post hoc results in edits primarily affecting that subject.

through the diffusion model  $\hat{\mathbf{x}}_0(\cdot)$ . We update a learnable delta  $\Delta\mathbf{e}$  that is added to the prompt embedding  $\mathbf{e}$  to minimize the regularized<sup>2</sup> reconstruction loss. One crucial detail is that the noise  $\epsilon$  is randomly re-drawn at every step.

We find that this approach, indeed, leads to learned prompt embedding deltas that capture semantic differences between the set of generated images given just the prompt and the target image. Fig. 2 shows that they substantially reduce the semantic gap between the image generated with the original prompt and the target image. Additionally, linear interpolation between the original prompt embedding and the embedding with the delta applied shows a clear semantic progression from the originally generated image toward the target image. This indicates that the tokenwise CLIP embedding space, as interpreted by the diffusion model, is at least locally smooth in semantic meaning (condition iii). However, we also observe “phase changes” where the image changes substantially during a short subset of the interpolation trajectory for large changes, indicating that the embedding space is not globally smooth.

**Subject Specificity of Semantic Edit Deltas.** We now investigate the relationship between these learned tokenwise semantic edit deltas  $\Delta\mathbf{e}$  and the initial prompt. At a high (semantic) level, our training method initially yields an adaptation of the full prompt embedding that substantially closes the gap between the originally generated images and the target images. We find (see

<sup>2</sup> Implemented as weight decay through AdamW [29] on  $\Delta\mathbf{e}$ .



**Fig. 3:** We observe that learned subject edit directions (such as for “woman” on the left, from Fig. 2) learned on one image/text pair can be transferred to other subjects. These can be transferred to novel prompts (unmodified images are marked in **orange**) that mention the same subject (a) or even ones that mention other subjects (b). In both cases, they cause the attributes (lip thickness, eyebrows, make-up, hairdo) of that subject to change in a similar direction.

Fig. 2) that only applying the subject-specific token edit  $\Delta e_{[S_j]}$  of the learned edit delta  $\Delta e$  suffices to obtain a mostly disentangled edit of  $S_j$ . This partial edit is semantically close to the full edit, affecting the rest of the image only minimally. Notably, we mask out the deltas only after training. This implies that our training process associates semantic information to the specific token corresponding to each subject  $S_j$  and that these token-specific deltas  $\Delta e_{[S_j]}$  are directly usable for influencing the subject’s attribute expression  $\text{expr}_{S_j}(\mathcal{A})$  on a semantic level. This provides a simple interpretable way to localize modifications (condition ii).

**Transferability of Semantic Edit Deltas.** Finally, we investigate the transferability of the learned  $\Delta e_{[S_j]}$  to other prompts. Instead of applying the subject-specific  $\Delta e_{[S_j]}$  to  $S_j$  in the same prompt, we transfer deltas learned on one text/image pair to a new prompt. First, we test transferring a delta  $\Delta e_{[S_j]}$  learned on one subject  $S_j$  (such as “woman”) from one prompt to the same word in a new prompt, as shown in Fig. 3a. This results in a change of attribute expression  $\text{expr}_{S'_j}(\mathcal{A})$  on the new subject  $S'_j$  semantically similar to the change seen on the original prompt (see Fig. 2). These deltas can also be transferred to other subjects  $S_k$  of a similar category (such as from “woman” to “man”, see Fig. 3b). There, they result in similar changes to attribute expression  $\text{expr}_{S_k}(\mathcal{A})$  again but keep the underlying subject change (such as the gender change in this example) from the base text token intact. This shows that our previously defined condition iii) is effectively fulfilled.

### 3.2 Identifying Specific Attribute Deltas from Contrastive Prompts

Sec. 3.1 showed that subject-specific attribute modifications can be achieved by modifying the token(s) of the noun corresponding to the subject without requiring additional words. This implies that the CLIP text encoder likely already performs aggregation of semantic attributes into the corresponding subject, which is corroborated by [41]. Otherwise, the diffusion model would likely not learn to

interpret the tokenwise embeddings this way. This is also corroborated by Li et al. [26], who find that concepts will be present in multiple places in the tokenwise embedding instead of being localized to just its specific tokens. Despite adding attributes typically affecting a multitude of token embeddings, our previous findings indicate that just modifying the token  $\mathbf{e}_{[S_j]}$  already suffices to enable substantial semantic changes.

Motivated by this finding, we propose identifying *semantic directions* in the tokenwise embedding space that affect *specific attributes*  $A_i$  from contrastive prompts similar to [14] (such as “a young person” vs. “an old person” for the “age” attribute, both prompts using the same noun for  $S_j$ ). We first obtain the tokenwise CLIP embeddings  $\mathcal{E}_{\text{CLIP}}(\text{prompt}_{A_i,+})$  and  $\mathcal{E}_{\text{CLIP}}(\text{prompt}_{A_i,-})$  for the positive and negative prompts, respectively. Then, we compute the difference between the token embedding of the subject in both embeddings

$$\Delta \mathbf{e}_{A_i} = (\mathcal{E}_{\text{CLIP}}(\text{prompt}_{A_i,+}))_{[S_j]} - (\mathcal{E}_{\text{CLIP}}(\text{prompt}_{A_i,-}))_{[S_j]}. \quad (2)$$

This directly yields a direction that corresponds to the target attribute  $A_i$ . To obtain more robust estimates of this direction, we average it over a multitude of contrastive prompt pairs that describe the same concept.

*Sampling.* During sampling, we simply add these learned deltas  $\Delta \mathbf{e}_{A_i}$  to  $\mathbf{e}_{[S_k]}$  of the target subject(s)  $S_k$  with the desired scale  $\alpha_i$ . We then pass this modified prompt embedding to the diffusion model to generate an image causing no additional computational cost over standard sampling. This yields similar behavior to adding adjectives such as “old” or “young” to the subject in the prompt but additionally enables control on a smooth scale while retaining subject-specificity.

Two examples of these learned directions are shown in Fig. 4. They illustrate that this approach works to identify directions that affect attributes such as a vehicle’s price but also often exhibits unrelated correlations such as car orientation or bike size. As the CLIP text encoder is causal, this approach is also limited to attributes that can be described as a prefix to the target subject noun.

### 3.3 Learning Robust Fine-Grained Attribute Deltas

Inspired by our findings from Sec. 3.1, we introduce a method for targeted, fine-grained, subject-specific control of attributes in T2I generation. We previously found that editing token embeddings of a specific subject  $S_j$  directly modulates its semantic attribute expression  $\text{expr}_{S_j}(\mathcal{A})$  and that these edits can be transferable. Thus, we propose to leverage this approach to introduce fine-grained control into T2I models without having to modify them, that is, learning edit directions  $\Delta \mathbf{e}_{A_i}$  to the embedded prompt  $\mathbf{e}$  that directly correspond to subject-specific fine-grained modulations of the expression  $\text{expr}_{S_j}(A_i)$  of a specific attribute  $A_i$ .

In general, to limit the  $\Delta \mathbf{e}_{A_i}$  to apply to a specific subject  $S_j$ , we only modify the part of  $\mathbf{e}$  that corresponds to  $S_j$ :

$$\mathbf{e}'(\mathbf{e}, \alpha_i \Delta \mathbf{e}_{A_i})_{[S_j]} = \mathbf{e}_{[S_j]} + \alpha_i \Delta \mathbf{e}_{A_i}. \quad (3)$$

This modified embedding is then passed to the diffusion model in place of  $\mathbf{e}$ .





**Fig. 4:** Two examples showing variations along the “vehicle price” direction identified using our CLIP embedding difference method (Sec. 3.2). Unmodified images are marked in **orange**. These directions successfully capture the target attribute and allow for fine-grained modulation but also show unwanted side-effects such as flipping the car’s orientation, which Sec. 3.3 addresses.



**Fig. 5:** The same two examples as in Fig. 4, but with our learned edit deltas instead of our CLIP embedding differences. Unmodified images are marked in **orange**. Spurious correlations are reduced substantially.

**Training** During training, we utilize the diffusion model’s world knowledge, specifically about which changes in the generated images  $\mathbf{x}_0$  correspond to modulations of specific high-level attributes  $A_i$ . We continue using contrastive text prompts that describe the target attribute  $A_i$  to elicit a fine-grained direction in the model’s noise prediction space that corresponds to that attribute. We then backpropagate this direction through the diffusion model, distilling it into our delta  $\Delta \mathbf{e}_{A_i}$ . Like in [14], this approach learns robust deltas  $\Delta \mathbf{e}_{A_i}$  from just a set of contrastive prompts and does not require training images.

