FluxSpace: Disentangled Semantic Editing in Rectified Flow Models

Yusuf Dalva Kavana Venkatesh Pinar Yanardag
Virginia Tech
https://fluxspace.github.io

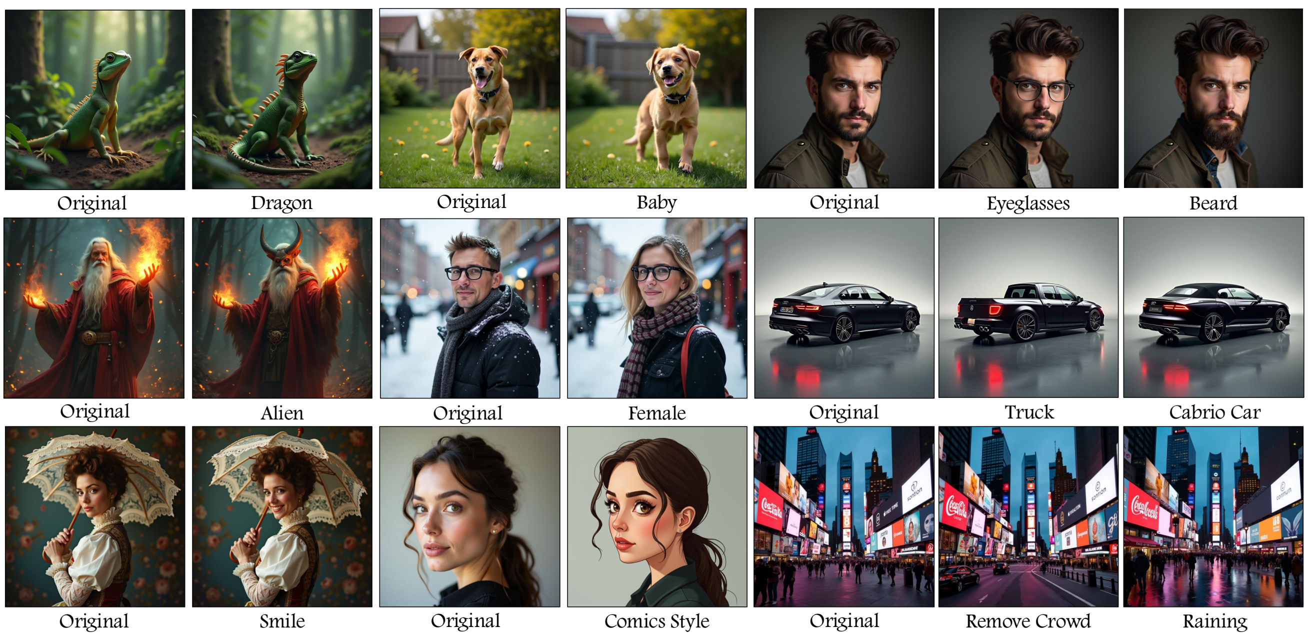

Figure 1. **FluxSpace.** We propose a text-guided image editing approach on rectified flow transformers [15], such as Flux. Our method can generalize to semantic edits on different domains such as humans, animals, cars, and extends to even more complex scenes such as an image of a street (third row, last example). FluxSpace can apply edits described as keywords (e.g. "truck" for transforming a car into a truck) and offers disentangled editing capabilities that do not require manually provided masks to target a specific aspect in the original image. In addition, our method does not require any training and can apply the desired edit during inference time.

Abstract

Rectified flow models have emerged as a dominant approach in image generation, showcasing impressive capabilities in high-quality image synthesis. However, despite their effectiveness in visual generation, rectified flow models often struggle with disentangled editing of images. This limitation prevents the ability to perform precise, attribute-specific modifications without affecting unrelated aspects of the image. In this paper, we introduce FluxSpace, a domain-agnostic image editing method leveraging a representation space with the ability to control the semantics of images generated by rectified flow transformers, such as Flux. By leveraging the representations learned by the transformer blocks within the rectified flow models, we propose a set of semantically interpretable representations that enable a wide range of image editing tasks, from fine-grained image editing to artistic creation. This work offers a scalable and effective image editing approach, along with its disentanglement capabilities.

