SAEDIT: TOKEN-LEVEL CONTROL FOR CONTINUOUS IMAGE EDITING VIA SPARSE AUTOENCODER

Anonymous authors
Paper under double-blind review

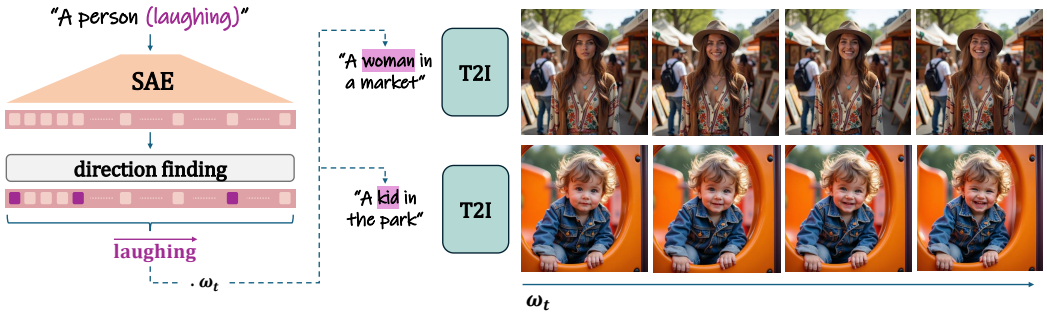

Figure 1: We train a Sparse AutoEncoder (SAE) to lift the text embeddings into a higher-dimensional space, where we identify disentangled semantic directions (e.g. for laughing). These directions can then be applied to specific tokens within the input of a text-to-image model to facilitate continuous image editing. As shown on the right, our token-level editing steers the model to incorporate the relevant attribute (laughing) into the subject in the image that corresponds to the chosen token (e.g., "woman" or "kid"), while allowing the attribute's intensity to be continuously adjusted through a scale factor, ω_t.

ABSTRACT

Large-scale text-to-image diffusion models have become the backbone of modern image editing, yet text prompts alone do not offer adequate control over the editing process. Two properties are especially desirable: disentanglement, where changing one attribute does not unintentionally alter others, and continuous control, where the strength of an edit can be smoothly adjusted. We introduce a method for disentangled and continuous editing through token-level manipulation of text embeddings. The edits are applied by manipulating the embeddings along carefully chosen directions, which control the strength of the target attribute. To identify such directions, we employ a Sparse Autoencoder (SAE), whose sparse latent space exposes semantically isolated dimensions. Our method operates directly on text embeddings without modifying the diffusion process, making it model agnostic and broadly applicable to various image synthesis backbones. Experiments show that it enables intuitive and efficient manipulations with continuous control across diverse attributes and domains.

1 INTRODUCTION

Large-scale text-to-image diffusion models have revolutionized the field of image synthesis (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Consequently, they have become a powerful foundation for a wide array of image manipulation and editing methods (Meng et al., 2022; Hertz et al., 2022; Tumanyan et al., 2023; Cao et al., 2023). These methods have demonstrated remarkable success in a range of edits, including adding new elements, replacing parts of the scene, and modifying the attributes of existing objects. Two properties are particularly desirable in such edits: disentanglement, which ensures that modifying one attribute does not unintentionally affect others, and continuous control, which allows adjusting the magnitude of the edit.

While there has been significant progress in achieving disentangled editing, finding controllable representations that enable edits which are both disentangled and continuous remains a major challenge. Text prompts alone struggle to provide this level of control, as their discrete nature prevents

Figure 2: Naïvely applying T5 edit direction (top) by interpolating T5 embedding of target edit, introduces entangled changes that may distort the scene. This can appear as an insufficient edit (left example) or as the modification of unwanted elements (right example). In contrast, edit directions found by the SAE (bottom) yield disentangled edits that preserve identity and achieve the intended modification.

smooth intensity adjustment and their holistic influence often leads to unintended changes. For example, to control the intensity of a smile, a user must resort to distinct coarse categorical descriptions like "a slight smile" versus "a wide grin", rather than smoothly varying the intensity. This limitation motivates research into underlying semantic control mechanisms that are both continuous and disentangled.

In pursuing this goal, some works have focused on general, training-free methods that manipulate the diffusion model's internal representations (Dalva et al., 2024; Guerrero-Viu et al., 2024; Baumann et al., 2025). While versatile, these techniques often struggle with disentanglement, where an edit intended to be local inadvertently causes widespread, undesirable changes to the overall image style and composition. To achieve higher fidelity, other approaches have pursued task-specific optimization, training a dedicated module for each edit (Gandikota et al., 2023; Sridhar & Vasconcelos, 2024; Dravid et al., 2024), with the module's weights acting as the controllable representation for the edit. However, while often producing high-quality results, this strategy is inherently unscalable, demanding a unique training pipeline for every possible modification.

In this work, we propose a method for disentangled and continuous image editing through the fine-grained manipulation of text embeddings at the token-level. Our approach leverages a Sparse Autoencoder (SAE) (Cunningham et al., 2023), an unsupervised model trained to reconstruct its input from a sparse, high-dimensional latent space. The sparsity of this latent space induces semantically disentangled dimensions, which in turn enable the discovery of meaningful editing directions for each token.

Specifically, we derive an edit direction in the SAE's space by comparing the sparse representations of two prompts that differ by the desired edit description (e.g., "a person" and "a smiling person"), identifying the entries most correlated with the change. We then construct an edit-specific direction as a sparse vector that modifies only these highly relevant entries.

This disentangled direction is added to the sparse representation of the prompt, and can be scaled to continuously control the magnitude of the target attribute, while preserving the rest of the image. This approach leverages the SAE to uncover disentangled directions that are difficult to identify directly in the raw embedding space, as qualitatively demonstrated in Figure 2.

Our method operates solely on the text embeddings, leaving the denoising process untouched. In this setup, the diffusion model serves merely as a renderer: it receives the edited semantic instructions and translates them into a visual output. As a result, the method is model-agnostic and can be applied to any text-to-image backbone that shares the same text encoder, without additional training or fine-tuning.

Through extensive experiments, we demonstrate the effectiveness of our method in providing both continuous and highly disentangled semantic edits. We validate the versatility of our approach by ap-

plying the same framework to various generative models, including two image synthesis backbones, without any model-specific training. Importantly, we show that our method enables a wide range of intuitive, magnitude-controlled manipulations from simple text commands, as demonstrated in Figure 1. We further show that our method can be applied to real images using inversion techniques.

2 Related Work

Image Editing with Diffusion Models The success of diffusion models in image synthesis (Ho et al., 2020; Song et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Song et al., 2021; Black Forest Labs, 2024; Podell et al., 2023) has led to their widespread adoption for the more challenging task of real image editing. Unlike pure generation, editing requires a careful balance between preserving an image's original attributes and introducing controlled, text-guided changes. Common strategies include manipulating the denoising process through feature injection (Hertz et al., 2022; Parmar et al., 2023; Tumanyan et al., 2023; Cao et al., 2023; Alaluf et al., 2023; Patashnik et al., 2023) or applying partial noise schedules with a new text condition (Meng et al., 2022; Huberman-Spiegelglas et al., 2023; Tsaban & Passos, 2023; Brack et al., 2023b; Deutch et al., 2024; Rout et al., 2024). A key requirement for applying these methods to real images is an inversion technique that can find an initial noise capable of reconstructing the image (Dhariwal & Nichol, 2021; Mokady et al., 2022; Miyake et al., 2023; Han et al., 2024; Garibi et al., 2024; Samuel et al., 2024; Jiao et al., 2025; Kadosh et al., 2025).

Continuous Image Editing with Diffusion Models A challenge in this area is achieving fine-grained, continuous control over semantic attributes. To achieve this kind of control some methods perform Task-specific optimization methods, which yield high-fidelity, disentangled edits but are not scalable, requiring a separate, costly process for each new attribute, such as training a dedicated LoRA adapter (Gandikota et al., 2023), optimizing a text token (Sridhar & Vasconcelos, 2024) or to train numerous person-specific DreamBooth LoRAs (Ruiz et al., 2023) and then trains a classifier in the latent space (Chang et al., 2025) or in the weights' space (Dravid et al., 2024). Conversely, training-free methods that discover semantic directions in existing latent spaces (Dalva et al., 2024; Guerrero-Viu et al., 2024; Baumann et al., 2025; Garibi et al., 2025; Dorfman et al., 2025; Brack et al., 2023a) are general-purpose but often struggle with the precision and disentanglement of specialized models. Other works (Gandikota et al., 2025; Dalva & Yanardag, 2023) explore unsupervised discovery of a model's latent variations but are not designed for direct, text-guided editing. Our work aims to bridge this gap, offering a general framework that provides the disentangled control of task-specific methods without the need for per-edit training.

3 Preliminary - Sparse AutoEncoders

Sparse Autoencoders (SAEs) are neural architectures designed to learn interpretable and disentangled high-dimensional latent representations (Cunningham et al., 2023). An SAE typically consists of a simple encoder, often a single linear layer with a non-negative activation, and a linear decoder. The model is trained with a dual objective:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \cdot \mathcal{L}_{\text{sparse}}, \tag{1}$$

where $\mathcal{L}_{\text{rec}}$ is a standard reconstruction loss, and $\mathcal{L}_{\text{sparse}}$ is a set of regularization terms that encourages the latent representation to be sparse. This sparsity constraint encourages the SAE to learn a dictionary-like representation, where a small set of active latent features often corresponds to a distinct semantic attribute of the input. This property makes SAEs a powerful tool for interpreting the otherwise dense and opaque hidden states of large language models.