For each optimization step, we start by generating a random image  $\mathbf{x}_{0,a}$  from the target category using a random prompt (such as “a photo of a person”) with standard sampling settings. This image will serve as the anchor for the optimization process. Starting from a noised version of this image  $\mathbf{x}_{t,a}$  at a random diffusion timestep  $t$ , we then generate three predictions with the diffusion model: one prediction  $\hat{\mathbf{x}}_{0,a}$  with the original anchor prompt and two predictions  $\hat{\mathbf{x}}_{0,+}$ ,  $\hat{\mathbf{x}}_{0,-}$  with modified prompts. For those, added adjectives either increase (such as “a photo of an old person”) or decrease (such as “a photo of a young person”) the expression of the target attribute  $\text{expr}(A_i)$  (such as age).

As is well-known, classifier-free guidance [19] can effectively merge multiple such predictions into a new one, allowing for continuous application of different



**Fig. 6: Expression of Global Correlations with Different Sampling Methods:**

When changing the age of the woman (age delta on “woman”) starting from the prompt “A photo of a beautiful woman sitting on a sofa in her flat”, the background can be expected to change with her age. Using different sampling methods (b and c) substantially reduces these correlations (see, e.g., the sofa & plants on the shelves). Unmodified images are marked in **orange**. Delta scales  $\alpha_i$  are identical across rows.

predictions. Starting from  $\hat{\mathbf{x}}_{0,a}$ , we can increase or decrease  $\text{expr}(A_i)$  with

$$\hat{\mathbf{x}}_{0,\text{target}}(\alpha_i) = \hat{\mathbf{x}}_{0,a} + \alpha_i \cdot (\hat{\mathbf{x}}_{0,+} - \hat{\mathbf{x}}_{0,-}) \quad (4)$$

Here, the guidance scale  $\alpha_i$  controls both the direction (sign) and magnitude (value) of the attribute expression change. We can then use these targets to train our semantic edit delta  $\Delta \mathbf{e}_{A_i}$ . Randomly sampling  $\alpha_i$  during training allows us to reliably continuously modulate  $A_i$  later on. When minimizing the difference between the diffusion model’s prediction and the target prediction, we then also adjust  $\Delta \mathbf{e}_{A_i}$  by  $\alpha_i$ , leading to our delta training loss formulation:

$$\mathcal{L}_{\text{delta}} = \mathbb{E}_{\alpha_i} \left[ w(t) \|\hat{\mathbf{x}}_{0,\text{target}}(\alpha_i) - \hat{\mathbf{x}}_0(\mathbf{x}_{t,a} | \mathbf{e}'(\mathbf{e}, \alpha_i \Delta \mathbf{e}_{A_i}), t)\|_2^2 \right]. \quad (5)$$

With a set of suitable prompts to describe increasing and decreasing expressions of a target attribute  $A_i$ , this objective learns robust attribute deltas  $\Delta \mathbf{e}_{A_i}$ .

During sampling, we use the same methodology for applying them as in Sec. 3.2. Compared to the results obtained using the CLIP text embedding differences (see Fig. 4), these results (see Fig. 5) also successfully capture the target attribute but exhibit fewer artifacts, such as the car’s orientation flipping and its age changing (see Sec. 4.1).

### 3.4 Global Correlations

Generally, image generation models learn correlations between different parts of the image. As we do not modify the diffusion model itself, directly applying our learned edit deltas  $\Delta \mathbf{e}_{A_i}$  to a specific subject  $S_j$  causes the desired change in attribute expression  $\text{expr}_{S_j}(A)$ , but additionally captures the entanglement between this subject’s attributes and the attributes of the rest of the generated image based on the diffusion model’s world knowledge. This means that this attribute



**Fig. 7: Continuous Attribute Modifications.** Unmodified images are marked in orange. **(a) Variety of Attributes:** Our learned edit deltas can capture a wide range of attributes and allow fine-grained control over their expression. All samples are generated using a delta scale from -2 to 2 without applying either of our two disentanglement strategies. **(b) Zero-Shot Transfer:** Our deltas can be learned on one model (SDXL) and transferred to others (including non-diffusion models) without re-training. **(c) Standard P2P [17]** does not allow for continuous control starting from the original image without adjectives, only from inserted adjectives (“young”, “old”).

change does not only apply to the part of the image  $\mathbf{x}_{0,[S_j]}$  corresponding to  $S_j$  but also the rest of the image (see Fig. 6a). This helps to enable the generation of plausible images, as the entanglement is based on real-world dependencies. Modeling these correlations, however, might not always be desirable.

If the modeling of dependencies between the subject and its surroundings is not desired, we can augment our sampling method (without having to re-learn  $\Delta\mathbf{e}_{A_i}$ ) during inference time. Following [14, 31], simply not applying the edit delta for the first few steps of the diffusion generation process already substantially reduces the expression of these correlations at no additional computational cost. This especially helps to retain the original global image structure, as it is determined in the first steps of the generation process [17]. Fig. 6b shows an example of this. Since our model only alters the text embedding, we can also directly pair it with Prompt-to-Prompt [17] to further improve appearance and structure disentanglement regarding the edit at the cost of doubling the inference cost. Fig. 6c shows the corresponding example.

## 4 Experiments

We evaluate our proposed method in a variety of settings, primarily on Stable Diffusion XL [38], a standard off-the-shelf large-scale T2I diffusion model. We sample with a standard classifier-free guidance [19] scale of 7.5. To test our method, we train a large variety of edit deltas for various different attributes, primarily focused on humans but also including vehicles and furniture. Fig. 7a shows a subset of these edit deltas. Additionally, Fig. 7b shows edit deltas learned on SDXL applied to other (diffusion and non-diffusion) models that use one of the two CLIP text encoders in a zero-shot manner. Quantitatively, we evaluate on human attributes. We sample 25 images per target subject noun, resulting in 100 images for each attribute delta at each scale. In addition to the experiments presented in this section, we also investigate transferring edit deltas learned on one set of nouns to novel nouns in Sec. F.2.

### 4.1 Learned Edit Deltas vs. CLIP Embedding Differences

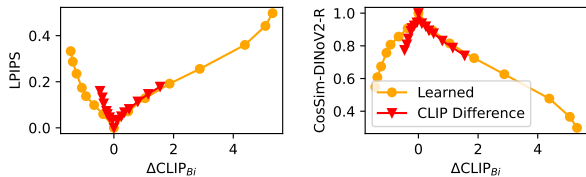
First, we compare the attribute edit directions based on simple CLIP embedding differences (Sec. 3.2) with the learning-based ones (Sec. 3.3). For this comparison, we limit ourselves to a subset of edit deltas that can easily be described by prefixes in front of the subject (“young person”, not “person wearing a colorful outfit”). Here, the causality of the CLIP text encoder inherently limits the method based on embedding differences. Using only the applicable subset avoids this drawback and ensures a fair comparison of the methods. Qualitatively, the directions obtained by taking differences of the CLIP token embeddings exhibit more unwanted side effects. Comparing Fig. 4 and Fig. 5, for example, the car flips its orientation in both directions in the difference-based setup. To quantitatively evaluate the occurrence of undesired side effects, we assess the attained expression  $\text{expr}(A_i)$  of the target attribute  $A_i$  against the change in the image  $\mathbf{x}$ . Following prior art [14, 32], we evaluate attribute expression change  $\Delta\text{expr}(A_i)$  using the change of CLIP similarity of the generated image  $I$  to a target prompt. As our attribute edit deltas are bi-directional, we compare to the prompt  $\text{prompt}_+$  that describes the positive direction (such as “old person”) and the prompt  $\text{prompt}_-$  for the negative direction (“young person”). We compute this bi-directional relative CLIP score as

$$\text{CLIP}_{Bi}(I) = \cos(\text{CLIP}_I(I), \text{CLIP}_T(\text{prompt}_+)) - \cos(\text{CLIP}_I(I), \text{CLIP}_T(\text{prompt}_-)), \quad (6)$$

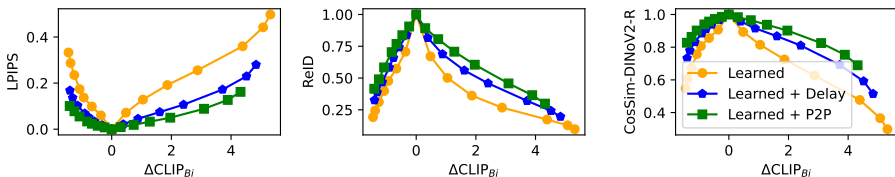
We then measure relative change to a reference image  $I_{ref}$  without delta applied

$$\Delta\text{CLIP}_{Bi}(I, I_{ref}) = \text{CLIP}_{Bi}(I) - \text{CLIP}_{Bi}(I_{ref}). \quad (7)$$

Comparing this score against the overall change in the image yields similar performance of learned and difference-based edit deltas in positive attribute expression directions (Fig. 8). In the negative direction, however, the difference-based method struggles to successfully capture attribute expression without substantially altering the overall appearance of the image.