1. Introduction

Recent research aimed at enhancing the interpretability of generative models has increasingly focused on disentangled editing capabilities, which allows precise control over specific features or attributes within generated images [21, 43, 45]. These capabilities are crucial for understanding and manipulating the outputs of generative models, such as Generative Adversarial Networks (GANs) and multi-step diffusion models like Stable Diffusion. In single-step models such as GANs [18], the editing process is facilitated by a structured and highly disentangled latent space. This space, composed of fixed-dimension vectors, linearly encodes meaningful concepts that can be individually adjusted to effectively alter specific aspects of the output images. However, achieving similar disentangled editing in multi-step diffusion models presents a significant challenge. Unlike GANs, diffusion models do not utilize a fixed-dimension latent space; instead, they generate images through a multi-step refinement process. This process gradually transitions from a random noise distribution to the data distribution, with each step involving complex interactions of learned noise patterns. These patterns are inherently difficult to map back to specific features in a latent space. Moreover, the sequential nature of diffusion models complicates the internal dynamics and obscures the direct correlations between dimensions of a hypothetical latent space and specific visual features in the generated output.

Disentangled latent spaces have been explored thoroughly in diffusion models, focusing on the UNet bottleneck layer [24, 31], text embedding space [3], noise space [9] and weight space [13, 17]. However, these approaches often suffer from significant limitations in terms of disentangled semantics. Manipulation of the UNet bottleneck layer [24, 31] may not fully capture the high-level semantic features necessary for fine-grained control, as bottleneck representations can still be highly entangled. Adjusting the text embedding space [3] relies heavily on alignment between textual and visual features, which may not always correspond accurately, leading to entangled semantics. Operating within the noise space [9] presents challenges due to the difficulty of associating specific noise patterns and semantics. Other works [13] explored the weight space to perform disentangled editing; however, they require time and resource consuming training of LoRA models to create a latent space for each domain. On the other hand, generative models based on flow-matching transformers (e.g. Flux) offer high-quality image generation; however, their disentanglement editing capabilities remain underexplored.

In this paper, we focus on the latent spaces in flow-matching transformers, characterized by their ability to generate images with high fidelity [15]. Our analysis reveals that the joint transformer blocks, integral to the denoising network, are adept at encoding highly disentangled semantic information. Since the architecture we target involves transformers, the attention layer outputs gradually add image content to the latent representation that gets denoised, without any residual connections between different blocks, unlike UNet-based architectures. Furthermore, the attention layers included in these blocks enable enhanced control over the model outputs, as they control the image content in isolation compared to succeeding and preceding blocks of the denoising network. By implementing a linear editing scheme across these attention outputs, we unlock the potential for semantic editing within flow-matching transformers, facilitating semantic navigation across their output space. Our approach supports precise, fine-grained edits, such as adding a smile, as well as broader, coarse-level modifications, such as stylization. We perform qualitative and quantitative experiments with various state-of-the-art methods and demonstrate the effectiveness of our method in enabling fine-grained image editing tasks. Our contributions are as follows.

- We introduce FluxSpace, a novel editing framework within flow-matching transformers using attention layer outputs, unlocking advanced semantic editing capabilities for precise navigation within the model's output space.
- We demonstrate the capability of joint transformer blocks to encode highly disentangled semantic information through incremental content refinement.
- We introduce support for both fine-grained edits, such as smile addition, and coarse-level modifications such as stylization, improving the model's versatility in image manipulation. Our method supports the editing of both real and generated images.

