

SAEdit: Token-Level Control for Continuous Image Editing via Sparse Autoencoder

Ronen Kamenetsky¹, Sara Dorfman¹, Daniel Garibi¹, Roni Paiss², Or Patashnik¹, and Daniel Cohen-Or¹

¹ Tel Aviv University

² Google DeepMind

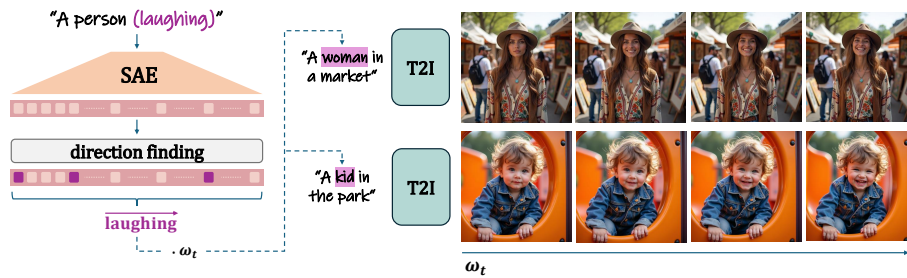


Fig. 1: We train a Sparse AutoEncoder (SAE) to lift the text embeddings into a higher-dimensional space, where we identify disentangled semantic directions (e.g. for laughing). These directions can then be applied to specific tokens within the input of a text-to-image model to facilitate continuous image 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.

Keywords: Diffusion Models · Image Editing



Fig. 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.

1 Introduction

Large-scale text-to-image diffusion models have revolutionized the field of image synthesis [50, 51, 54]. Consequently, they have become a powerful foundation for a wide array of image manipulation and editing methods [10, 31, 41, 61]. 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 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 [4, 15, 29]. 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 [21, 24, 58], 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) [13], 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 applying 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 [6, 32, 48, 50, 54, 56, 57] 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 [1, 10, 31, 46, 47, 61] or applying partial noise schedules with a new text condition [8, 18, 34, 41, 52, 60]. 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 [19, 27, 30, 35, 36, 42, 43, 55].

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 [24], optimizing a text token [58] or to train numerous person-specific DreamBooth LoRAs [53] and then trains a classifier in the latent space [12] or in the weights’ space [21]. Conversely, training-free methods that discover semantic directions in existing latent spaces [4, 7, 15, 20, 28, 29] are general-purpose but often struggle with the precision and disentanglement of specialized models. Another line of work conditions an instruction-based editor on an edit-strength scalar for continuous intensity control [45], at the cost of training on synthesized strength-annotated data. Other works [16, 25] 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 [13]. 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:

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

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 [9, 13, 26]. For example, [9] 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 [2, 5]. The basic SAE framework can be extended with more advanced variants and sparsity regularization techniques, which are detailed further in the supplementary materials.

Recent works have integrated Sparse Autoencoders (SAEs) into diffusion models for various distinct purposes. For instance, *One-Step is Enough* [59] trains SAEs on SDXL cross-attention layers to extract interpretable features from the model’s internal representations. Furthermore, *Concept-Steerers* [38] and *SAEuron* [14] 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 [49], 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 [6] 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

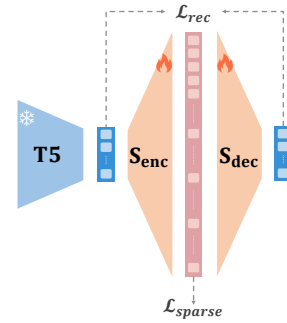


Fig. 3: We train the Sparse Autoencoder on token embeddings obtained from a frozen T5 encoder, using reconstruction and sparsity losses.

are the exact representations that are continuously processed by the Diffusion Transformer (DiT) throughout the denoising steps.

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 [13, 26], 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 a priori which tokens hold the semantic information for a concept [37], 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 $\text{maxpool}(\mathcal{S}_{enc}(\mathcal{P}_{src}))$ and $\text{maxpool}(\mathcal{S}_{enc}(\mathcal{P}_{tgt}))$ 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:

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

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, $R^{norm} = R / \max(R)$,

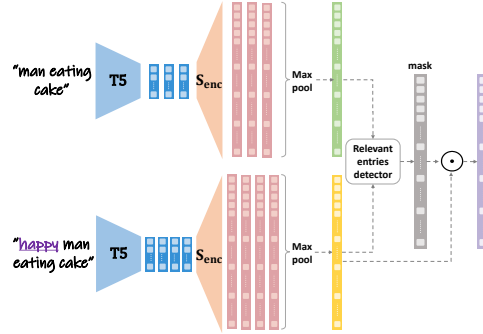


Fig. 4: Extracting Edit Directions. We derive an edit direction from a prompt pair that isolates a single attribute. Both prompts are encoded with the SAE, and their token representations are aggregated via max-pooling. By comparing the two resulting sparse vectors, we identify the key features corresponding 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.