**Fig. 8:** Attribute expression change (horizontal axis) against overall change in the image (vertical axis) as similarity/difference between the modified image at some scale and the reference image at scale 0. We compare learned edit deltas (Sec. 3.3) with CLIP text difference deltas (Sec. 3.2) at scales from -5 to 5 and find that our learned deltas allow the same attribute expression control at lower image deviation.



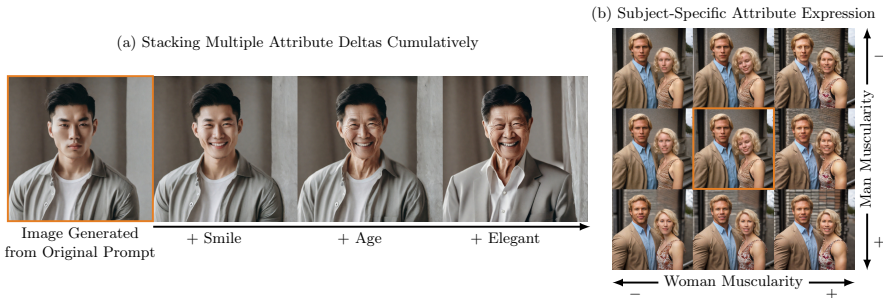
**Fig. 9:** Attribute expression change (horizontal axis) and overall change in the image (vertical axis) for our learned edit deltas (Sec. 3.3) with different sampling methods. For edit deltas targeting people, we also report person identity similarity (ReID).

## 4.2 Global Correlations

To quantitatively evaluate the capturing of global correlations (see Sec. 3.4) when using different sampling methods, we use the same evaluation methodology as in Sec. 4.1. Additionally, we use ArcFace-based [10] person re-identification for edit deltas on human attributes to measure how much the person’s identity is affected by the attribute change. Fig. 9 shows that, for the same level of attribute expression, delayed application of the learned delta helps to substantially reduce unrelated changes in the image (especially as measured by LPIPS) at no additional cost during sampling. Combining our method with Prompt-to-Prompt [17] further reduces these unrelated changes, but at the cost of doubling the sampling time.

## 4.3 Compositional & Subject-Specific Attribute Editing

In the general case, we find that our learned edit deltas are directly composable. This means that multiple edit deltas can be applied to the same subject, where their effects stack yielding fine-grained control over multiple attributes simultaneously as shown in Fig. 10a, with a large number of additional samples shown in appendix Sec. F.4. As they are subject-specific, multiple learned edit deltas can also be applied to different subjects (even ones of the same subject category) by applying each edit delta to just one of the multiple subjects mentioned in a prompt with individual scales as shown in Fig. 10b and Fig. 1a. Numerous additional examples are shown in appendix Sec. F.3.



**Fig. 10:** (a) Multiple attribute edit deltas can be composed simply by adding them. (b) Attribute deltas can be applied to different subjects with different magnitudes. Unmodified images are marked in **orange**.

## 5 Conclusion

This work uncovers the powerful capabilities of the tokenwise CLIP [39] text embedding for exerting control over the image generation process in T2I diffusion models. Instead of just acting as a discrete space of embeddings of words, we find that diffusion models are capable of interpreting local deviations in the tokenwise CLIP text embedding space in semantically meaningful ways. We use this insight to augment the typically rather coarse prompt with fine-grained, continuous control over the attribute expression of specific subjects by identifying semantic directions that correspond to specific attributes. Since we only modify the tokenwise CLIP text embedding along pre-identified directions, we enable more fine-grained manipulation at no additional cost in the generation process.

**Limitations & Future Work** This work is a step towards revealing the hidden capabilities of the text embedding input to common large-scale diffusion models and making them usable in straightforward ways. While our approach works for different off-the-shelf models without modifying them, it is also inherently limited by their capabilities. Specifically, our method inherits the limitation that diffusion models sometimes mix up attributes between different subjects. Complementary methods [7, 41] reduce these problems substantially, and future work could investigate their combination with our method in depth.

**Impact Statement** This work aims to improve the capabilities of text-to-image (T2I) diffusion models by introducing an efficient, simple-to-use method of influencing the expression of attributes of specific subjects in the generated images in a fine-grained manner. While, in general, similar targeted control has been possible before by utilizing editing-based methods on the original generated images, where generated images are inverted and re-generated with changes localized with custom edit masks, this method’s simplicity and efficiency potentially enables such a level of control for a wider audience. This, like other works in the space of improving the control over image synthesis models, carries the risk of further enabling the generation of harmful or deceptive content.

## Acknowledgement

We thank Timy Phan for proofreading and feedback. This project has been supported by the German Federal Ministry for Economic Affairs and Climate Action within the project “NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung”, the German Research Foundation (DFG) project 421703927, and the bidt project KLIMA-MEMES. The authors acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS at JSC.

## References

1. Abdal, R., Zhu, P., Femiani, J., Mitra, N., Wonka, P.: Clip2stylegan: Unsupervised extraction of stylegan edit directions. SIGGRAPH '22 (2022) [4](#)
2. Baba, T., Nishida, K., Nishida, K.: Robust text-driven image editing method that adaptively explores directions in latent spaces of stylegan and clip. arXiv (2023) [4](#), [39](#)
3. Baykal, A.C., Anees, A.B., Ceylan, D., Erdem, E., Erdem, A., Yuret, D.: Clip-guided stylegan inversion for text-driven real image editing. ACM Trans. Graph. **42**(5) (aug 2023) [4](#)
4. Bodur, R., Gundogdu, E., Bhattarai, B., Kim, T.K., Donoser, M., Bazzani, L.: iedit: Localised text-guided image editing with weak supervision. arXiv (2023) [3](#), [4](#)
5. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing text-to-image models using semantic guidance. Advances in Neural Information Processing Systems (2024) [3](#), [5](#)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) [3](#)
7. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph. **42**(4) (jul 2023) [14](#)
8. Chen, X., Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances In Neural Information Processing Systems 29 (2016) [3](#)
9. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024) [41](#)
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019) [13](#), [41](#)
11. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems (2021) [5](#)
12. Du, Y., Durkan, C., Strudel, R., Tenenbaum, J.B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., Grathwohl, W.S.: Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 8489–8510 (2023) [2](#), [4](#), [5](#)

13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023) [2](#), [3](#), [5](#)
14. Gandikota, R., Materzyńska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: Lora adaptors for precise control in diffusion models. arXiv (2023) [1](#), [2](#), [3](#), [4](#), [5](#), [8](#), [9](#), [11](#), [12](#), [19](#), [39](#), [40](#)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014) [3](#)
16. Han, I., Yang, S., Kwon, T., Ye, J.C.: Highly personalized text embedding for image manipulation by stable diffusion. arXiv (2023) [5](#)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023) [1](#), [2](#), [3](#), [4](#), [5](#), [10](#), [11](#), [13](#), [19](#), [20](#), [40](#)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020) [4](#)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) [5](#), [9](#), [12](#), [39](#)
20. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022) [4](#)
21. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) [4](#)
22. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023) [2](#), [3](#), [4](#)
23. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2426–2435 (2022) [2](#), [4](#), [5](#)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023) [3](#)
25. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: The Eleventh International Conference on Learning Representations (2023) [2](#), [4](#), [5](#)
26. Li, S., van de Weijer, J., taihang Hu, Khan, F., Hou, Q., Wang, Y., jian Yang: Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In: The Twelfth International Conference on Learning Representations (2024) [3](#), [8](#)
27. Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. *Advances in Neural Information Processing Systems* **34**, 23166–23178 (2021) [2](#), [3](#), [5](#)
28. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439 (2022) [2](#), [3](#), [5](#)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019) [6](#), [39](#), [40](#)



30. Mao, Q., Chen, L., Gu, Y., Fang, Z., Shou, M.Z.: Mag-edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance. arXiv (2023) [3](#)
31. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022) [3](#), [4](#), [11](#)
32. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6038–6047 (June 2023) [2](#), [3](#), [4](#), [5](#), [12](#)
33. Nie, W., Vahdat, A., Anandkumar, A.: Controllable and compositional generation with latent-space energy-based models. Advances in Neural Information Processing Systems **34**, 13497–13510 (2021) [3](#), [5](#)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024) [41](#)
35. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) [3](#)
36. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2085–2094 (2021) [4](#)
37. Patil, S., Berman, W., Rombach, R., von Platen, P.: amused: An open muse reproduction. arXiv (2024) [11](#)
38. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024) [12](#), [39](#), [41](#)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) [2](#), [4](#), [14](#), [41](#)
40. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016) [3](#)
41. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) [7](#), [14](#)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) [4](#), [11](#)
43. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) [3](#)