2. Related Work

Latent Space Exploration of Diffusion Models. Diffusion models, which have been the dominant approach for the tasks of text-to-image generation and editing, encapsulate rich semantics within their latent representations. To facilitate new applications, research has focused on leveraging the semantics encoded by such models. In addition to the explorations performed on latent spaces, certain methods [24, 42] explored image editing techniques that alter the backward diffusion process, targeting the learned latent variables. Specifically, [24] bases its approach on the features learned by the bottleneck block of the denoising model, which has a UNet-based architecture, while [42] modifies the intermediate latent variables using stochastic diffusion models. As a further advancement toward this area, [31] introduces a method that aims to identify latent-specific directions representing various semantics, inspired by latent space exploration methods in GANs [10, 22, 32, 45]. Although such methods promise the discovery of semantic directions in domain-specific DDPMs [20], their application to large-scale diffusion models [15, 34, 37] is limited. In addition, various approaches have studied alternative latent spaces to propose directions for semantic editing, such as the noise space [9], the text embedding space [3], and the weight space [13, 17]. Despite these efforts that propose different formulations of semantic directions on diffusion models, such an approach still does not exist for flow-matching transformer models, which is addressed by our proposed FluxSpace framework.

Image Editing with Diffusion Models. Due to their high-fidelity generation capabilities, text-to-image diffusion models have been used frequently for image generation and editing tasks. As a straightforward strategy, one can perform image editing with a target text prompt that effectively describes the desired visual effect. However, this strategy can lead to entangled edits, where the change described as a text prompt modifies certain aspects of the image that are not intended to be edited. Addressing this problem, studies such as [19, 46] enabled increased precision within the editing process. Among these works, [46] uses a conditional diffusion model that guides the generation process based on user-specified controls. Similarly, [41] attempts to preserve image content by fine-tuning the diffusion model. On the other hand, several methods benefit from the representations learned by the diffusion models. Among these methods, [19, 40] use attention control mechanisms within the diffusion model, [6, 7] use the semantic guidance mechanism over the noise space, and [3, 44] utilize the text-embedding representations to perform edits. Furthermore, studies such as [11, 17] succeeded in applying disentangled edits by trained semantic directions. However, these methods are limited in terms of requiring training-per-edit to perform the desired image-to-image transformation, which is either specified as a set of textual conditions or paired images reflecting the desired edit. Addressing the problems in both directions, we focus on disentangled editing on rectified flow transformers, without any additional training.

Rectified Flow-Based Models. Succeeding over diffusion models, transformers trained with the flow-matching objective [25, 26], such as Flux, now serve as the state-of-the-art text-to-image generation models [15]. Despite their superior generation capabilities, the methods applied to diffusion models for image editing are not directly transferable to rectified flow models, since the MM-DiT architecture [33] involves continuous interaction between text and image features, which results in changes in both representations, unlike diffusion models [34, 37]. Furthermore, disentangled editing with existing methods is not possible in a straightforward manner. As a preliminary effort towards text-based image editing with rectified flows, [38] offers an editing method with target captions specifying the desired edit. However, existing methods are limited in terms of the types of edits available and edit scale adjustment, whereas our method offers such functionalities.

3. Preliminaries

3.1. Rectified-Flow Models

Generative models aim to define a mapping from samples x_1 from a noise distribution p_1 to samples x_0 from a data distribution p_0, where p_0 represents real images in image generation tasks. Rectified flows [25, 26] define a time-dependent forward process that constructs paths between distributions p_0 and p_1 as straight trajectories, as shown in Eq. 1, where $p_1 = \mathcal{N}(0, 1)$ and timestep t.

$$x_t = (1-t)x_0 + t\epsilon, \quad \epsilon \sim N(0,1) \tag{1}$$

To learn this mapping, a network is trained with parameters θ, to estimate the velocity v of the rectified flow, represented by v_θ. This velocity prediction network can serve as a noise prediction network, ϵ_θ, optimized using the Conditional Flow Matching (CFM) objective formulated as Eq. 2, with signal-to-noise ratio λ'_t and time-dependent weighting function w_t [15].