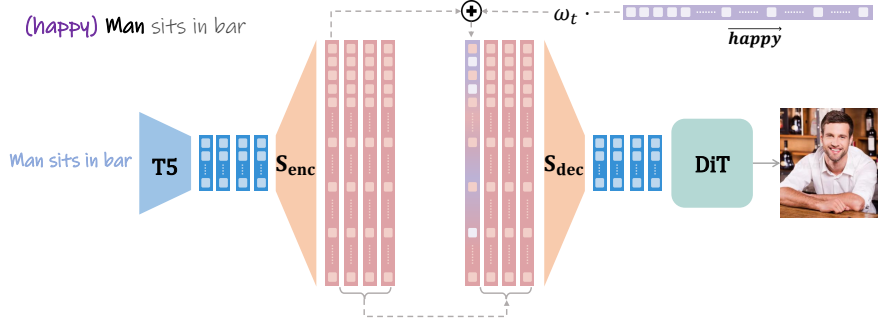


Fig. 5: Applying the edit direction. An aggregated edit direction is scaled to adjust edit magnitude and applied to the sparse representation of the relevant source token (e.g., man). The result is then decoded back into the T5 embedding space, and used to condition the text-to-image model.

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:

$$M = \{i \mid R_i^{norm} > \rho\}. \quad (3)$$

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:

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

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: $D = [d_1, \dots, d_N]^T$. 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} .



Fig. 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), change style (cartoon) and to perform highly localized edits, such as modifying the age of only one person in a scene or adding cracks to a vase.

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 $\mathcal{S}_{dec}$:

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

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 [3, 11, 33, 47, 62]. 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 [34, 35].

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:

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

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 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 [6] and Stable Diffusion 3.5 [22]. 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 [63] and semantic accuracy with a VQA-Score [40]. Implementation details in the Appendix.

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



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

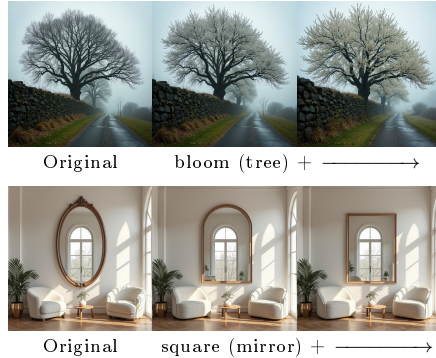


Fig. 8: Our method’s versatility extends beyond human subjects, enabling continuous and disentangled control over object attributes like seasonal appearance and object shape.

produce consistent, high-quality results when applied to a different T5-based model, Stable Diffusion 3.5. We provide additional qualitative results in Appendix, including examples of our method applied to real images, and used for gradual change in style.

5.2 Ablation

Figure 10 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) 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. We complement these qualitative observations with a quantitative ablation in Figure 9 (top-right is better). Beyond our direction-aggregation choices, we assess the necessity of the SAE itself by applying our pipeline directly to the raw T5 embeddings. Since this dense space lacks the sparse structure our method exploits, we adapt the token-level aggregation and report both mean and max pooling variants. This no-SAE baseline suffers from attribute entanglement and insufficient edits, confirming that the SAE’s sparse latent space is necessary for deriving disentangled edit directions. We further ablate replacing our SVD-based direction with mean aggregation, and replacing max-pooling with mean token pooling; full details are provided in the Appendix.