Consequently, SAEs have been successfully applied to the internal states of large language models to uncover meaningful, semantic features (Bricken et al., 2023; Gao et al., 2024; Cunningham et al., 2023). For example, Bricken et al. (2023) found that certain features in the sparse representation are active only when specific entities, such as "US presidents," are mentioned in the text. Identifying which features correspond to specific concepts enables model steering, allowing for direct control over model behavior by manipulating its internal activations (Arad et al., 2025; Bayat et al., 2025). The basic SAE framework can be extended with more advanced variants and sparsity regularization techniques, which are detailed further in Section C.

Recent works have integrated Sparse Autoencoders (SAEs) into diffusion models for various distinct purposes. For instance, *One-Step is Enough* (Surkov et al., 2025) trains SAEs on SDXL cross-

attention layers to extract interpretable features from the model's internal representations. Furthermore, *Concept-Steerers* (Kim & Ghadiyaram, 2025) and *SAEuron* (Cywiński & Deja, 2025) utilize SAEs to steer generation toward safer outputs and enable concept unlearning, respectively. However, these methods do not support controllable, continues image editing.

4 METHOD

We present a method for text-driven image editing that provides both disentanglement and continuous control. Our approach is based on manipulating the text embeddings of a frozen text-to-image model. We train a Sparse Autoencoder (SAE) on these embeddings, which provides a space in which disentangled directions corresponding to semantic attributes can be found. Editing is then performed by adjusting the embeddings along these directions to achieve controlled manipulations.

Specifically, given a frozen text encoder, we train a Sparse Autoencoder (SAE) on its output embedding space (details in Sec. 4.1). The SAE is composed of an encoder, $\mathcal{S}_{enc}$, and a decoder $\mathcal{S}_{dec}$. The encoder maps dense text embeddings into a high-dimensional, disentangled latent space where distinct semantic concepts are isolated, while the decoder reconstructs the original embedding from this sparse representation. Once trained, manipulations are applied directly in this sparse SAE's space by adjusting specific entries in the latent representation. The modified representation is then passed through the SAE's decoder to recover an edited text embedding, which can be fed into any compatible text-to-image model (e.g., Flux) that uses the same text encoder architecture. In this way, the SAE acts as a lightweight, pluggable module that enables disentangled and semantic control over the final generated image.

The editing direction is obtained from a source prompt $\mathcal{P}_{src}$ (e.g. a "man") and target prompt $\mathcal{P}_{tgt}$ (e.g. "a smiling man") , details in Sec. 4.2. We apply the edit direction by multiplying it with a scale factor and adding it to the sparse representation of the specific source token in $\mathcal{P}_{src}$ to be edited (e.g. the "man" token). The magnitude of the edit is dictated by this scale factor, allowing for continuous control over the attribute's intensity (details in Sec. 4.3).

We demonstrate our method on the T5 text encoder (Raffel et al., 2023), which is widely adopted as the text conditioning module in many state-of-the-art text-to-image models. For the image generation backbone, which acts as a renderer for our text embedding manipulations, we primarily use the Flux (Black Forest Labs, 2024) diffusion transformer (DiT).

4.1 SAE TRAINING

We train our Sparse Autoencoder (SAE) on a dataset of text embeddings. To create this dataset, we first process a corpus of text prompts through the frozen T5 text encoder and collect the resulting token embeddings, excluding padding tokens. Notably, unlike typical SAE applications that focus on intermediate transformer layers, we train our SAE on the final output of the text encoder, as these are the exact representations that are continuously processed by the Diffusion Transformer (DiT) throughout the denoising steps.

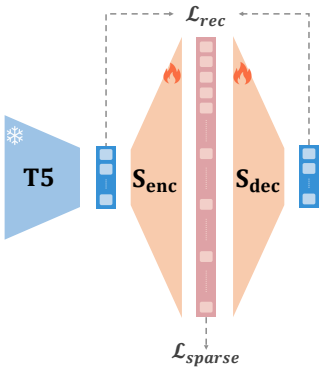

Figure 3: We train the Sparse Autoencoder on token embeddings obtained from a frozen T5 encoder, using reconstruction and sparsity losses.

The SAE is trained on the embeddings of individual text tokens, using the objective function from Eq. 1, the process illustrated in Fig. 3. Here, $\mathcal{L}_{rec}$ is the standard reconstruction loss (e.g., Mean Squared Error) between the SAE's input and output embeddings. We control the target level of sparsity via another hyperparameter which sets the desired number of non-zero activations for each token's latent code.

4.2 OBTAINING AN EDIT DIRECTION

Motivated by prior work on SAEs in language models, which shows that specific entries in the sparse representation activate only in the presence of particular semantic attributes (Cunningham et al., 2023; Gao et al., 2024), we aim to detect such entries to construct disentangled directions in the SAE's latent space for image editing. To do so, we use a source prompt $\mathcal{P}_{src}$ (e.g. "a woman") and a target prompt $\mathcal{P}_{tgt}$ (e.g. "a woman laughing"). We first encode all text tokens in both prompts using the SAE encoder, $\mathcal{S}_{enc}$, to obtain sparse token representations. Since it is unknown apriori

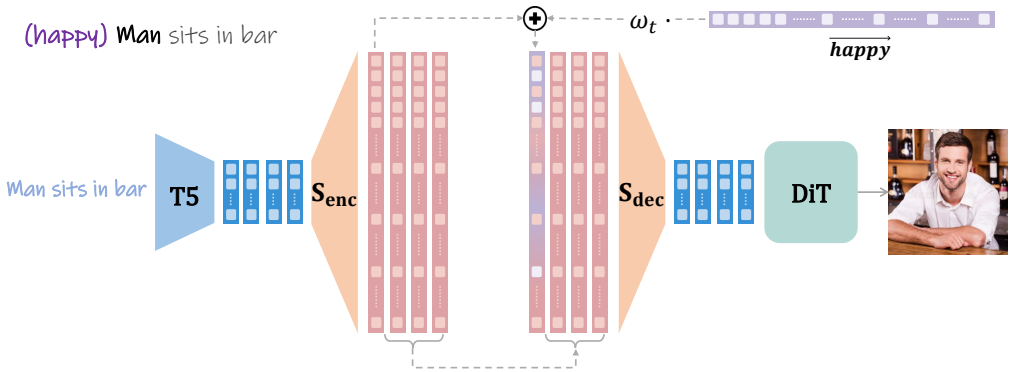

Figure 5: Applying the edit direction. An aggregated edit direction is scaled to adjust edit magnitude and applied to the sparse representation of the relevant source token (e.g., man). The result is then decoded back into the T5 embedding space, and used to condition the text-to-image model.

which tokens hold the semantic information for a concept (Kaplan et al., 2025), we use element-wise max-pooling to aggregate their sparse representations into a single, sparse vector for each prompt. As $\mathcal{P}_{src}$ and $\mathcal{P}_{tgt}$ are semantically similar except for the edited attribute, the activated entries in maxpool$(S_{enc}(\mathcal{P}_{src}))$ and maxpool$(S_{enc}(\mathcal{P}_{tgt}))$ should overlap substantially, with their differences centering around entries corresponding to the edit-specific attribute.

To identify the entries that correlate with the requested edit, we compute an entry-wise ratio, R, between the source and target prompt:

$$R = \frac{\text{maxpool}(S_{enc}(\mathcal{P}_{tgt}))}{\text{maxpool}(S_{enc}(\mathcal{P}_{src})) + \epsilon}, \tag{2}$$

where ϵ is a small constant added for numerical stability. The entries in R with the highest values correspond most strongly to the edit-specific attribute. Next, to isolate these key entries , we normalize the ratio vector, $R^{norm} = R/\max(R)$, and apply a predefined threshold $\rho \in [0, 1]$. This yields a set of indices, M, corresponding to the most relevant entries for the edit:

$$M = \{i \mid R_i^{norm} > \rho\}. \tag{3}$$

Finally, we use this set of indices to construct the disentangled edit direction, d_{edit}, as a sparse vector, as illustrated in Fig. 4. The direction is defined to be zero everywhere except at the identified indices, where it takes its values from the target representation:

$$[d_{edit}]_i = \begin{cases} [S_{enc}(\mathcal{P}_{tgt})]_i & \text{if } i \in M, \\ 0 & \text{if } i \notin M \end{cases} \tag{4}$$

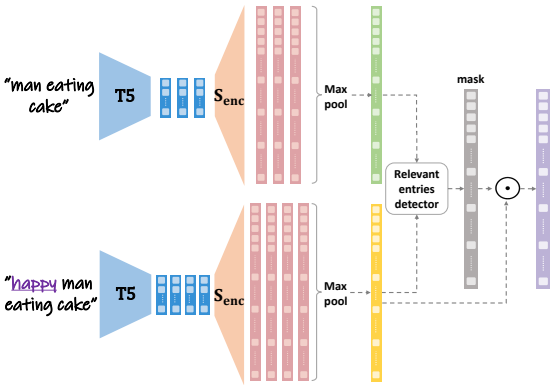

Figure 4: Extracting Edit Directions. We derive an edit direction from a prompt pair that isolates a single attribute. Both prompts are encoded with the SAE, and their token representations are aggregated via max-pooling. By comparing the two resulting sparse vectors, we identify the key features corresponding to the desired change. The final edit direction is a sparse vector composed of only these key features, taken from the target prompt's representation.