$$\mathcal{L}_{CFM} = -\frac{1}{2}\mathbb{E}_{t\sim\mathcal{U}(t),\epsilon\sim\mathcal{N}(0,I)}[w_t\lambda'_t||\epsilon_\theta(x_t,t)-\epsilon||^2] \tag{2}$$

3.2. Multi-Modal Diffusion Transformers

The multi-modal diffusion transformer architecture, multi-modal diffusion transformers (MM-DiT) [33], integrates both text and image modalities to generate images aligned with the semantics implied by text inputs. Among the state-of-the-art text-to-image generation models, rectified flow transformers (e.g. Flux) utilize the adaptation of this architecture introduced in [15], which involves modulated attention and MLP layers, followed by attention layers l_θ parametrized by parameters θ, conditioning image generation on both pooled and token-wise text embeddings.

To guide image generation with text inputs, rectified flow transformers use two text embeddings: c_{pool} and c_{ctxt}. The pooled embedding c_{pool}, derived from the CLIP Text Encoder [35], is used in the modulation mechanism to scale and shift features fed as input to the attention layers. The token-wise embeddings c_{ctxt}, obtained from a T5 Text Encoder [36], ensure alignment with prompt semantics, thus enhancing the relevance of the generated images. The complete conditioning set, $\{c_{pool}, c_{ctxt}\}$, improves prompt alignment, as shown in previous work [15, 34]. Both embeddings are integrated into joint transformer blocks, where text and image features interact through distinct query (Q), key (K), and value (V) transformations. These blocks enable influence across modalities (text and image) in a bidirectional manner, forming the foundation for the image editing mechanism introduced in this work.

4. Methodology

Our approach leverages the architecture of flow-matching transformers, specifically Flux, to introduce a structured method for controlled image editing. In Flux, image generation progresses by refining content from initial noise through multi-modal transformer blocks. At each timestep t, the noise prediction network $\epsilon_\theta(x_t, c, t)$ outputs a prediction conditioned on noisy latent x_t and text-based conditioning c (composed of c_{pool} and c_{ctxt}). This stepwise refinement occurs within the attention layers l_θ, creating a