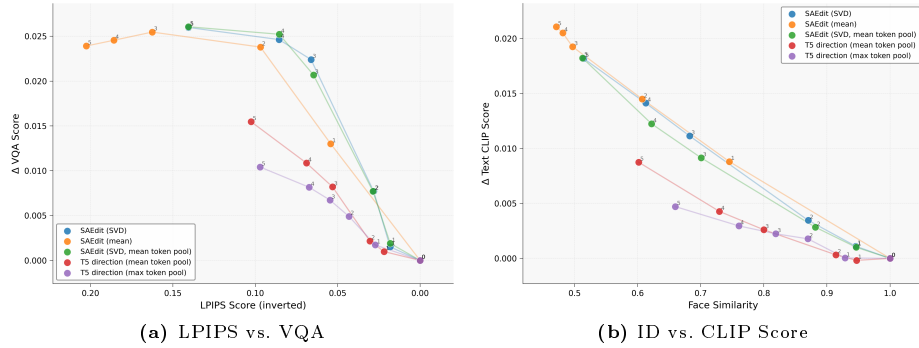


Fig. 9: Quantitative ablations (top-right is better). We compare our full method (SAEdit w/ SVD) against replacing SVD with mean aggregation, replacing max-pooling with mean token pooling, and applying the pipeline directly on dense T5 embeddings (no SAE). Results confirm the necessity of the SAE and our heuristic choices for optimal disentanglement.

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 [24] 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 [58], a textual-inversion [23] based method that learns concept embeddings used for adjustable control. In the training-free category, we compare against FluxSpace [15], adjusting its λ_{fine} parameter to control edit magnitude. We also evaluate against Attribute Control [4], a method that proposes both a training-free and an optimization-based variant. For Flux Kontext [39], which does not natively support continuous edit scaling, we implement two proxy baselines for magnitude control over the edit strength: the

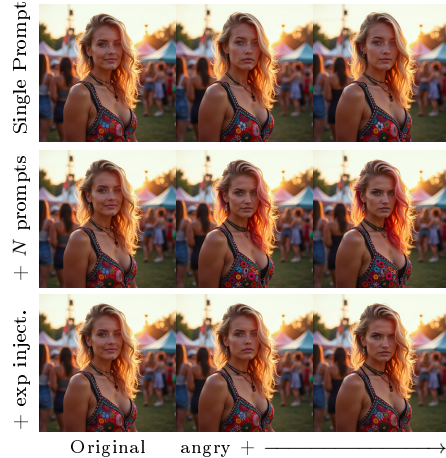


Fig. 10: Ablation study. We demonstrate how each component progressively improves the quality of an 'angry' edit. A direction from a single prompt pair results in a weak edit with unintended modifications. Aggregating N prompts produces a more robust and semantically accurate direction, but can still 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.

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.

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 [44]. 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. 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 [63]. To measure prompt adherence with the edit, we compute a VQA-based score [40]. 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 [17] to calculate the cosine similarity between the subject in the source image and the edited output.

Figure 11 presents the quantitative comparison between our method and other methods at varying levels of edit intensity. Additional metrics showcasing facial similarity preservation are in the appendix. 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.

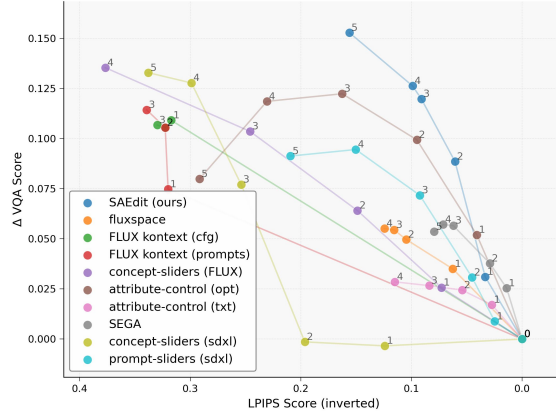


Fig. 11: Quantitative comparison. We compare our method to other baselines.

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

Opponent Method	Image Prompt		Overall
	Pres.	Adher.	
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.

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 13 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). 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 the appendix. 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 the appendix.

5.4 Integration With Flux-Kontext

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 (in the appendix). 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,

its 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 12, 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.



Fig. 12: 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).

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.