Improving direction's robustness To enhance the robustness of our derived edit directions, we aggregate information from a set of multiple source-target prompt pairs rather than relying on a single pair. Given a desired edit, defined by the pair of texts descriptions $\mathcal{P}_{src}$ and $\mathcal{P}_{tgt}$, we use an LLM to construct N sentence pairs that share the same underlying semantic relationship. This process, generalizes the specific edit into an abstract concept. For example, to create a direction for "happiness", the LLM generates pairs that add this attribute to various contexts, such as ("man on the beach", "happy man on the beach") and ("man eating cake", "happy man eating cake"). We then apply our direction-finding procedure to each of the N prompt pairs, resulting in a set of N steering vectors $\{d_i\}_{i=1}^N$. These vectors are stacked to form a direction matrix: $D = [d_1, \ldots, d_N]^T$. To extract the most prominent and consistent direction representing the shared attribute across all examples, we perform Singular Value Decomposition (SVD) on D. The singular vector corresponding to the largest singular value is then selected as our final, robust edit direction d_{edit}.

Figure 6: Qualitative Results. Our method enables a diverse range of continuous and disentangled semantic edits across various image styles. We demonstrate the ability to add attributes (e.g., mustache, glasses), change expressions (smile, laugh), and perform highly localized edits, such as modifying the age of only one person in a scene.

4.3 Applying the Edit Direction

Once the edit direction d_{edit} is derived, we apply it to the source prompt $\mathcal{P}_{src}$. To ensure the manipulation is localized, we modify only the embedding of the specific token to be edited (e.g., the "woman" token), which we denote as e_{tgt}. The magnitude of the edit is controlled by a scalar factor ω, allowing for continuous, fine-grained control over the attribute's intensity.

The final, edited text embedding for the token, e'_{tgt}, is produced by first encoding the original token's embedding with $\mathcal{S}_{enc}$, adding the scaled direction in the sparse latent space, and then decoding the result with $\mathcal{S}_{dec}$:

$$e'_{token} = \mathcal{S}_{dec}(\mathcal{S}_{enc}(e_{tgt}) + \omega \cdot d_{edit}). \tag{5}$$

Setting $\omega = 0$ recovers the original embedding, while progressively increasing ω strengthens the visual effect. This new token embedding, e'_{tgt}, replaces the original in the prompt.

Finally, the manipulated text embeddings are used to condition the renderer. Specifically, for diffusion models, we follow the standard editing approach to preserve the overall structure of the source image. This involves using the same initial noise, x_T, that was used to generate the source image, and only substituting the original token embeddings with our modified ones. This ensures that the changes in the final generated image are driven exclusively by our disentangled edit. Fig. 5 provides a schematic of this entire editing pipeline.

4.4 Injection Schedule

The denoising process in diffusion models operates hierarchically: early timesteps are crucial for establishing the global structure and layout of an image, while later steps refine fine-grained details and textures (Patashnik et al., 2023; Balaji et al., 2023; Cao et al., 2025; Huberman et al., 2025; Yehezkel et al., 2025). Consequently, for fine-grained edits that aim to preserve the original structure, prior work has shown that it is often optimal to begin the editing manipulation only at later timesteps, after the core layout is formed (Huberman-Spiegelglas et al., 2023; Jiao et al., 2025).

Building on this insight, we introduce an exponential injection schedule that applies the edit direction with increasing intensity over time. For a base scale factor ω and diffusion step t, we define the time-dependent scale ω_t as:

$$\omega_t = \min\left(e^{t\cdot\omega} - 1, \tau\right), \tag{6}$$

where $\tau \in \mathbb{R}$ is a hyperparameter that acts as an upper bound on the edit strength. This exponential formulation offers a key advantage over linear schedules: it applies the edit very gently in the early, structure-defining timesteps and progressively increases its influence as the process moves into the later, detail-refining stages. This gradual application better aligns with the hierarchical nature of image synthesis, preserving global structure while enabling powerful, fine-grained modifications.

5 Experiments

We conduct extensive experiments to evaluate our method's ability to provide continuous control and disentangled edits that preserve the subject's identity. Similar to prior work, we focus our evaluation

Figure 7: Results with SD3.5. These demonstrate that our method integrates seamlessly with models relying on T5, enabling consistent and faithful edits across architectures.

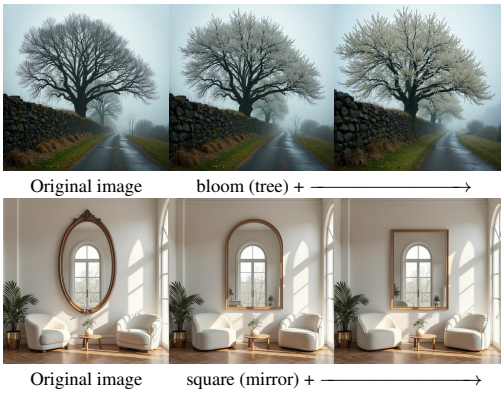

Figure 8: Our method's versatility extends beyond human subjects, enabling continuous and disentangled control over object attributes like seasonal appearance and object shape.

on human subjects, a challenging domain that demands strong disentanglement to preserve identity and offers the most meaningful application of continuous magnitude control. To demonstrate its model-agnostic nature, we apply our approach to both Flux (Black Forest Labs, 2024) and Stable Diffusion 3.5 (Esser et al., 2024). Unless otherwise specified, all results are generated using Flux. We also show its applicability to real image editing through integration with standard inversion techniques. For quantitative evaluation, we measure preservation with LPIPS (Zhang et al., 2018) and semantic accuracy with a VQA-Score (Lin et al., 2024). Implementation details in the Appendix A.

5.1 Qualitative Results

We present qualitative results generated by our method, SAEdit, in Figures 1, 6, 7 and 8. Figure 6 shows a wide variety of continuous edits on human subjects. Our method successfully changes expressions (e.g., adding a smile), modify attributes (e.g., making hair blonde), and add accessories (e.g., hats or glasses). Crucially, these edits are highly localized. For instance, we demonstrate the ability to modify the age of a single person in a multi-subject image while leaving the other person and the background entirely untouched.

The results also highlight the continuous nature of our control. As shown in the examples, attributes such as the intensity of a laugh or the degree of age can be smoothly scaled. This allows users to precisely tune the magnitude of the desired effect while the rest of the image content is faithfully preserved.

The approach is not limited to human subjects and generalizes to a broad range of semantic concepts, as shown in Figure 8. Finally, to demonstrate the model-agnostic nature of SAEdit, Figure 7 shows that the same edit directions produce consistent, high-quality results when applied to a different T5-based model, Stable Diffusion 3.5. We provide additional qualitative results in Appendix B.1, including examples of our method applied to real images in Appendix B.2 and used for gradual change in style in Figure 26.

5.2 Ablation

Figure 9 provides a qualitative ablation study of our method's components, demonstrating their respective contributions to the final result. As a baseline, deriving an edit direction from a single prompt pair (top row) preserves the subject's identity, but the intended semantic change to the expression is weak and insufficient. Aggregating the direction from N prompt pairs (middle row)

Figure 9: Ablation study. We demonstrate how each component progressively improves the quality of an 'angry' edit. A direction from a single prompt pair results in a weak edit with unintended modifications. Aggregating N prompts produces a more robust and semantically accurate direction, but can still alter fine details. Adding our exponential injection schedule preserves the original image's details (e.g., the necklace and hair color), yielding the most faithful and disentangled result.

successfully strengthens the edit as required, but causing minor unwanted changes to the hair color, the necklace, and the dress texture. Finally, incorporating our exponential injection schedule (bottom row) resolves this issue by preserving these fine-grained details while maintaining the strong semantic edit, thus achieving a high-quality and disentangled result. Our quantitative ablation study is detailed in Appendix B.4.

5.3 COMPARISONS

We evaluate our method against several state-of-the-art approaches for continuous image editing, highlighting its ability to provide disentangled control without per-edit optimization. We compare against methods from both optimization-based and training-free categories. From the optimization-based group, we evaluate Concept Sliders (Gandikota et al., 2023) by using their official SDXL-trained LoRAs as well as LoRAs we trained on the Flux architecture for a direct comparison. We additionally include Prompt Sliders (Sridhar & Vasconcelos, 2024), a textual-inversion (Gal et al., 2022) based method that learns concept embeddings used for adjustable control. In the training-free category, we compare against FluxSpace (Dalva et al., 2024), adjusting its λ_{fine} parameter to control edit magnitude. We also evaluate against Attribute Control (Baumann et al., 2025), a method that proposes both a training-free and an optimization-based variant. For Flux Kontext (Labs et al., 2025), which does not natively support continuous edit scaling, we implement two proxy baselines for magnitude control over the edit strength: the first involves varying the Classifier-Free Guidance (CFG) strength, while the second uses an LLM to generate prompts corresponding to 'light' and 'extreme' versions of each edit (instruction prompts and more details provided in Appendix B.5).

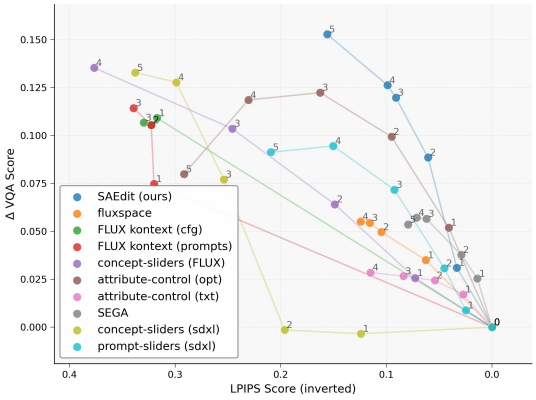

Figure 10: Quantitative comparison. We compare our method to other baselines on image preservation and prompt fidelity (top-right is better).