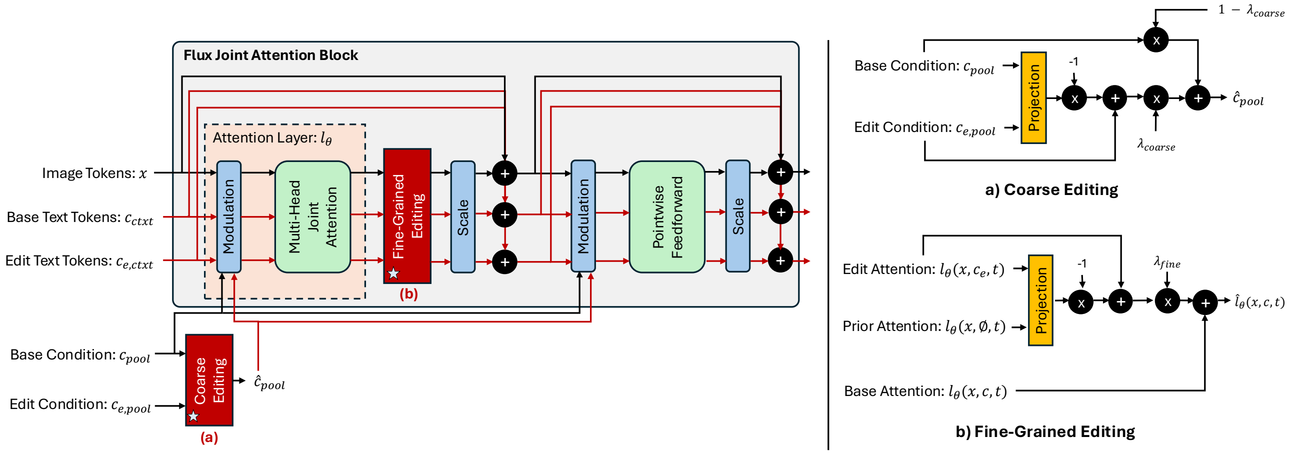

Figure 2. **FluxSpace Framework.** The FluxSpace framework introduces a dual-level editing scheme within the joint transformer blocks of Flux, enabling coarse and fine-grained visual editing. Coarse editing operates on pooled representations of base (c_{pool}) and edit ($c_{e,pool}$) conditions, allowing global changes like stylization, controlled by the scale λ_{coarse} (a). For fine-grained editing, we define a linear editing scheme using base, prior, and edit attention outputs, guided by scale λ_{fine} (b). With this flexible design, our framework is both able to perform coarse-level and fine-grained editing, with a linearly adjustable scale.

progressively evolving representation space suitable for semantic manipulations. For simplicity, we represent the conditional input to the attention layer as c, where it contains both the modulating text embedding c_{pool} and the token-wise text condition c_{ctxt}, which both influences the attention mechanism within Flux.

Our method designates the outputs of Flux's joint attention layers as *FluxSpace*, a linear representation space where semantic image edits can be performed in a disentangled manner. Within this space, we enable transformations on the attention outputs, allowing semantic modifications such as detailed adjustments to object attributes or overall changes in style, without altering the model parameters.

Our primary objective is to introduce an inference-time image editing algorithm with flexible levels of control. The first option focuses on detailed edits, allowing fine-grained adjustments, such as adding a smile to a face. The second option provides control over the image's coarse appearance, such as changing the overall style.

4.1. Fine-Grained Editing

Given the pre-trained attention layer l_θ and the image tokens from the noisy image input x_t, we employ three different outputs obtained by this layer. With a minor use of notation, we use x for image tokens from the noisy input, both for simplicity and to illustrate the relation between x_t and image tokens. First, we utilize the output $l_\theta(x, c, t)$ where c is the text condition used to generate the unedited image. Following, we compute two additional features, $l_\theta(x, c_e, t)$ and $l_\theta(x, \phi, t)$ corresponding to the outputs w.r.t. editing condition c_e (e.g. "eyeglasses" for the addition of eyeglasses) and null text ϕ. In our editing scheme, we treat $l_\theta(x, \phi, t)$ as an image prior, which reflects the attention layers' knowledge of the image features without any text condition.

Our framework relies on the linearity assumption of attention outputs, which enables us to define latent directions on a given input condition c_e. First, to isolate the conditional output $l_\theta(x_t, c_e, t)$ from the image-related details, we apply an orthogonal projection with the image prior retrieved as the null prediction $l_\theta(x_t, \phi, t)$, to obtain $\text{proj}_\phi l_\theta(x, c_e, t)$ formulated over Eq. 3.

$$\text{proj}_\phi l_\theta(x, c_e, t) = \frac{l_\theta(x, c_e, t) \cdot l_\theta(x, \phi, t)}{||l_\theta(x, \phi, t)||^2} l_\theta(x, \phi, t) \quad (3)$$

Given the projection of attention outputs, we identify the orthogonal component of $l_\theta(x, c_e, t)$ over $l_\theta(x, \phi, t)$ as $l'_\theta(x, c_e, t)$ in Eq. 4. With this vector, we identify a semantic direction that is effective in latent pixels to shift image content w.r.t. editing prompt c_e.

$$l'_\theta(x, c_e, t) = l_\theta(x, c_e, t) - \text{proj}_\phi l_\theta(x, c_e, t) \quad (4)$$

Using this linear direction, we formulate our editing scheme in Eq. 5. As we define our editing method in the form of linear interpolation between the editing vector $l'_\theta(x, c_e, t)$, our method is also able to perform edit interpolation with the edit strength parameter λ_{fine}.

$$\hat{l}_\theta(x, c, c_e, t) = l_\theta(x, c, t) + \lambda_{fine}(l'_\theta(x, c_e, t)) \quad (5)$$

Content Preservation with Attention Masking. To facilitate further disentanglement over the performed edit, we introduce a self-supervised mask, based on the interaction of