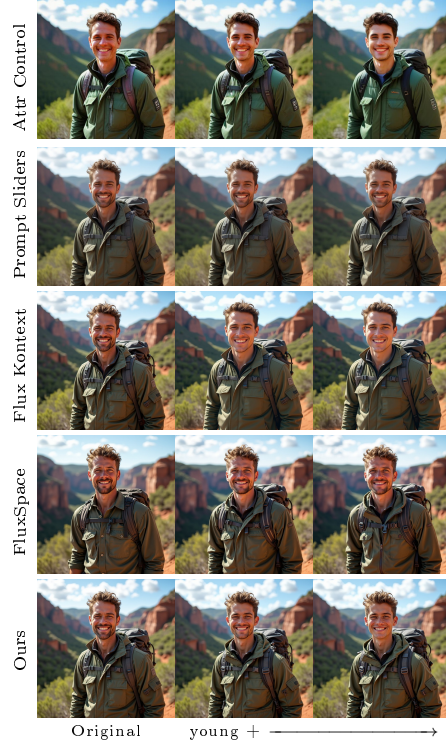


Fig. 13: Each row showcases the results of a different editing method for the same edit. Our method (bottom row) produces a more disentangled result that better preserves the subject’s identity.

References

1. Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., Cohen-Or, D.: Cross-image attention for zero-shot appearance transfer (2023)
2. Arad, D., Mueller, A., Belinkov, Y.: Saes are good for steering – if you select the right features (2025), <https://arxiv.org/abs/2505.20063>
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers (2023), <https://arxiv.org/abs/2211.01324>
4. Baumann, S.A., Krause, F., Neumayr, M., Stracke, N., Sevi, M., Hu, V.T., Omer, B.: Continuous, subject-specific attribute control in t2i models by identifying semantic directions (2025), <https://arxiv.org/abs/2403.17064>
5. Bayat, R., Rahimi-Kalahroudi, A., Pezeshki, M., Chandar, S., Vincent, P.: Steering large language model activations in sparse spaces (2025), <https://arxiv.org/abs/2503.00177>
6. Black Forest Labs: Flux (2024), <https://github.com/black-forest-labs/flux>, <https://github.com/black-forest-labs/flux>, accessed: 28 June 2026
7. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing text-to-image models using semantic guidance (2023), <https://arxiv.org/abs/2301.12247>
8. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. arXiv preprint arXiv:2311.16711 (2023)
9. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T., Olah, C.: Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread (2023), <https://transformer-circuits.pub/2023/monosemantic-features/index.html>, accessed: 28 June 2026
10. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
11. Cao, Y., Zhao, Z., Patras, I., Gong, S.: Temporal score analysis for understanding and correcting diffusion artifacts (2025), <https://arxiv.org/abs/2503.16218>
12. Chang, Y., Yao, Y., Qin, T., Wang, M., Tsang, I., Dai, G.: Instructing text-to-image diffusion models via classifier-guided semantic optimization (2025), <https://arxiv.org/abs/2505.14254>
13. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models (2023), <https://arxiv.org/abs/2309.08600>
14. Cywiński, B., Deja, K.: Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders (2025), <https://arxiv.org/abs/2501.18052>
15. Dalva, Y., Venkatesh, K., Yanardag, P.: Fluxspace: Disentangled semantic editing in rectified flow transformers (2024)
16. Dalva, Y., Yanardag, P.: Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models (2023), <https://arxiv.org/abs/2312.05390>

17. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(10), 5962–5979 (Oct 2022). <https://doi.org/10.1109/tpami.2021.3087709>, <http://dx.doi.org/10.1109/TPAMI.2021.3087709>
18. Deutch, G., Gal, R., Garibi, D., Patashnik, O., Cohen-Or, D.: Turboedit: Text-based image editing using few-step diffusion models (2024), <https://arxiv.org/abs/2408.00735>
19. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
20. Dorfman, S., Cohen-Bar, D., Gal, R., Cohen-Or, D.: Ip-composer: Semantic composition of visual concepts (2025), <https://arxiv.org/abs/2502.13951>
21. Dravid, A., Gandelsman, Y., Wang, K.C., Abdal, R., Wetzstein, G., Efros, A.A., Aberman, K.: Interpreting the weight space of customized diffusion models (2024), <https://arxiv.org/abs/2406.09413>
22. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), <https://arxiv.org/abs/2403.03206>
23. 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 (2022), <https://arxiv.org/abs/2208.01618>
24. Gandikota, R., Materzynska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: Lora adaptors for precise control in diffusion models (2023), <https://arxiv.org/abs/2311.12092>
25. Gandikota, R., Wu, Z., Zhang, R., Bau, D., Shechtman, E., Kolkin, N.: Sliderspace: Decomposing the visual capabilities of diffusion models. In: *Proceedings of the IEEE/CVF international conference on computer vision* (2025), arXiv:2502.01639
26. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders (2024), <https://arxiv.org/abs/2406.04093>
27. Garibi, D., Patashnik, O., Voynov, A., Averbuch-Elor, H., Cohen-Or, D.: Renoise: Real image inversion through iterative noising (2024), <https://arxiv.org/abs/2403.14602>
28. Garibi, D., Yadin, S., Paiss, R., Tov, O., Zada, S., Ephrat, A., Michaeli, T., Mosseri, I., Dekel, T.: Tokenverse: Versatile multi-concept personalization in token modulation space (2025), <https://arxiv.org/abs/2501.12224>
29. Guerrero-Viu, J., Hasan, M., Roullier, A., Harikumar, M., Hu, Y., Guerrero, P., Gutiérrez, D., Masia, B., Deschaintre, V.: Texsliders: Diffusion-based texture editing in clip space. In: *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. p. 1–11. SIGGRAPH '24, ACM (Jul 2024). <https://doi.org/10.1145/3641519.3657444>, <http://dx.doi.org/10.1145/3641519.3657444>
30. Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., 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)
31. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022)
32. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
33. Huberman, S., Patashnik, O., Dahary, O., Mokady, R., Cohen-Or, D.: Image generation from contextually-contradictory prompts. *arXiv preprint arXiv:2506.01929* (2025)

34. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations (2023)
35. Jiao, G., Huang, B., Wang, K.C., Liao, R.: Unedit-flow: Unleashing inversion and editing in the era of flow models (2025), <https://arxiv.org/abs/2504.13109>
36. Kadosh, E., Goren, N., Patashnik, O., Garibi, D., Cohen-Or, D.: Tight inversion: Image-conditioned inversion for real image editing (2025), <https://arxiv.org/abs/2502.20376>
37. Kaplan, G., Toker, M., Reif, Y., Belinkov, Y., Schwartz, R.: Follow the flow: On information flow across textual tokens in text-to-image models (2025), <https://arxiv.org/abs/2504.01137>
38. Kim, D., Ghadiyaram, D.: Concept steerers: Leveraging k-sparse autoencoders for test-time controllable generations (2025), <https://arxiv.org/abs/2501.19066>
39. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742>
40. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation (2024), <https://arxiv.org/abs/2404.01291>
41. 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 (2022)
42. Miyake, D., Iohara, A., Saito, Y., Tanaka, T.: Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models (2023)
43. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models (2022)
44. OpenAI: Chatgpt (gpt-5). <https://chat.openai.com/> (2025), accessed: 28 June 2026
45. Parihar, R., Patashnik, O., Ostashev, D., Babu, R.V., Cohen-Or, D., Wang, K.C.: Kontinuuous kontext: Continuous strength control for instruction-based image editing (2025), <https://arxiv.org/abs/2510.08532>
46. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. SIGGRAPH '23, ACM (Jul 2023). <https://doi.org/10.1145/3588432.3591513>, <http://dx.doi.org/10.1145/3588432.3591513>
47. Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
48. 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 (2023)
49. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023), <https://arxiv.org/abs/1910.10683>
50. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)

52. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic image inversion and editing using rectified stochastic differential equations (2024), <https://arxiv.org/abs/2410.10792>
53. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation (2023), <https://arxiv.org/abs/2208.12242>
54. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
55. Samuel, D., Meiri, B., Maron, H., Tewel, Y., Darshan, N., Avidan, S., Chechik, G., Ben-Ari, R.: Lightning-fast image inversion and editing for text-to-image diffusion models (2024), <https://arxiv.org/abs/2312.12540>
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022)
57. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
58. Sridhar, D., Vasconcelos, N.: Prompt sliders for fine-grained control, editing and erasing of concepts in diffusion models (2024), <https://arxiv.org/abs/2409.16535>
59. Surkov, V., Wendler, C., Mari, A., Terekhov, M., Deschenaux, J., West, R., Gulcehre, C., Bau, D.: One-step is enough: Sparse autoencoders for text-to-image diffusion models (2025), <https://arxiv.org/abs/2410.22366>
60. Tsaban, L., Passos, A.: Ledits: Real image editing with ddpm inversion and semantic guidance (2023), <https://arxiv.org/abs/2307.00522>
61. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1921–1930 (June 2023)
62. Yehezkel, S., Dahary, O., Voynov, A., Cohen-Or, D.: Navigating with annealing guidance scale in diffusion space. arXiv preprint arXiv:2506.24108 (2025)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)