44. Simsar, E., Tonioni, A., Xian, Y., Hofmann, T., Tombari, F.: Lime: Localized image editing via attention regularization in diffusion models. arXiv (2023) **3**
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021) **39**
46. Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Muller, M., Lal, V.: Ldm3d: Latent diffusion model for 3d. In: 3DMV: Learning 3D with Multi-View Supervision (CVPRW'23) (2023) **11**
47. Tunanyan, H., Xu, D., Navasardyan, S., Wang, Z., Shi, H.: Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. arXiv (2023) **3**
48. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation. arXiv (2023) **5**
49. Yang, F., Yang, S., Butt, M.A., van de Weijer, J., et al.: Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. Advances in Neural Information Processing Systems **36** (2024) **2, 3, 4**
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) **40**
51. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. arXiv (2023) **3**

# Appendix

F Additional Results . . . . .	19
G Implementation Details . . . . .	39
H Image Copyright . . . . .	41

## F Additional Results

### F.1 Fine-Grained Control

Compared to Concept Sliders [14], which enables fine-grained control over attribute expression on a *global* level (i.e., shared over all instances), our method allows targeting the attribute expression control to specific subjects by selecting the target subject via the prompt. Compared to Prompt-to-Prompt [17] (P2P), which provides subject-specific control like our method in Fig. 11, we find that our method offers substantially more fine-grained control over attribute expression when starting from the image generated from a prompt such as “a photo of a person”. For P2P, we start from the prompt “a photo of a beautiful man”, insert an adjective that describes the target direction (e.g., “a photo of a beautiful old man”), and then re-weight the adjective to control its expression<sup>3</sup>. Here, it is obvious that the initial, coarse changes (such as from a neutral age to “old”) work well, but fine-grained modulations do not in the general case. Specifically, these re-weightings allow slight modulations around “old”, but they do not allow fine-grained control starting from the original image. This makes sense intuitively, as we previously observed (see Sec. 3.2) that attributes are aggregated in the subject by the CLIP text encoder and that the diffusion model uses these aggregated attributes. This means that simply reducing the weight of the adjective can not suffice to enable smooth interpolation between the original attribute expression (as it was in the prompt that did not contain any adjectives relating to the target concept) and the changed attribute expression, at least for models that use text encoders that exhibit this aggregating behavior.

<sup>3</sup> This approach is directly modeled after the one used in the official implementation at [https://github.com/google/prompt-to-prompt/blob/main/prompt-to-prompt\\_stable.ipynb](https://github.com/google/prompt-to-prompt/blob/main/prompt-to-prompt_stable.ipynb) to modulate the inserted word “fried” in the fried potatoes example and uses the same hyperparameters.



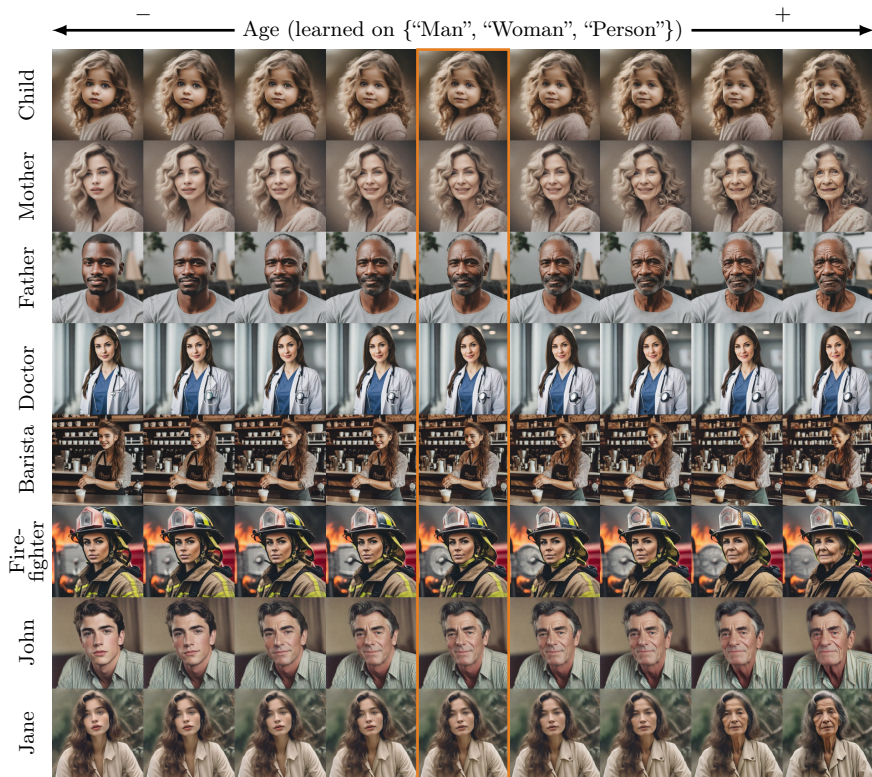
**Fig. 11: Comparison of Capabilities for Continuous Attribute Modulation.**

We compare the capabilities of our learned deltas for continuous attribute modulation with that of Prompt-to-Prompt [17]. The unmodified image is marked in orange. Our samples are generated using attribute deltas being applied with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling). Those for P2P are generated starting from the same image by adding an adjective and then modulating its weight post hoc.

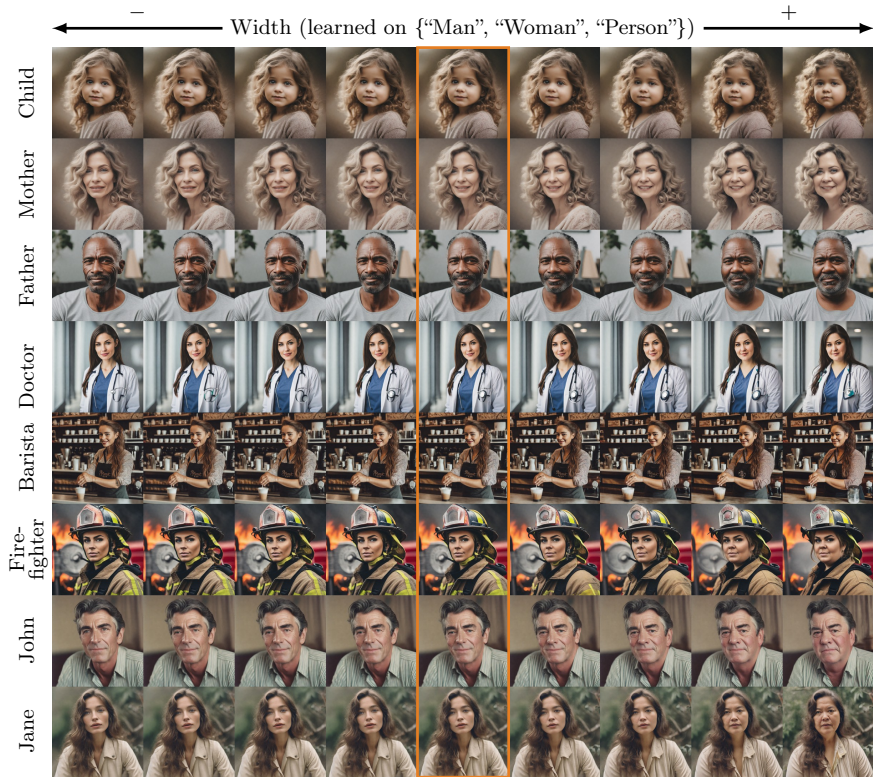
## F.2 Subject Noun Transferability

We investigate how much the attribute deltas can generalize across different nouns that describe the same subject. We generally learn them on a set of different nouns that describe a subject of a specific category (e.g., deltas for people with the words “man”, “woman”, and “person”). However, these words typically do not cover the whole range of possible nouns that can be used to describe subjects of a general category. Ideally, one could learn one delta for one concept, such as age, on a small set of nouns and generalize across all nouns of a category or even to subjects of other categories.

First, we test the generalization of deltas learned for people on “man”, “woman”, and “person” and apply them to increasingly more specific nouns that describe people. Results are shown in Figs. 12 and 13, and all prompts are “a photo of a beautiful <noun>”. As a baseline, we apply them to “child”, “mother”, and “father”, three words that are previously unseen but still describe very high-level sub-categories of people. We find that the learned deltas still work as expected. Similarly, for categories of jobs such as “doctor”, “barista”, or “firefighter”, which are substantially more specific and also substantially affect their clothing and the rest of the image, we find that they also work well. Finally, applying these learned deltas to very specific nouns such as the names “John” and “Jane” also works as expected. This demonstrates that these learned deltas can generalize well across a wide range of unseen nouns describing instances of a specific category, even if they were only learned on a small set of high-level, potential nouns.



**Fig. 12: Subject Noun Transferability.** We stress-test applying deltas that have been learned only on the nouns “man“, “woman“, and “person“ to various other nouns that describe people. The unmodified image is marked in **orange**. All samples are generated using attribute deltas being applied with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 13: Subject Noun Transferability.** We stress-test applying deltas that have been learned only on the nouns “man”, “woman”, and “person” to various other nouns that describe people. The unmodified image is marked in **orange**. All samples are generated using attribute deltas being applied with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).