Figure 3. **Qualitative Results on Face Editing.** Our method can perform a variety of edits from fine-grained face editing (e.g. adding eyeglasses) to changes over the overall structure of the image (e.g. comics style). As our method utilizes disentangled representations to perform image editing, we can precisely edit a variety of attributes while preserving the properties of the original image.

the image features and the editing condition. First, we introduce the mask $M_{i,edit}$ based on the query features Q_i, computed with the image features, and the key features K_{edit} with the editing condition in Eq. 6. Intuitively, we query the image for the pixels that respond strongly to the given text condition as an intermediate estimate, which is used to mask out the pixels with low attention values. Following the existing work on the T5 text encoder, which is used in Flux, we utilize the attention map of the first text token [29].

$$M_{i,edit} = softmax\left(\frac{Q_i \cdot K_{edit}^T}{\sqrt{d}}\right) \tag{6}$$

Given the mask $M_{i,edit}$, we introduce a soft decision boundary $M'_{i,edit}$ with boundary coefficient $b = 10$ [14], sigmoid operator σ, and min-max normalization operator $normalize$, formulated as Eq. 7.

$$M'_{i,edit} = \sigma(b * (normalize(M_{i,edit}) - 0.5)) \tag{7}$$

As the final step, we perform a thresholding operation with the parameter τ_m to retrieve the thresholding mask $M''_{i,edit}$ in Eq. 8.

$$M''_{i,edit} = \begin{cases} 1 & \text{if } M'_{i,edit} \geq \tau_m \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Using this thresholding mask, we modify the content editing equation presented in Eq. 5 by masking out the latent pixels that receive low attention from the editing direction $l'_\theta(x, c_e, t)$.

4.2. Editing Coarse Level Details

Before performing the attention calculation, Flux applies a modulation based on the pooled CLIP embeddings that provides information about the coarse structure of the generated image [34]. Since certain edits need to change the overall structure and appearance of the image, we introduce an additional control mechanism for appearance-based changes. Specifically, we extend our editing approach based on the orthogonal projection of attention features in the CLIP embeddings, where we edit the pooled generation condition c_{pool} in the direction of the pooled editing condition $c_{e,pool}$ to retrieve $\hat{c}_{pool}$, which is used to normalize the attention features.

Since we want to edit the coarse condition in a disentangled manner, we perform our editing scheme based on the application of the linear representation hypothesis [1, 2, 16, 27, 30] on pooled CLIP embeddings [5]. To do so, we first perform an orthogonal projection in Eq. 9, for the editing concept $c_{e,pool}$ on c_{pool} to retrieve the editing direction for the coarse representation of the image. However, different from the attention features, we use the base generation condition as the basis of the projection, as we now operate on textual representations rather than multimodal representations.

$$c'_{e,pool} = c_{e,pool} - \frac{c_{pool} \cdot c_{e,pool}}{||c_{pool}||^2} c_{pool} \tag{9}$$

Given the linear direction for editing the coarse condition, we formulate the edit as a linear interpolation between the original condition and the editing direction with a scale parameter λ_{coarse}, which controls the extent of editing of the modulation condition.