Quantitative Comparisons To evaluate the fine-grained and continuous control of our method, we constructed a custom evaluation set. This set is based on 63 images, each generated from a unique prompt created by a large language model (OpenAI, 2025). Each prompt describes a scene containing a person. For each source image, we applied a set of 6 to 8 different semantic edit directions, resulting in 432 unique edit scenarios. To assess the continuity of these edits, we then generated each scenario at 3-5 distinct magnitude levels, producing a final evaluation set of at least 1,296 images per method. The complete list of prompts and edit directions is provided in Appendix B.3. We quantitatively assess our method along three axes: image preservation, prompt adherence and identity preservation. To measure the image preservation of original content, we use LPIPS (Zhang et al., 2018). To measure prompt adherence with the edit, we compute a VQA-based score (Lin et al., 2024). This score is the delta between the VQA score of the edited image against the target prompt and that of the source image against the same prompt, which isolates the semantic change introduced by the edit. Finally, to evaluate identity preservation, we employ ArcFace (Deng et al., 2022) to calculate the cosine similarity between the subject in the source image and the edited output.

Figure 10 and 20 presents the quantitative comparison between our method and other methods at varying levels of edit intensity. The results demonstrate that our method outperforms all other approaches.

Notably, our zero-shot method is superior even to task-specific techniques that are explicitly trained for each edit type. This indicates that our approach successfully achieves the dual goals of high semantic accuracy for the required edit and strong preservation of the original content. Furthermore, the metrics show a smooth and predictable progression as the edit magnitude increases, confirming that our method provides true continuous control and allows users to precisely tune the intensity of an effect.

Table 1: User Study. Pairwise win rate of our method against other methods.

Opponent Method	Image Pres.	Prompt Adher.	**Overall**
Flux Kontext (CFG)	73%	71%	70%
Flux Kontext (LLM)	60%	68%	70%
ConceptSlider (Flux)	71%	67%	71%
Flux Space	59%	92%	93%

User Study To complement our quantitative analysis, we conducted a user study to evaluate the perceptual quality of our method against competing approaches. For fairness, we limited our comparison to methods that also use the Flux model, ensuring the source images were as similar as possible. In a pairwise comparison, we presented participants with results from our method and a competing method, showing three distinct levels of edit intensity for each to assess continuous control. Users were asked to state their preference based on three criteria: Image Preservation, Prompt Alignment (which included the gradualness of the effect), and Overall Quality. In total, our user study gathered 390 pairwise comparison responses. More details in Appendix B.7

The results, summarized in Table 1, show that our method was significantly preferred over all other approaches in all categories. This suggests that users found our edits achieve a better balance of successfully applying the desired change while faithfully preserving the original image content.

Qualitative Comparisons Figure 12 presents a qualitative comparison between our method and other approaches. While the results for most methods are taken directly from our quantitative evaluation set, we manually optimized the prompts for the Flux Kontext baselines to ensure the strongest possible comparison, as their default outputs were often suboptimal (see Appendix B.6). For example, for the CFG-based baseline, we found the prompt "Make the man look slightly like a kid" with CFG scales of 1.5 and 1.6 yielded the most plausible results. While most of the compared methods operate on the Flux model, Attribute Control and Prompt Sliders use SDXL-based pipelines. For Prompt-Sliders we implemented an inversion procedure to enable real-image editing within their framework, and for Attribute-Control we use their official real-image editing pipeline based on SDXL Turbo. In both cases, we present the reconstructed inversion output as the "original image" for a fair comparison across methods. Additional qualitative comparisons with Attribute Control and Prompt Sliders are provided in Figure 25. In this configuration, inversion is applied only by our method, while the SDXL-based baselines generate their results directly. The visual comparison highlights the superior disentanglement of our method. For instance, in contrast to Concept-Sliders, our approach achieves a perfect reconstruction of the subject's jacket while applying the desired edit. Similarly, when compared to Flux Kontext, our method successfully modifies the subject's age in a more natural and gradual manner, demonstrating more precise control over the semantic attributes. More results in Appendix B.6.

5.4 Integration With Flux-Kontext

Figure 11: Our method struggles with out-of-distribution (OOD) edits that conflict with strong priors in the base model. For example, applying a "beard" edit changes the woman into a man (left), while making the dog "green" results in an unnatural, animated-style dog (right).

Since our method operates directly on the semantic text-embedding space, it can be seamlessly incorporated into instruction-based models. We demonstrate this capability using Flux-Kontext for gradual real-image editing (Figure 24). The application mechanism, however, differs from the standard text-to-image setting. The direction finding mechanism remains the same, but the target token changes. In the text-to-image case the edit direction is injected into the embedding of the target subject token, whereas instruction-based models require modifying the instruction tokens that encode the editing operation itself.

To enable continuous control in this setting, we apply the edit direction to the tokens corresponding to the action described by the instruction. For example, for an instruction such as "make the man laugh", we subtract the calculated laugh direction from the associated instruction token ("laugh").

This reduces the strength of the directive, providing fine-grained and gradual control over the resulting edit intensity.

5.5 LIMITATIONS

While our method identifies robust and disentangled edit directions, we observe that further refinement is sometimes possible. For certain complex edits, manually selecting or de-selecting a few specific entries in the sparse direction vector can yield even more disentangled results.

In addition, our method's ability to disentangle is constrained by the inherent biases of the underlying text-to-image model. When an edit is requested that is strongly out-of-distribution (OOD), our approach can fail to maintain disentanglement. As shown in Figure 11, attempting to add a 'beard' to a "woman" results in the subject's perceived gender being changed to male. Similarly, making a dog "green" alters its texture to appear unnatural and cartoon-like. We hypothesize these failures occur because the SAE cannot fully separate concepts that are fundamentally entangled in the base model's worldview.

Figure 12: Each row showcases the results of a different editing method for the same edit. Our method (bottom row) produces a more disentangled result that better preserves the subject's identity compared to the competing approaches.

6 CONCLUSIONS

In this work, we introduced a novel framework that provides both disentangled and continuous control for text-to-image editing. Our method leverages a Sparse Autoencoder (SAE) on text embeddings to create a sparse representation where semantic attributes are isolated. This sparse representation is the key to our method's success. Having isolated individual attributes facilitates disentangled edits, where the subject's core identity is preserved. Our approach enables token-level manipulation, providing fine-grained and continuous control over the magnitude of a given attribute.

A key advantage of our design is that editing is decoupled from rendering: we modify only the text embedding, enabling any compatible text-to-image backbone model to act as the renderer. SAEs are primarily known for their role in interpretability of language models, yet in this work we demonstrate that they can be harnessed for image generation, yielding fine-grained editing capabilities. Image editing has recently seen remarkable progress, yet precise fine-grained control remains an open challenge, and we believe this work will encourage further advances in that direction.

Ethics Statement

We acknowledge the ethical considerations inherent in powerful text-to-image models. Our method, like other generative technologies, is a dual-use tool that can be used for both beneficial creative purposes and potential misuse. The underlying text and image models we build upon are trained on large-scale internet data and are known to contain societal biases, which can lead to the generation of stereotypical or harmful content. Furthermore, the ability to manipulate images carries the risk of being used for creating misinformation or other malicious synthetic media.

We have developed this work with the goal of empowering creative applications. We strongly advocate that any deployment of this technology must be accompanied by robust safety filters and content moderation systems to mitigate the risks of generating harmful or biased outputs. We are committed to the principles of responsible AI development and encourage continued research into the safety, fairness, and transparency of generative models.

Reproducibility Statement

To ensure the reproducibility of our results, we will make our code, pre-trained SAE models, and the derived edit directions publicly available upon publication. Our method is built upon the publicly available T5-XXL text encoder, and our experiments use the official model weights for Flux.dev and Stable Diffusion 3.5. Key details for training the Sparse Autoencoder, including all hyperparameters such as the learning rate, sparsity coefficient α, and the target number of active features, are provided in Appendix A. The edit directions were derived using 100 prompt pairs generated by GPT for each concept. The complete set of prompts and edits used in our custom evaluation benchmark is included in the supplementary materials to facilitate direct and fair comparisons.

References

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering – if you select the right features, 2025. URL https://arxiv.org/abs/2505.20063.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. URL https://arxiv.org/abs/2211.01324.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions, 2025. URL https://arxiv.org/abs/2403.17064.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces, 2025. URL https://arxiv.org/abs/2503.00177.

Black Forest Labs. Flux, https://github.com/black-forest-labs/flux, 2024. URL https://github.com/black-forest-labs/flux.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance, 2023a. URL https://arxiv.org/abs/2301.12247.

Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. *arXiv preprint arXiv:2311.16711*, 2023b.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex

Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders, 2025. URL https://arxiv.org/abs/2503.17547.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023.

Yu Cao, Zengqun Zhao, Ioannis Patras, and Shaogang Gong. Temporal score analysis for understanding and correcting diffusion artifacts, 2025. URL https://arxiv.org/abs/2503.16218.

Yuanyuan Chang, Yinghua Yao, Tao Qin, Mengmeng Wang, Ivor Tsang, and Guang Dai. Instructing text-to-image diffusion models via classifier-guided semantic optimization, 2025. URL https://arxiv.org/abs/2505.14254.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.18052.

Dawei Dai, Xu Long, Li Yutang, Zhang Yuanhui, and Shuyin Xia. Humanvlm: Foundation for human-scene vision-language model, 2024. URL https://arxiv.org/abs/2411.03034.

Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models, 2023. URL https://arxiv.org/abs/2312.05390.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow transformers, 2024.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3087709. URL http://dx.doi.org/10.1109/TPAMI.2021.3087709.

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models, 2024. URL https://arxiv.org/abs/2408.00735.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts, 2025. URL https://arxiv.org/abs/2502.13951.

Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A. Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models, 2024. URL https://arxiv.org/abs/2406.09413.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

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, 2022. URL https://arxiv.org/abs/2208.01618.

Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models, 2023. URL https://arxiv.org/abs/2311.12092.

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. arXiv:2502.01639.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising, 2024. URL https://arxiv.org/abs/2403.14602.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space, 2025. URL https://arxiv.org/abs/2501.12224.

Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutiérrez, Belen Masia, and Valentin Deschaintre. Texsliders: Diffusion-based texture editing in clip space. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH '24, pp. 1–11. ACM, July 2024. doi: 10.1145/3641519.3657444. URL http://dx.doi.org/10.1145/3641519.3657444.

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4291–4301, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Saar Huberman, Or Patashnik, Omer Dahary, Ron Mokady, and Daniel Cohen-Or. Image generation from contextually-contradictory prompts. *arXiv preprint arXiv:2506.01929*, 2025.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations, 2023.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models, 2025. URL https://arxiv.org/abs/2504.13109.

Edo Kadosh, Nir Goren, Or Patashnik, Daniel Garibi, and Daniel Cohen-Or. Tight inversion: Image-conditioned inversion for real image editing, 2025. URL https://arxiv.org/abs/2502.20376.

Guy Kaplan, Michael Toker, Yuval Reif, Yonatan Belinkov, and Roy Schwartz. Follow the flow: On information flow across textual tokens in text-to-image models, 2025. URL https://arxiv.org/abs/2504.01137.

Dahye Kim and Deepti Ghadiyaram. Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations, 2025. URL https://arxiv.org/abs/2501.19066.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models, 2023.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

OpenAI. Chatgpt (gpt-5). https://chat.openai.com/, 2025.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*, SIGGRAPH '23. ACM, July 2023. doi: 10.1145/3588432.3591513. URL http://dx.doi.org/10.1145/3588432.3591513.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations, 2024. URL https://arxiv.org/abs/2410.10792.

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, 2023. URL https://arxiv.org/abs/2208.12242.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models, 2024. URL https://arxiv.org/abs/2312.12540.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Deepak Sridhar and Nuno Vasconcelos. Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models, 2024. URL https://arxiv.org/abs/2409.16535.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models, 2025. URL https://arxiv.org/abs/2410.22366.

Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance, 2023. URL https://arxiv.org/abs/2307.00522.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. pp. 1921–1930, June 2023.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL https://arxiv.org/abs/2210.14896.

Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. Navigating with annealing guidance scale in diffusion space. *arXiv preprint arXiv:2506.24108*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

APPENDIX

A IMPLEMENTATION DETAILS

We illustrate our method with the T5-XXL text encoder (Raffel et al., 2023), which is utilized by state-of-the-art text-to-image models such as Flux.dev (Black Forest Labs, 2024) and Stable Diffusion 3.5 (Esser et al., 2024). To train the SAE, we compiled a dataset from two sources: the DiffusionDB dataset (Wang et al., 2022), containing 2M general image captions, and the HumanCaption-10M dataset (Dai et al., 2024), which provides 10M captions focused on humans. The combined training set consists of 12M text prompts, totaling approximately 800M text tokens after filtering.

The dimension of the SAE's latent space is set to 65,536, and the target number of active entries for each token is 300. We trained the SAE for 200,000 steps using the Adam optimizer with a learning rate of 0.003. The weight for the sparsity loss, α (from Eq. 1), was set to $\frac{1}{32}$.

For each edit, the corresponding direction was derived using a set of $n = 100$ source and target prompt pairs. These prompt pairs were generated using GPT-5. The parameter τ (from Eq. 6) used for the exponential injection mechanism was set to be a function of the scale parameter: $\tau = 15 \cdot \omega$.

B EXPERIMENTS

B.1 ADDITIONAL QUALITATIVE RESULTS

Figure 13 showcases the universality of our learned edit directions. We apply the exact same set of four directions (smile, angry, surprised, and old) to four diverse source images, demonstrating that a single direction vector can generalize effectively across different subjects, scenes, and identities.

Figure 14 demonstrates the compositionality of our learned directions, where we independently control a "smile" on the horizontal axis and the addition of "glasses" on the vertical axis. It is evident that these manipulations are highly disentangled, as the subject's identity and all background details remain perfectly consistent across the grid, with only the intended attributes changing. Figure 28 further demonstrates the ability of our method to compose up to 7 edit directions on a single image.

We further demonstrate the compositionality and advanced localization capabilities of our method in Figure 16. The figure showcases the simultaneous application of two distinct edits targeted at different subjects within the same scene. A "laugh" direction is applied to the woman, while an "old" direction is applied to the man. The results across the grid show that each manipulation is successfully confined to its intended target, preserving the background and the non-targeted attributes of each subject without interference.

Figures 23 and 17 present additional qualitative results for continuous editing on human and non-human subjects, respectively.

Figure 27 shows that our method generalizes across seeds, we apply the same edit direction on the prompt "A close up portrait of a man wearing a black coat and a yellow hat, vivid colors, daylight", and we show our results for random seeds 0-3.

B.2 REAL IMAGE EDITING

Our method's applicability extends to the challenging task of real image editing. To achieve this, we first use a state-of-the-art inversion technique, Uni-Inv Flow (Jiao et al., 2025), to obtain the initial noise corresponding to a given source image. Our SAE-based manipulation is then applied to the text embeddings as previously described. Figure 15 presents several results of this combined approach. As shown, we can apply high-fidelity, continuous edits to real photographs, successfully modifying expressions (cry, laughing) and attributes (old). Importantly, these edits preserve the subject's core identity and background details, demonstrating that our disentangled control is effective even in the demanding context of real image manipulation.

Figure 13: Each row shows a different source image (leftmost column) and its edits along four semantic directions: smile, angry, surprised, and old. The images in each column are generated by adding the same direction, showcasing the generality of the directions found by our method

B.3 Benchmark Details

As mentioned in the main paper, we constructed a custom benchmark for our comparative evaluation. The process began with a large language model (LLM) (OpenAI, 2025), which we used to generate 21 diverse source prompts. For each of these prompts, we generated images using 3 different random seeds, resulting in a set of 63 unique source images. Finally, we applied between 6 to 8 different semantic edits to each source image, depending on the applicability of the edit to the subject. The complete list of source prompts and the specific edits applied to each are detailed in Table 3.

B.4 Quantitative Ablation

To quantitatively measure the contribution of each component of our method, we conduct an ablation study on our benchmark, with results shown in Figure 18. We evaluate three variants of our approach: (1) deriving an edit direction from a single prompt pair, (2) aggregating directions from N prompts but without our proposed injection schedule, and (3) our full method which includes the exponential injection schedule.

The plot of VQA score (prompt alignment) versus LPIPS score (image preservation) reveals the contribution of each component. The single-prompt version serves as our initial baseline and produces a less pronounced semantic change, resulting in a significantly lower VQA score. Aggregating N prompts drastically improves prompt alignment, yielding a much higher VQA score. Our full method, which adds the exponential injection schedule, maintains the high prompt alignment gained from using N prompts while significantly improving image preservation, achieving superior LPIPS scores at all intermediate intensity levels. This validates that both components are crucial for achieving a state-of-the-art balance between edit accuracy and preservation.

.

Figure 14: Composing Disentangled Edits. We demonstrates the compositionality of our learned edit directions. Starting from the source image (top-left), we independently control two attributes of the same subject. The horizontal axis continuously controls the "smile" attribute, while the vertical axis adds "glasses". The smooth and accurate results in the grid showcase our method's ability to combine edits.

Figure 15: Real Image Editing with Image Inversion. Our method seamlessly integrates with inversion techniques, allowing for high-fidelity edits on real-world images. Leveraging UniFlow (Jiao et al., 2025) to invert the source image into the diffusion model's latent space, we demonstrate continuous control over expressions and attributes. The edits maintain the subject's identity and background fidelity across all intensity levels.

B.5 Flux Kontext Baseline

Since Flux Kontext (Labs et al., 2025) lacks a native mechanism for continuous edit scaling, we implemented two distinct proxy baselines to evaluate different edit intensities. The first, which we term Flux Kontext[1] (LLM), controls the edit magnitude by using three different instruction prompts ('light', 'medium', and 'extreme') generated by an LLM, as detailed in Table 2. The second baseline, Flux Kontext[2] (CFG), uses the fixed 'medium' instruction prompt and instead varies the Classifier-Free Guidance (CFG) scale to achieve different levels of edit strength.

Figure 16: Composition of Edits on Multiple Subjects. We demonstrate our method's ability to apply and compose edits targeted at different subjects within the same image. Starting from the source image (top-left), the horizontal axis applies a "laugh" edit exclusively to the woman, while the vertical axis applies an "old" edit only to the man. The results showcase a high degree of localization and disentanglement, as each edit affects only its intended target without interfering with the other subject or the background.