### F.3 Multi-Subject Attribute Editing

Figs. 14 to 17 show examples of modulating attributes in a subject-specific manner using our learned deltas. These show that various attributes can be applied to subjects individually, even if both subjects are of the same category (e.g., “people”). A slight correlation between, e.g., the age of the man and the age of the woman in Fig. 14 is visible and expected, as the diffusion model also models these dependencies between different subjects in the generated image. By applying both deltas with different strengths, the whole spectrum of combinations can be achieved.

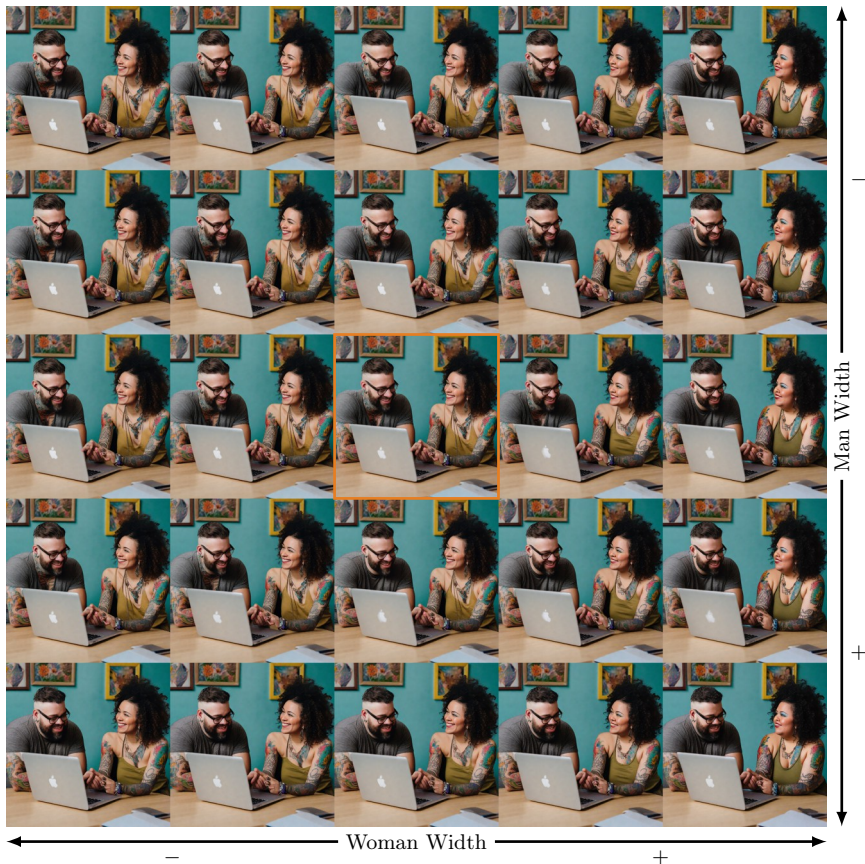


**Fig. 14: Multi-Subject Attribute Modifications.** The unmodified image is marked in orange. All samples are generated using one attribute delta each being applied to the two subjects mentioned in the prompt with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).





**Fig. 15: Multi-Subject Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using one attribute delta each being applied to the two subjects mentioned in the prompt with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 16: Multi-Subject Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using one attribute delta each being applied to the two subjects mentioned in the prompt with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



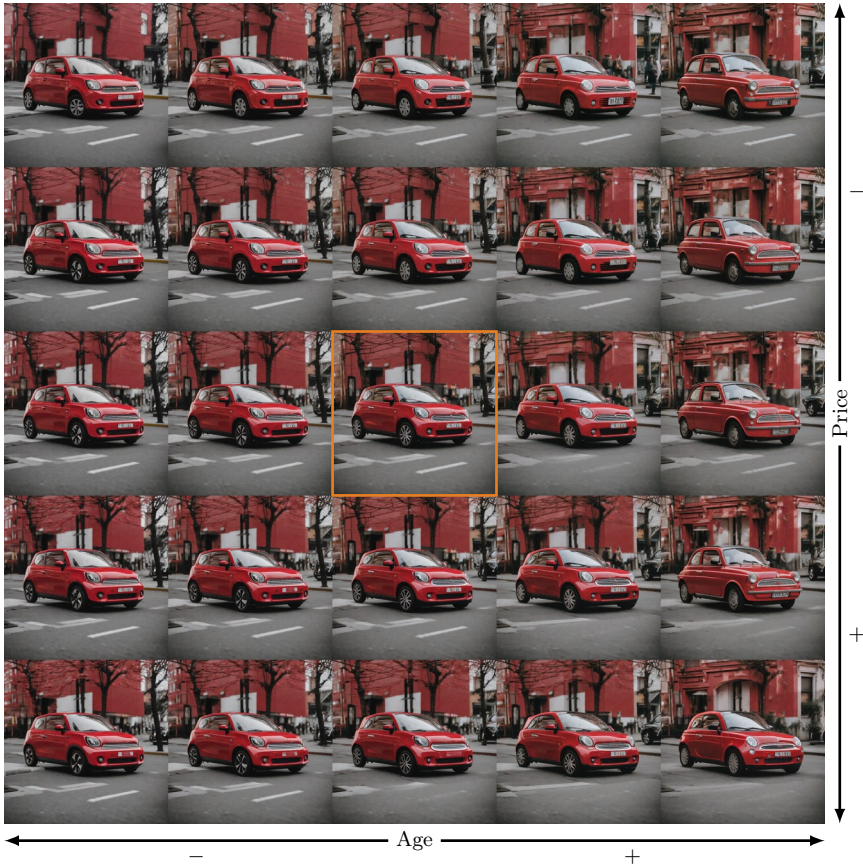
**Fig. 17: Multi-Subject Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using one attribute delta each being applied to the two subjects mentioned in the prompt with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).

## F.4 Compositional Attribute Editing

We show some 2d grids where two attributes are modulated for the same target subject in an additive manner in Figs. 18 to 23. Both attribute deltas interact with each other according to the world knowledge of the diffusion model to produce a realistic image for every combination. This can especially be seen in Fig. 20, where the variant with increased age *and* makeup has substantially reduced wrinkles compared to the version with reduced makeup.



**Fig. 18: Compositional Attribute Modifications.** The unmodified image is marked in orange. All samples are generated using two attribute deltas being applied additively with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



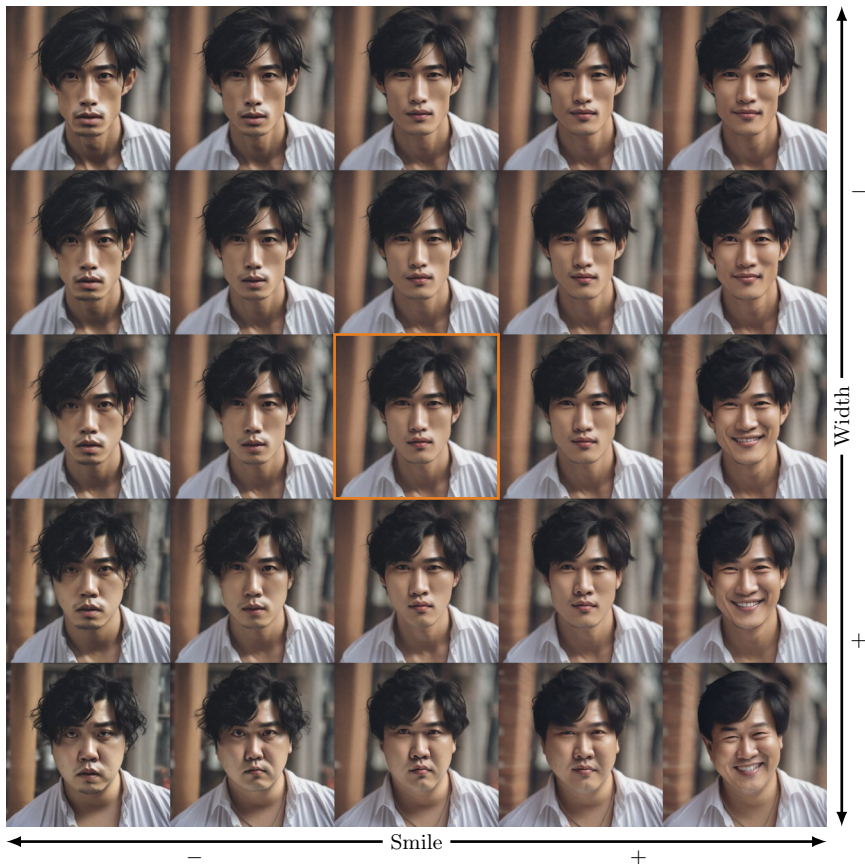
**Fig. 19: Compositional Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using two attribute deltas being applied additively with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 20: Compositional Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using two attribute deltas being applied additively with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 21: Compositional Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using two attribute deltas being applied additively with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 22: Compositional Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using two attribute deltas being applied additively with a linear delta scale from  $-2$  to  $2$  across each, with the deltas being applied after  $10/50$  steps (Delayed Sampling).