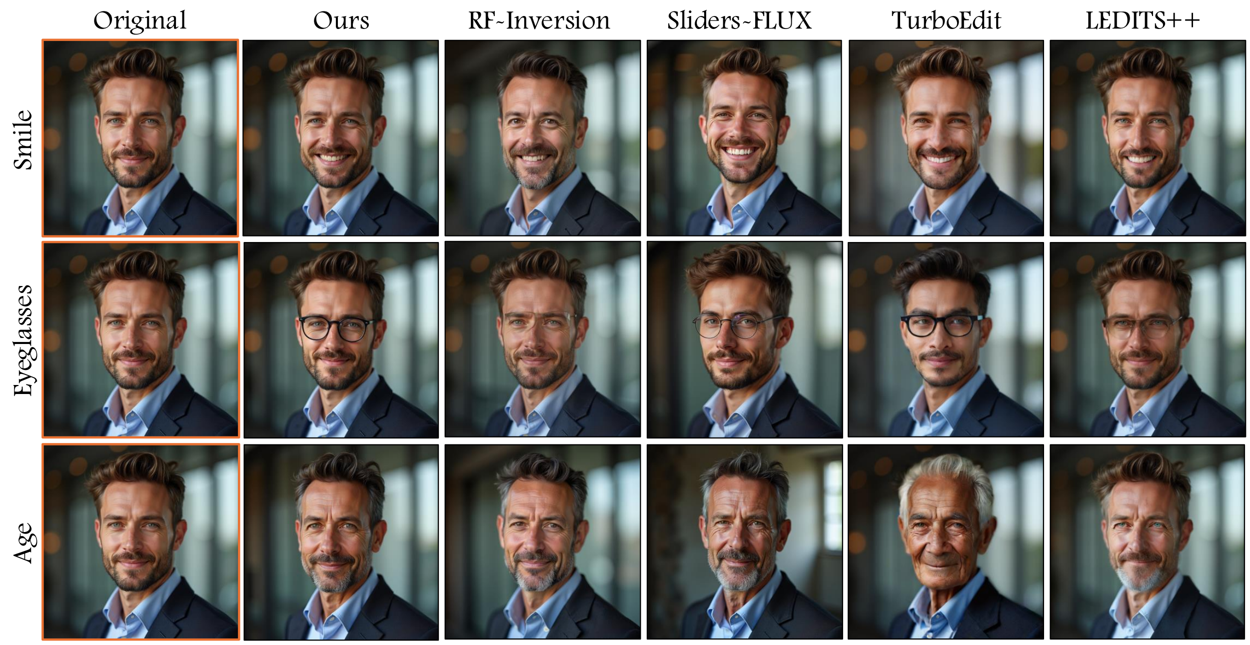

Figure 4. **Qualitative Comparisons.** We compare our method both with latent diffusion-based approaches (LEDITS++ [7] and TurboEdit [12]) and flow-based methods (Sliders-FLUX [17] and RF-Inversion [38]) in terms of their disentangled editing capabilities. We present qualitative results for smile, eyeglasses, and age edits where our method succeeds over competing methods in both reflecting the semantic and preserving the input identity.

$$\hat{c}_{pool} = (1 - \lambda_{coarse}) * c_{pool} + \lambda_{coarse} * c'_{e,pool} \tag{10}$$

During generation, to preserve the image content and influence the image with the desired edit, we use different modulations for text and image features. Specifically, we normalize the image and text features inputted to the attention layer with c_{pool} and $\hat{c}_{pool}$, respectively.

5. Experiments

To assess the effectiveness of FluxSpace in disentangled semantic editing, we conducted a series of qualitative and quantitative experiments. Furthermore, we compared our approach with the state-of-the-art image editing methods in both flow-based and latent diffusion-based models to highlight our method's versatility and superior performance.

5.1. Experimental Setup

We utilize `FLUX.1-dev`[1] for all experiments, evaluating both the capabilities of the proposed framework and the control parameters introduced. For all of our experiments, we use 30 generation steps and $\lambda_{coarse} = 0.5$ and $\tau_m = 0.5$ unless stated otherwise. Since the required editing strength changes for every editing task, we use varying values of λ_{fine} and starting timestep for every experiment. We provide the complete list of used hyperparameters in the supplementary material, Sec. C. To guarantee the reproducibility of our experiments, we maintained a consistent random seed of 0. Our method requires approximately 20 seconds to generate an image with the desired edit on a single NVIDIA L40 GPU.