Attribute	1.0 (Low)	2.0 (Medium)	3.0 (High)
Bald	make the person balding	make the person bald	make the person completely bald
Beard	make the person have short beard	make the person have a beard	make the person have a long thick beard
Curly Hair	make the person have slightly curly hair	make the person have curly hair	make the person have very curly hair
Laughing	make the person giggle	make the person laugh	make the person laugh hysterically
Old	make the person middle-aged	make the person old	make the person very old
Smiling	make the person smile slightly	make the person smile	make the person smile broadly
Surprised	make the person slightly surprised	make the person surprised	make the person extremely surprised
Young	make the person slightly young	make the person young	make the person very young

Table 2: Textual descriptions of attribute scales used in our comparison with Flux Kontext

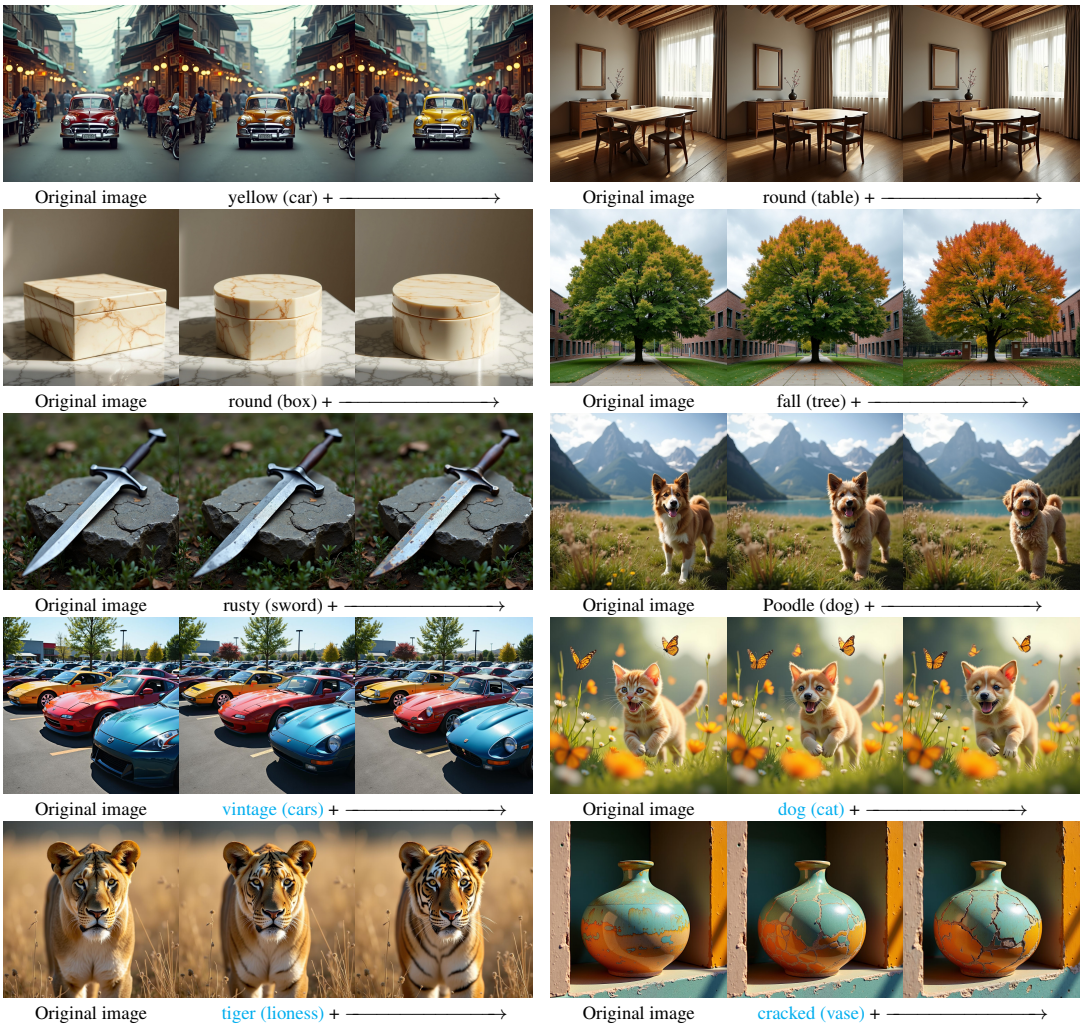

Figure 17: Examples of continuous edits on non-human subjects, showcasing control over seasonal changes, color, and object shape and other attributes.

B.6 Qualitative Comparisons (Continued)

To further evaluate our approach, we provide qualitative comparisons against existing methods, including FluxSpace (Dalva et al., 2024), Concept-Sliders (Gandikota et al., 2023), and two variants of Flux Kontext (Labs et al., 2025): Flux Kontext[1] (LLM), which leverages an LLM to craft prompts for gradual editing, and Flux Kontext[2] (Cfg), which uses the cfg score to guide edits. Results are presented in Figures 21 and 22.

In Figure 21 (left), competing methods fail to introduce a meaningful edit, whereas our method produces a clear and consistent modification. On the right, several baselines either fail to perform the edit or induce significant identity changes. Notably, both Flux Kontext variants are unable to achieve gradual edits and distort subject proportions, often enlarging the head unnaturally. By contrast, our method generates edits that are gradual and identity-preserving.

Figure 22 further illustrates these differences. On the left, competing methods fail to add a beard, produce abrupt transitions, or generate unnatural appearances. Our approach successfully creates a gradual, natural-looking beard. On the right, most baselines again yield non-gradual changes or identity shifts, while our method produces clear, progressive edits that maintain subject identity.

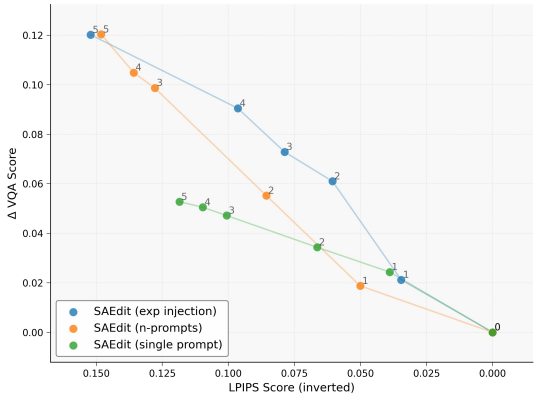

Figure 18: Quantitative Ablation. We compare different versions of our method. (top-right is better).

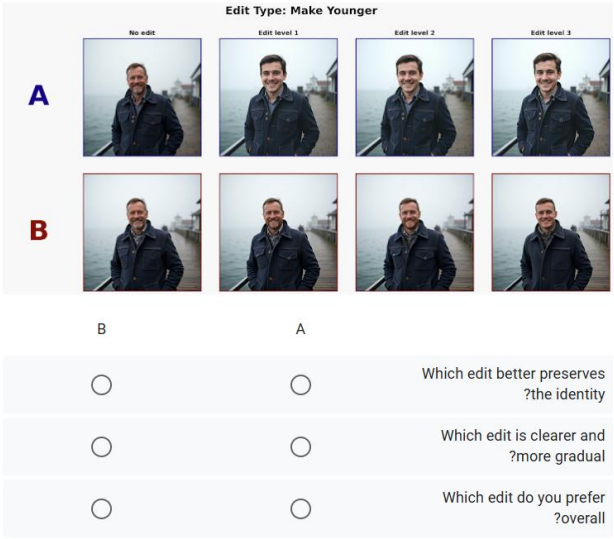

Figure 19: An example of a question in the user study

B.7 User study

As reported in the main text, we conducted a user study to further evaluate the perceptual quality of our method. For this study, we randomly sampled 20 edit scenarios from our quantitative evaluation benchmark.

In each question, we performed a pairwise comparison. Participants were shown the three levels of edit intensity from our method alongside the corresponding three levels from a single competing method. They were then asked to choose which set of edits they preferred based on three criteria:

- **Image Preservation:** Which edits better preserves the identity?
- **Prompt Alignment & Graduality:** Which edits is clearer and more gradual?
- **Overall Preference:** Which edits do you prefer overall?

The exact format of the user study interface is shown in Figure 19.

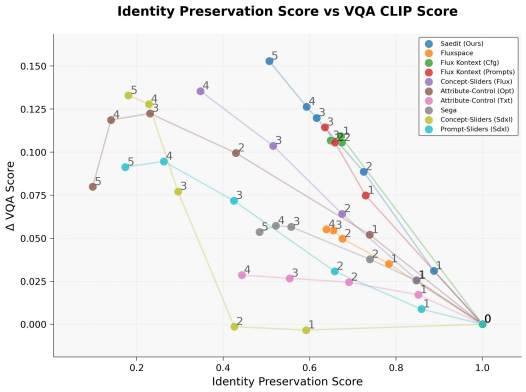

Figure 20: Quantitative comparison. ArcFace is used to compute the identity similarity score between source and target images.

C Sparse AutoEncoders - Continue

Enforcing Sparsity Enforcing sparsity in an SAE's latent space is a central challenge that has led to specialized techniques. One prominent method is the BatchTopK operator (Bussmann et al., 2024), a computationally efficient approach that retains only the top $B \times K$ strongest entries across an entire training batch of size B. At inference, this operator is replaced by a pre-calibrated global threshold (θ) for consistent behavior on single inputs. A common failure mode with such strong sparsity is the emergence of dead latents, which are entries that cease to activate and in turn degrade the SAE's reconstruction performance. To mitigate this, an auxiliary loss, $\mathcal{L}_{\text{aux}}$, can be incorporated (Gao et al., 2024), which encourages these inactive latents to "revive" by tasking them with explaining a portion of the reconstruction error.