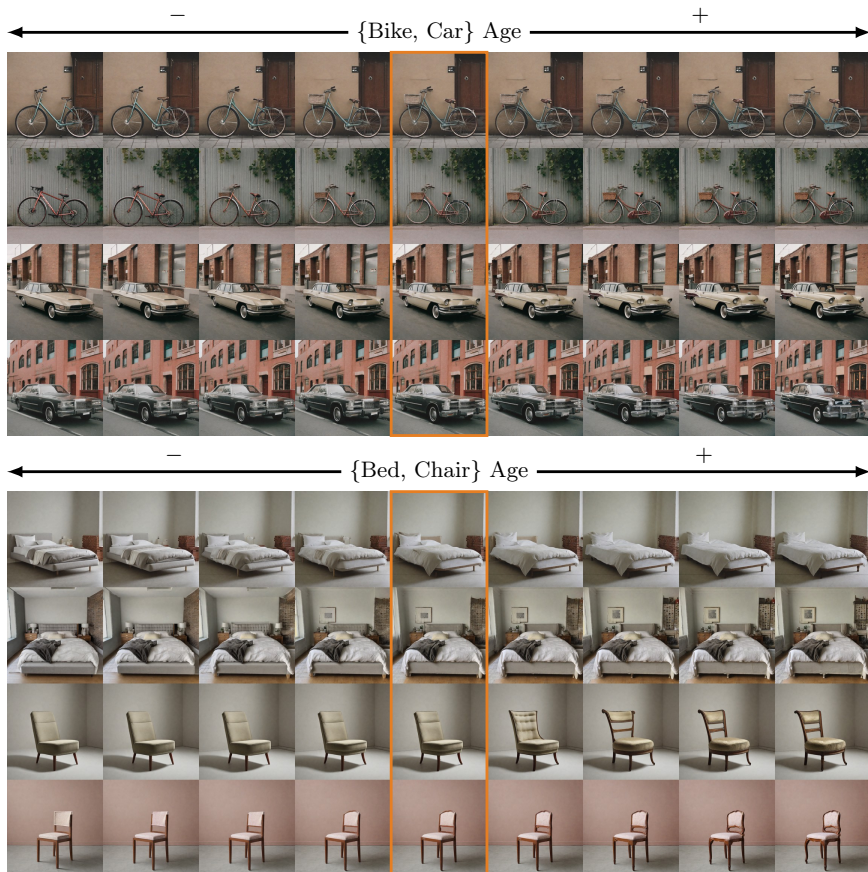




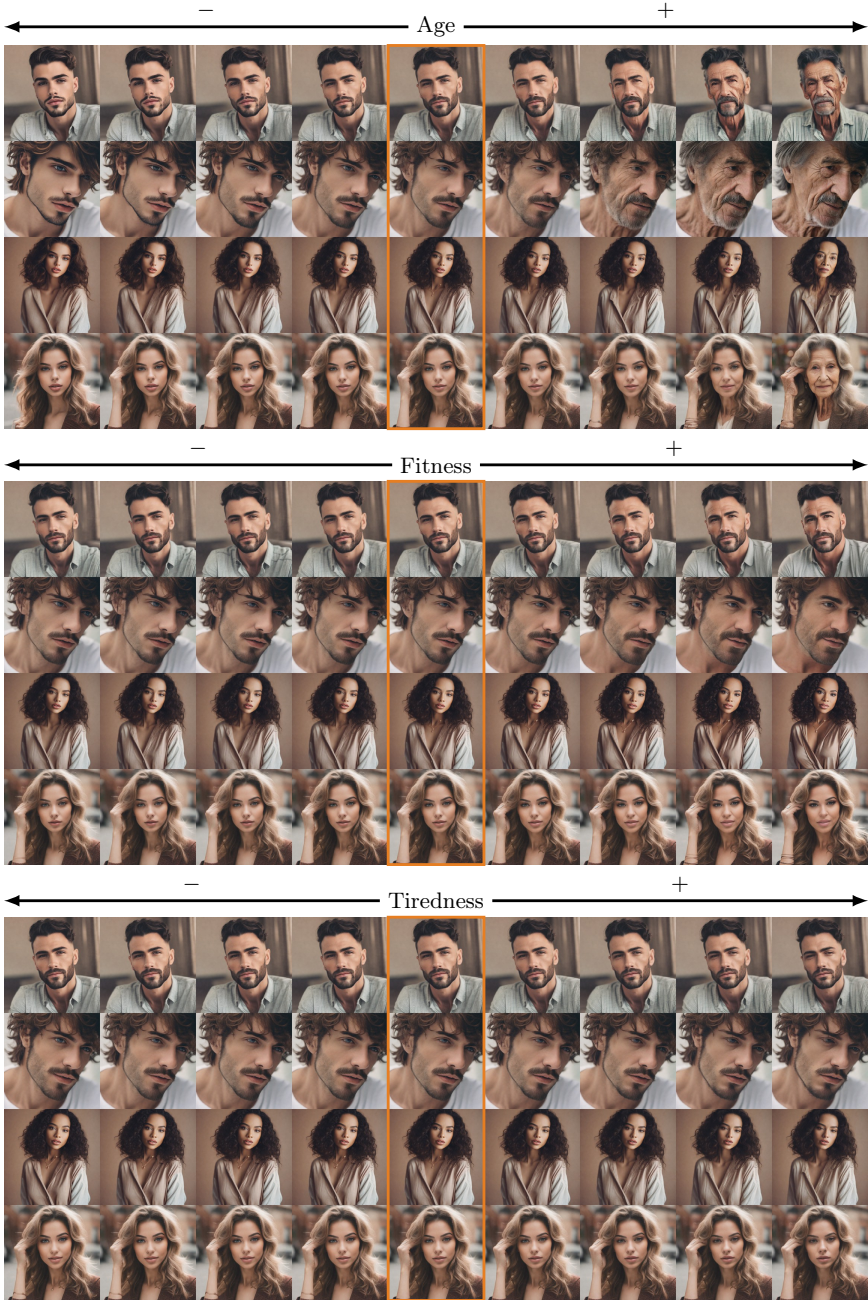
**Fig. 23: Compositional Attribute Modifications.** The unmodified image is marked in **orange**. All samples are generated using two attribute deltas being applied additively with a linear delta scale from -2 to 2 across each, with the deltas being applied after 10/50 steps (Delayed Sampling).

## F.5 Continuous Attribute Modification

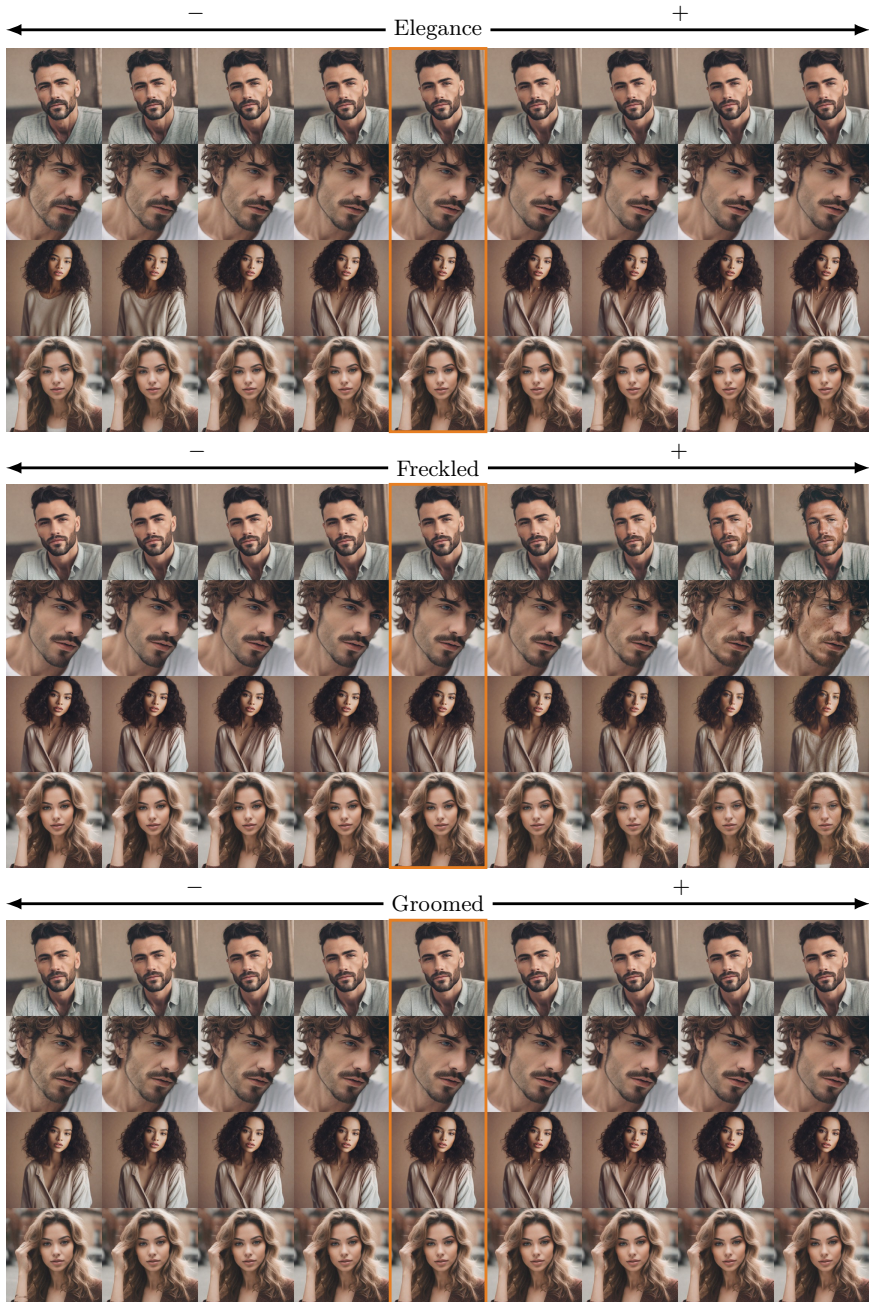
To illustrate the breadth of attributes that can be modulated and how continuous the attribute changes are, we show a range of our learned delta attributes being continuously modulated. Figs. 24 to 27 show examples where attribute deltas are applied with our delayed sampling, Fig. 28 shows attribute deltas applied for the full sampling time. For every category, we re-use the same sample instances as a starting point.