Matryoshka Sparse Autoencoders (MSAEs) Bussmann et al. (2025) extend SAEs by learning a single, hierarchical feature dictionary that provides nested representations at multiple levels of granularity. This is achieved by training the model to reconstruct the input using a sequence of nested dictionary subsets of sizes $\mathcal{M} = \{m_1, \ldots, m_n\}$. The training objective minimizes the sum of reconstruction losses across all these levels, along with standard sparsity and auxiliary losses:

$$\mathcal{L} = \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{rec}}(m) + \alpha \mathcal{L}_{\text{sparse}}, \tag{7}$$

where $\mathcal{L}_{\text{rec}}(m)$ is the reconstruction loss using only the first m entries. This encourages the most important features to appear early in the dictionary, creating an ordered representation.

D LLM Usage Statement

We utilized a Large Language Model (LLM) to improve the grammar, spelling, and clarity of this manuscript. The authors critically reviewed and edited all suggestions and bear full responsibility for the accuracy and integrity of the final content.

Prompt	Applied Attributes
Portrait of a woman in a flowing sundress in a field of wildflowers at golden hour	smiling, curly hair, laughing, old, surprised, young
Close-up of a woman in traditional Japanese kimono with cherry blossoms framing her face	smiling, curly hair, laughing, old, surprised, young
woman in business attire portrait in modern glass office building with city skyline	smiling, curly hair, laughing, old, surprised, young
Female pilot in leather jacket portrait next to vintage biplane	smiling, curly hair, laughing, old, surprised, young
woman in rain jacket portrait at lighthouse during coastal storm	smiling, curly hair, laughing, old, surprised, young
Portrait of a woman in bohemian clothing at outdoor art market in Paris	smiling, curly hair, laughing, old, surprised, young
Rock climber woman portrait with climbing gear and canyon background	smiling, curly hair, laughing, old, surprised, young
woman in winter coat portrait with Northern Lights in Finnish Lapland	smiling, curly hair, laughing, old, surprised, young
Female chef in whites portrait in busy restaurant kitchen	smiling, curly hair, laughing, old, surprised, young
Portrait of a woman in wetsuit on surfboard with ocean waves behind	smiling, curly hair, laughing, old, surprised, young
a portrait of a woman violinist in elegant gown in candlelit baroque chamber	smiling, curly hair, laughing, old,, surprised, young
woman in hiking gear portrait at mountain summit with valley vista	smiling, curly hair, laughing, old, surprised, young
Portrait of a man in a worn leather jacket with misty fjord background at dawn	smiling, curly hair, laughing, old, surprised, young, beard, bald
Portrait of a man in traditional samurai armor in a zen garden setting	smiling, curly hair, laughing, old, surprised, young, beard, bald
Portrait of a man wearing hiking gear with tropical canyon vista behind him	smiling, curly hair, laughing, old, smiling, surprised, young, beard, bald
man in fisherman's sweater portrait with foggy dock and sea background	smiling, curly hair, laughing, old, surprised, young, beard, bald
Young man in vintage band t-shirt leaning against 1967 Mustang in desert	smiling, curly hair, laughing, old, surprised, young, beard, bald
Portrait of a man in Renaissance clothing at an easel in Italian courtyard	smiling, curly hair, laughing, old, surprised, young, beard, bald
man in red flannel shirt portrait outside log cabin with falling snow	smiling, curly hair, laughing, old, surprised, young, beard, bald
male chef in whites at sushi counter, portrait with minimalist restaurant background	smiling, curly hair, laughing, old, surprised, young, beard, bald
man wearing panama hat portrait in Marrakech market with colorful spices	smiling, curly hair, laughing, old, surprised, young, beard, bald

Table 3: The complete set of source prompts and their corresponding edit attributes used for our quantitative evaluation and user study.

Figure 21: Each row showcases the results of a different editing method for the same edit. We now show two side-by-side runs (6 images per row). Our method (bottom row) produces a more disentangled result that better preserves the subject's identity compared to the competing approaches.

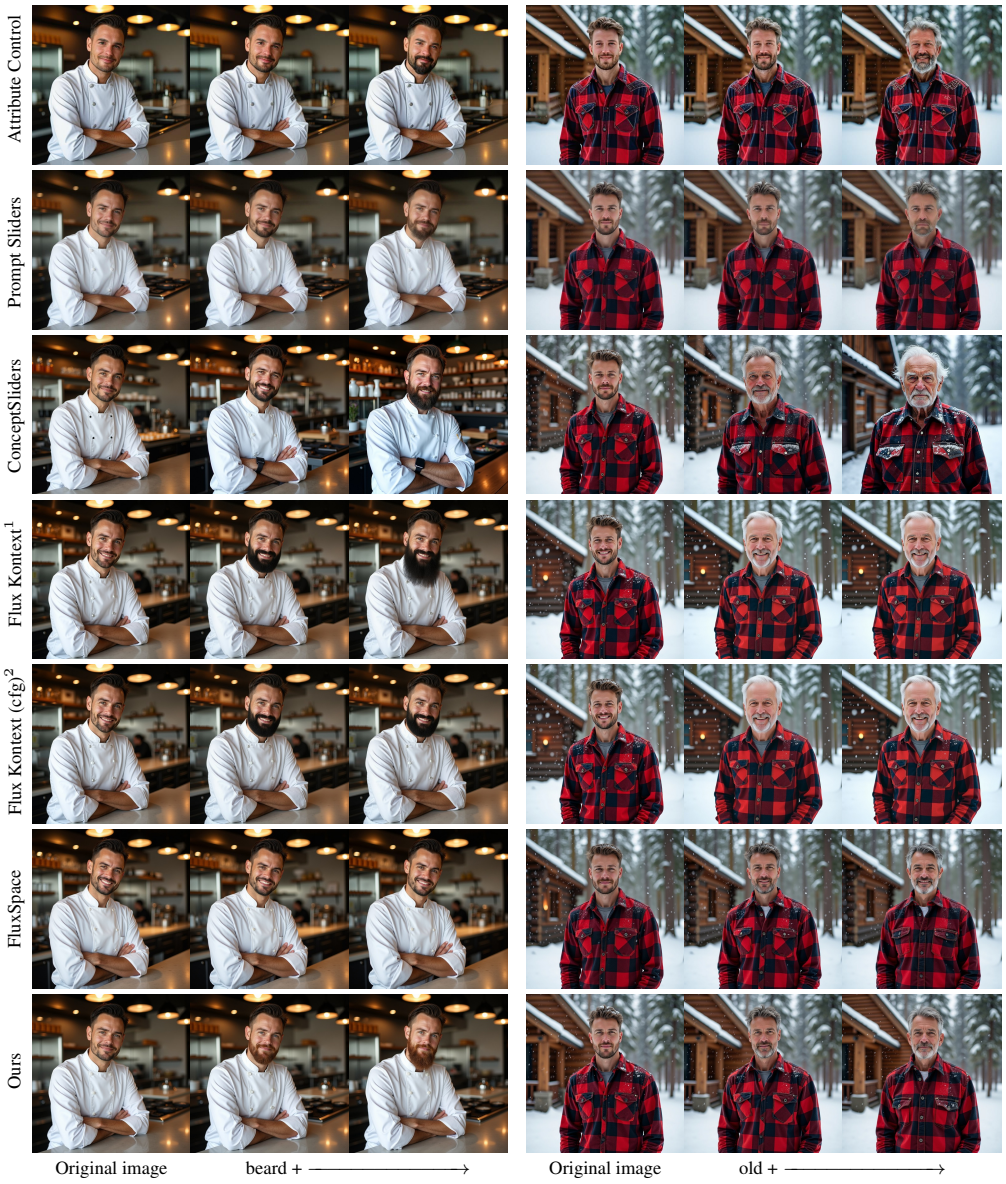

Figure 22: Each row showcases the results of a different editing method for the same edit. We now show two side-by-side runs (6 images per row). Our method (bottom row) produces a more disentangled result that better preserves the subject's identity compared to the competing approaches.

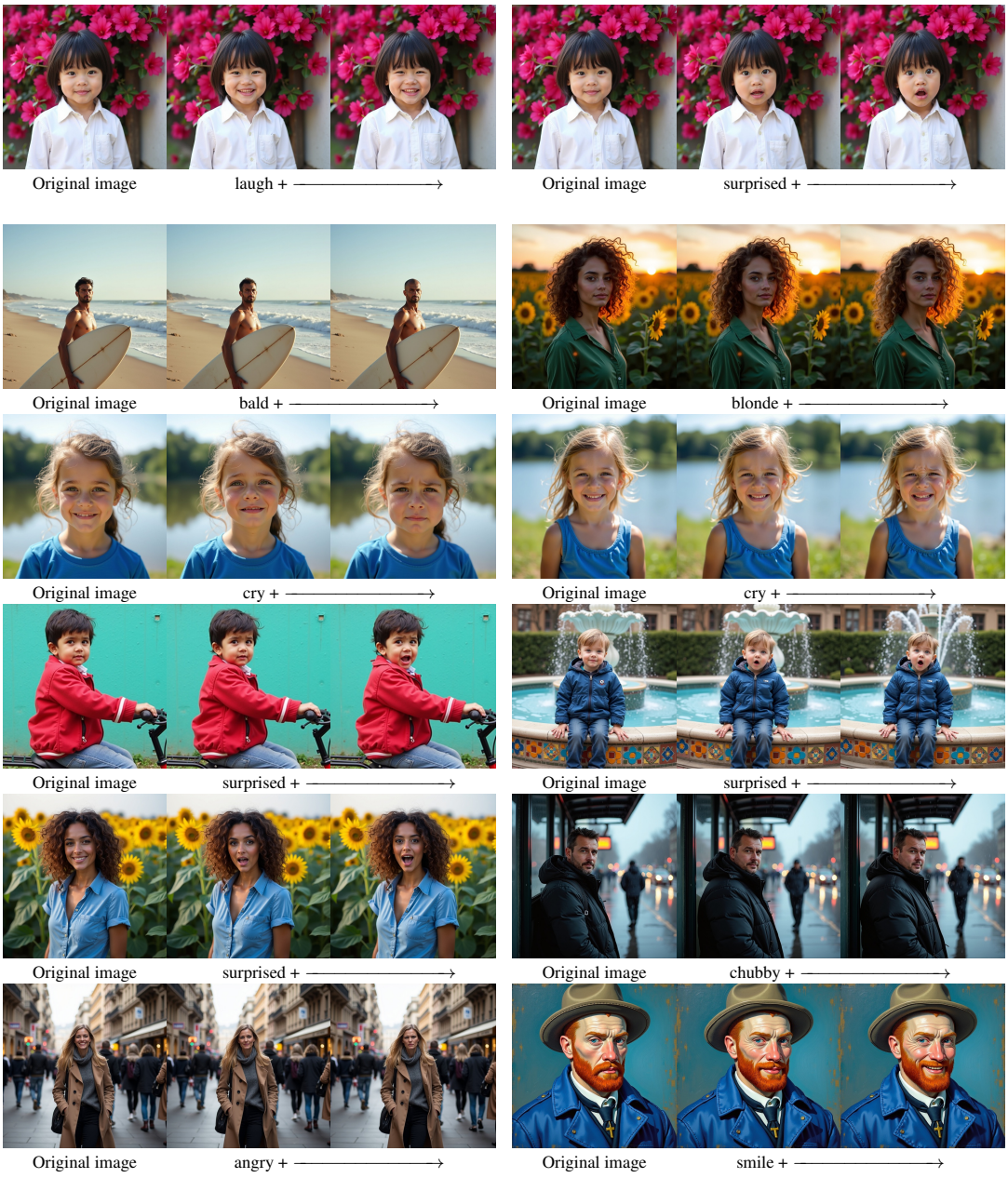

Figure 23: Additional results of our text-based Sliders.

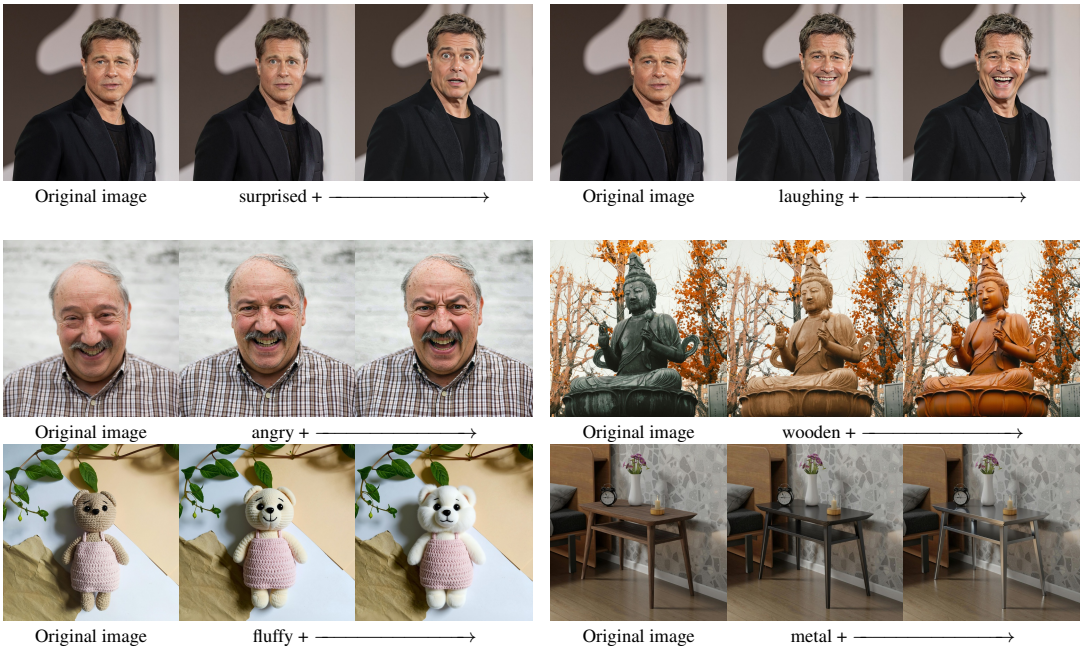

Figure 24: Additional results of our text-based Sliders employed on Flux-Kontext showcasing slider based manipulation on text to image instruction based models.

Figure 25: Comparing with SDXL-based methods. Since our method is Flux-based, we employ inversion on the SDXL-generated reference image when running our method to ensure a fair comparison. The SDXL baselines either fail to preserve identity (top) or produce a weak edit (2nd row). Our method (bottom) maintains the subject's identity while achieving a controllable attribute shift.

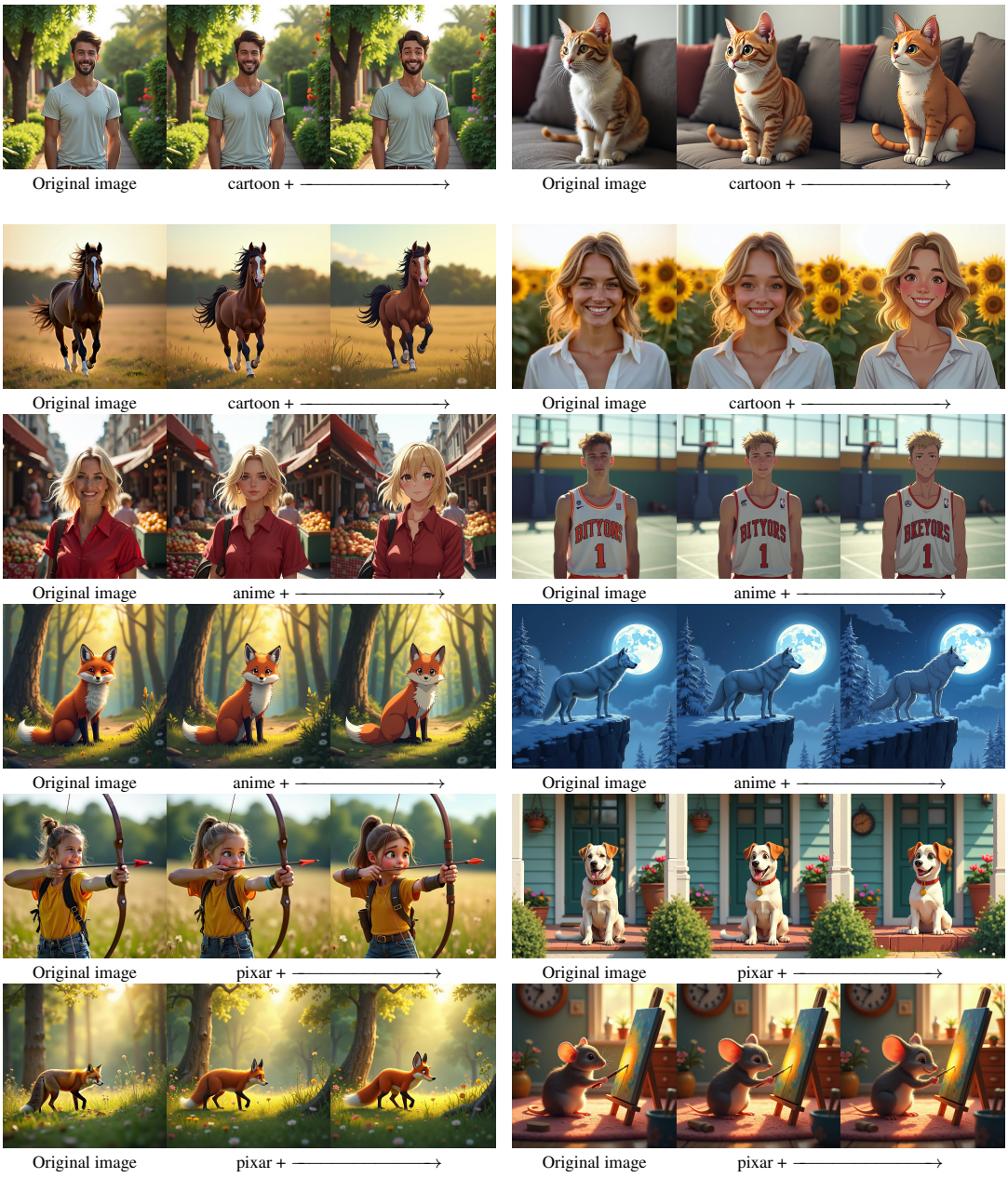

Figure 26: Additional results of our text-based sliders used for gradual style transfer

Figure 27: Consistency across Random Seeds. We apply the identical edit direction to images generated from four consecutive random seeds. This demonstrates the robustness of our method, which consistently applies the intended semantic attribute regardless of the initial noise or resulting scene variations.

Figure 28: We show composition capabilities of our method. We apply an increasing number of sliders from left to right, showcasing our methods ability to concertante multiple edits on the same image. In the second row the added attributes are beard, smile, glasses, blonde hair, old, blue eyes, wide nose. In the third row the added attributes are glasses, smile, beard, cartoon, blonde, blue eyes.