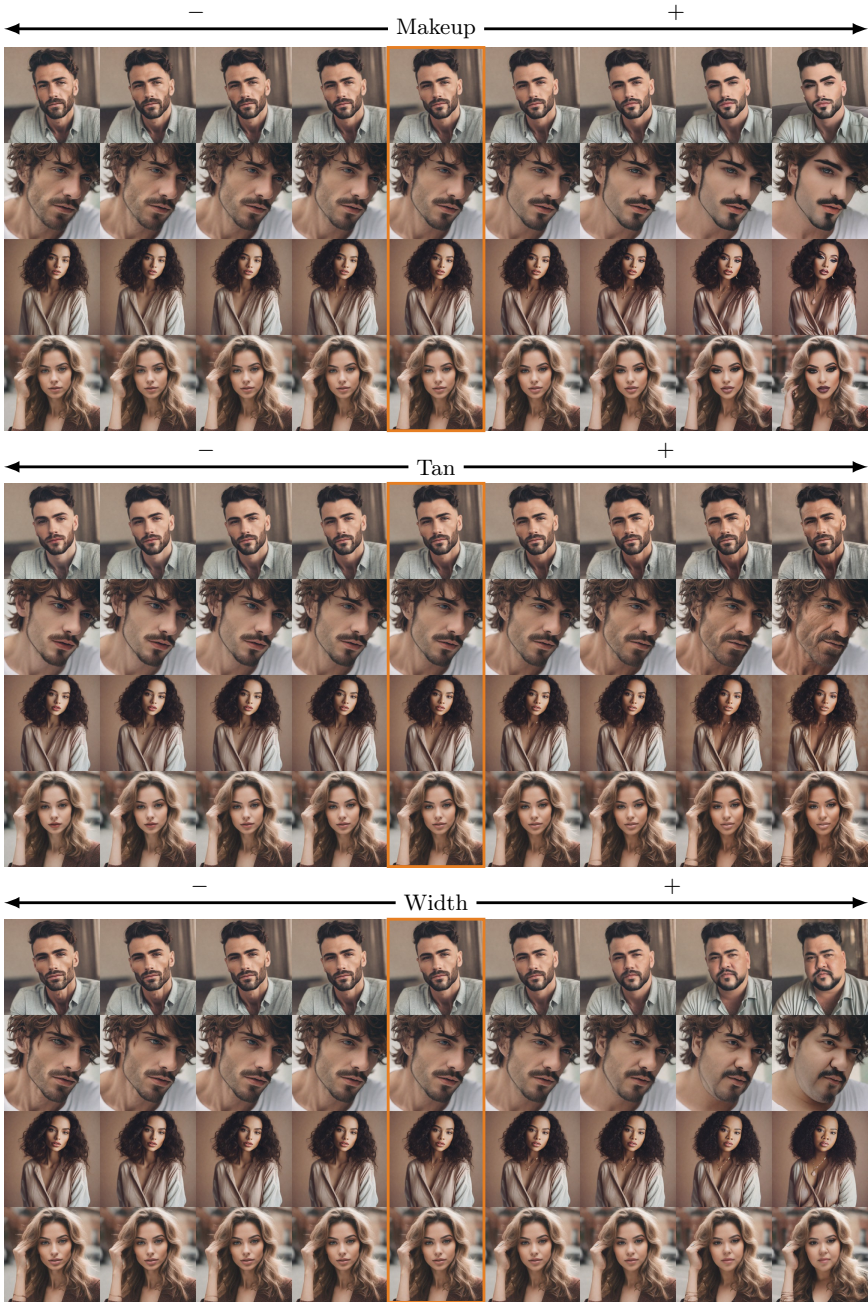
**Fig. 24: Continuous Attribute Modifications.** Unmodified images are marked in orange. All samples are generated using a linear delta scale from -2 to 2, with the deltas being applied after 10/50 steps (Delayed Sampling).



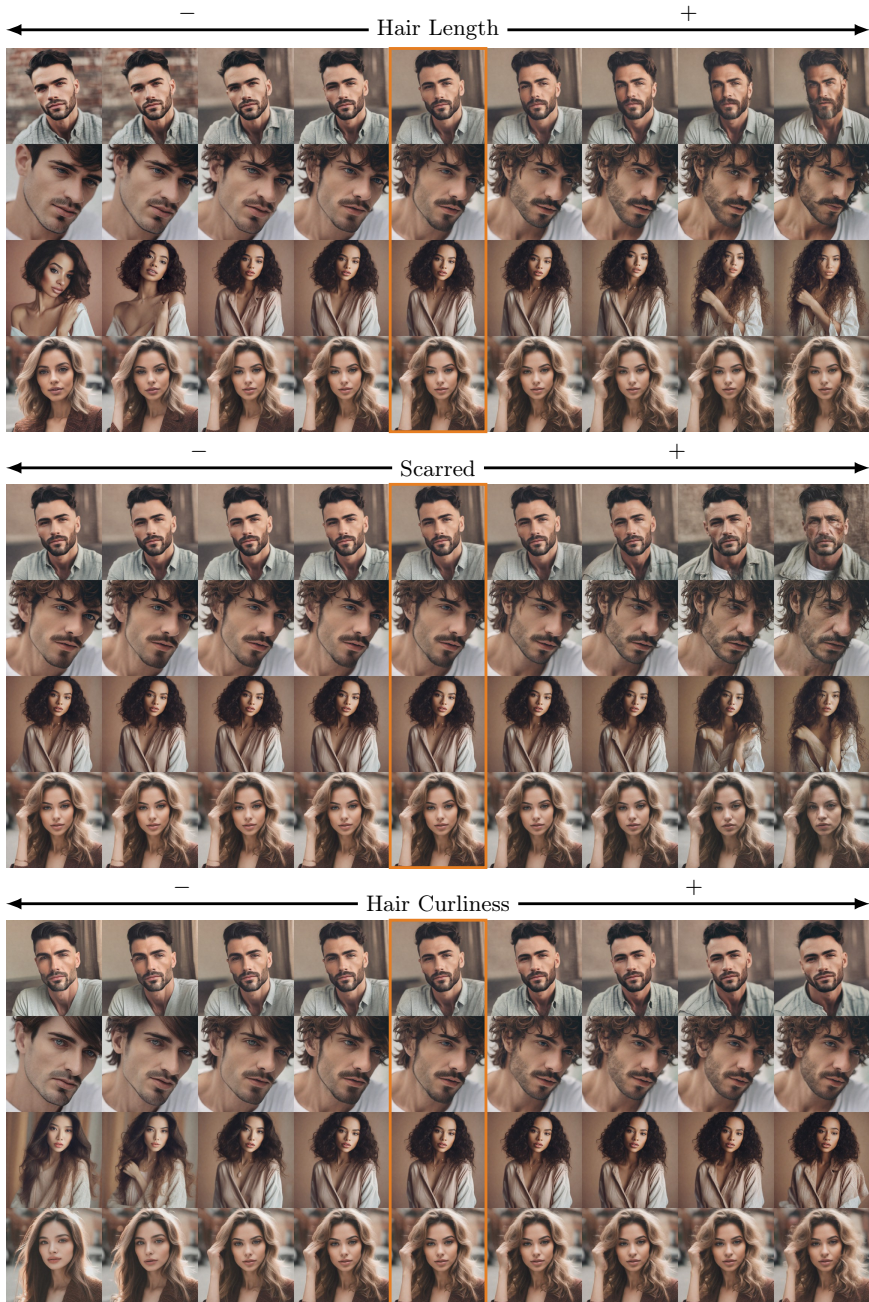
**Fig. 25: Continuous Attribute Modifications.** Unmodified images are marked in orange. All samples are generated using a linear delta scale from -2 to 2, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 26: Continuous Attribute Modifications.** Unmodified images are marked in orange. All samples are generated using a linear delta scale from -2 to 2, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 27: Continuous Attribute Modifications.** Unmodified images are marked in orange. All samples are generated using a linear delta scale from -2 to 2, with the deltas being applied after 10/50 steps (Delayed Sampling).



**Fig. 28: Continuous Attribute Modifications.** Unmodified images are marked in orange. All samples are generated using a linear delta scale from -2 to 2, with the deltas being applied for all steps.

## G Implementation Details

This section gives details about the implementation of our method. We generally use the default settings as set in `diffusers`<sup>4</sup>-v0.25.0 with a classifier-free guidance [19] scale of 7.5 and 50-step DDIM [45] sampling unless specified otherwise.

### G.1 Main Method

**Data: Contrastive Prompts** As the basis from which our edit deltas (both the learned ones in Sec. 3.3 and the difference-based ones in Sec. 3.2), we use sets of contrastive prompts and prefixes (the same general setup as in [2, 14]). These are composed of multiple tuples of 3 prompts with a negative, neutral, and positive prompt each (e.g., (“young person”, “person”, “old person”) or (“weak woman”, “woman”, “strong woman”). These tuples generally include both multiple ways of expressing the attribute (e.g., “muscular”, “strong”, “body builder” for the “muscularity” attribute) and, if applicable, multiple ways of naming the target subject (e.g., “person”, “woman”, “man” for deltas targeting people). Additionally, we keep a list of prompt prefixes to achieve more variation. These span some aspects orthogonal to the target attribute (e.g., “a photo of a { $\emptyset$ , mountain, road, BMX, folding, big, small, green, red, blue}” for the “age” attribute for “bike”). These variations of expressing the attribute and prefixes result in a large number of combinations for contrastive prompts, which, in turn, enables a more robust estimation of the underlying direction that modulates this target attribute in the prompt embedding.

**Training (Sec. 3.3)** The tokenwise edit deltas  $\Delta \mathbf{e}_{A_i}$  are implemented as learnable parameters of shape  $1 \times d_{\text{CLIP}}$ , with  $d_{\text{CLIP}}$  being the embedding dimension of the CLIP text encoder. For SDXL [38], this is 2048. This delta is applied additively with scaling according to  $\alpha_i$  to the target subject tokens (e.g., “person” in the case of “a photo of a strong person”) in the original text embedding  $\mathbf{e}$ . If the target subject consists of multiple tokens, we broadcast  $\Delta \mathbf{e}_{A_i}$  across those tokens, although this is only rarely the case in practice.

We train our learned edit deltas  $\Delta \mathbf{e}_{A_i}$  for 1000 steps at a batch size of 10. We use AdamW [29] with a learning rate of 0.1,  $(\beta_1, \beta_2) = (0.5, 0.8)$ , and weight decay of 0.333. All learned deltas are trained on a single A100 with 40GB of VRAM using a bfloat16 version of SDXL [38].

For every entry in the batch, we use a random combination of prefix prompt and prompt tuple and sample an image with the neutral prompt and a random seed, stopping at a random timestep. We then compute the “vanilla” prediction starting from that step for all three prompts, resulting in  $\hat{\mathbf{x}}_{0,a}, \hat{\mathbf{x}}_{0,+}, \hat{\mathbf{x}}_{0,-}$ . In contrast to Gandikota et al. [14], who use a similar approach, we sample our starting samples using standard sampling instead of a modified generation process.

We then sample four values for  $\alpha_i \sim \mathcal{U}([-5, 5] \setminus (-0.1, 0.1))$  and compute  $\mathcal{L}_{\text{delta}}$  (Eq. (5)) using them. We found that sampling multiple values for  $\alpha_i$  here

<sup>4</sup> <https://github.com/huggingface/diffusers>

boosts the performance of our learned deltas at little overhead cost (as the online sampling of the original images is the most costly part) and that values for  $\alpha_i$  very close to zero were not particularly useful for the training process. Empirically, we find that most of our learned edit deltas are already close to convergence after 5 optimization steps, but we keep training for the full time for simplicity.

**CLIP Embedding Differences (Sec. 3.2)** For the edit deltas determined via CLIP embedding differences, we apply the same idea as for the trained version (Sec. G.1), but simply compute the difference between the subject token embeddings as in Eq. (2) and average them over the set of prompts.

**Inference** During inference, we add the learned deltas  $\Delta\mathbf{e}_{A_i}$  to the subject target tokens. The methodology here is the same as described in Sec. G.1.

*Composition* When combining multiple deltas  $\Delta\mathbf{e}_{A_i}$  during a single generation, we add each delta with the chosen scale to the chosen subject. Multiple deltas that are to be applied to the same subject are simply summed, making the results invariant to the order in which they are applied.

*Sampling* We use three sampling variations in Sec. 3.4. For the “Normal Sampling” version, we simply apply  $\Delta\mathbf{e}_{A_i}$  during the whole inference process. For “Delayed Sampling”, we do not apply it for the first few steps (e.g., 10 for a 50-step sampling process) and then apply it, as done in [14]. For our combination with Prompt-to-Prompt [17], we use a public reference implementation<sup>5</sup> and use the word replacement methodology to replace the original subject prompt embedding with our modified subject prompt embedding (original plus  $\alpha_i\Delta\mathbf{e}_{A_i}$ ).

## G.2 Text/Image Pair Deltas

For learning full image deltas on a text/image pair, we use the same optimizer setup as in Sec. G.1 with AdamW [29] with a learning rate of 0.1,  $(\beta_1, \beta_2) = (0.5, 0.8)$ , and weight decay of 0.333. We learn deltas of shape  $N \times d_{\text{CLIP}}$ , with  $d_{\text{CLIP}}$  being the embedding dimension of the CLIP text encoder and  $N$  being the number of tokens in the tokenwise prompt embedding. We do not optimize the start-of-sequence and end-of-sequence tokens or the pooled embeddings in the case of SDXL. We train for 75 steps at a batch size of 1 and randomly select the noise level at each step.

## G.3 Evaluation

To compute perceptual image differences, we use LPIPS [50] as implemented in the `lpips`<sup>6</sup> package with default settings at a resolution of  $256^2$  (interpolated bilinearly). For CLIP scores, we use the standard implementation in `torchmetrics`<sup>7</sup>

<sup>5</sup> <https://github.com/RoyiRa/prompt-to-prompt-with-sdxl>

<sup>6</sup> <https://github.com/richzhang/PerceptualSimilarity>

<sup>7</sup> <https://github.com/Lightning-AI/torchmetrics>



(which outputs cosine similarities scaled to  $[0, 100]$ ) with default settings, including the default CLIP choice of the CLIP-ViT-L/14 trained by OpenAI [39]. For image-image similarity evaluations with DINOv2 [34], we use the ViT-L/14 variant with registers [9] and bi-linearly resize to  $224^2$  before passing them to the model and comparing the cosine similarity of the CLS token outputs. Finally, for ReID evaluations, we use the ArcFace [10] implementation provided by the `insightface`<sup>8</sup> python package with the default `buffalo_1` model, where we return the cosine similarity of the embeddings of the detected faces.

## H Image Copyright

Fig. 2 uses two photos obtained from Unsplash, one of a blue car in the garage<sup>9</sup> by Martin Katler and one of a woman with a bird on her hand<sup>10</sup> (which is also used in Fig. 3) by Ali Esfehaniyan. Both are licensed under the Unsplash license<sup>11</sup>. All other images shown in the paper are generated using Stable Diffusion XL [38] unless noted otherwise.

---

<sup>8</sup> <https://github.com/deepinsight/insightface>

<sup>9</sup> <https://unsplash.com/photos/a-blue-car-parked-in-a-parking-garage-Roq9jZmPetA>

<sup>10</sup> <https://unsplash.com/photos/a-woman-with-a-bird-on-her-shoulder-vpIvmZBurgQ>

<sup>11</sup> <https://unsplash.com/license>