[1] https://huggingface.co/black-forest-labs/FLUX.1-dev

5.2. Qualitative Results

Fig. 3 and 4 show the qualitative results of our editing experiments. These results demonstrate our method's capability to execute disentangled edits on fine-grained attributes, such as adding eyeglasses or sunglasses, as well as beards, smiles, and detailed facial expressions such as surprise (see Fig. 3, top row). Additionally, our method can conduct broader transformations, including altering a person's perceived age or gender and applying stylistic changes like converting images to comic or 3D cartoon styles (see Fig. 3, bottom row). Beyond faces, our method handles broader image contexts, where we provide examples in the supplementary material, Sec. E. Our approach also extends to other domains, including cars and complex scenes such as street views (see Fig. 1).

Furthermore, we conducted a qualitative comparison with several state-of-the-art methods, including approaches

	Eyeglasses			Smile			Overall
	CLIP-T ↑	CLIP-I ↑	DINO ↑	CLIP-T ↑	CLIP-I ↑	DINO ↑	User Pref. ↑
LEDITS++ [7]	**0.3723**	0.8639	0.9177	0.3634	0.8657	0.9111	3.04
TurboEdit [12]	0.3626	0.8315	0.8954	**0.3692**	0.8479	0.9082	2.75
Sliders-FLUX [17]	0.2878	0.8356	0.7858	0.3618	0.8268	0.8683	3.34
RF-Inversion [38]	0.3046	0.9354	0.9359	0.3438	0.8834	0.9134	2.91
Ours	0.3180	**0.9417**	**0.9402**	0.3503	**0.9038**	**0.9347**	**4.19**

Table 1. **Quantitative Results.** We quantitatively measure the editing performance of our method over competing approaches both in terms of text alignment using CLIP-T [35], and content preservation using CLIP-I [35] and DINO [8] metrics where higher is better for all metrics. We compare our method with both latent diffusion [7, 12], and flow-matching-based approaches [17, 38]. Overall, our method strikes a good balance in terms of alignment with the editing prompt and content preservation. Supplementary to these metrics, we also present a user study as a perceptual evaluation that aligns with our claims regarding edit performance, where our method outperforms the competing approaches.

based on latent diffusion models like LEDITS++ [7] (with SDXL [34]), and TurboEdit [12] (with SDXL-Turbo [39]), as well as flow-based methods such as Sliders-FLUX [17] and RF-Inversion [38]. Fig. 4 illustrates that our method consistently achieves disentangled edits, such as adding eyeglasses or smiles without altering unrelated features, while other methods often struggle with this aspect. For instance, RF-Inversion incorrectly positions the eyeglasses, placing them in unnatural locations on the face. Similarly, while performing edits, Concept Sliders (Sliders-FLUX) and TurboEdit tend to significantly alter the subjects' identities, as seen in the eyeglass and age transformations. Although LEDITS++ manages to execute reasonable smile and eyeglass edits, it inadvertently changes the subject's identity in age-related edits. These results highlight the ability of our method to perform disentangled edits and maintain the identity of the subject while applying targeted edits. More comparisons with other editing methods are provided in the supplementary material, Sec. E.

5.3. Real Image Editing

To be able to apply *FluxSpace* to real images, we integrate our method into the inversion mechanism proposed by RF-Inversion [38] and replace their editing with our editing algorithm. We provide real image editing results in Fig. 5. However, we observe that, due to the nature of their inversion algorithm, the input image is not fully mapped into the latent distribution of the generator model (e.g. Flux), but rather consistency is enabled with a correction term. Since we focus on the interpretability of the latent space of rectified flow transformers, we focus on the models' output distribution and perform our experiments with images generated by Flux.

5.4. Quantitative Results

To further validate the effectiveness of our method, we performed quantitative experiments comparing it with state-of-the-art editing techniques, using Eyeglass and Smile semantics on a set of 60 images (see Table 1). We used metrics such as CLIP-T, CLIP-I [35], and DINO [8] for evaluation. CLIP-T assesses textual alignment by measuring the semantic similarity between the target description and the generated image, ensuring that the intended edits align with textual descriptors. CLIP-I and DINO evaluate image fidelity by measuring the visual similarity between the original and edited images, focusing on maintaining the identity and context of the original subjects.

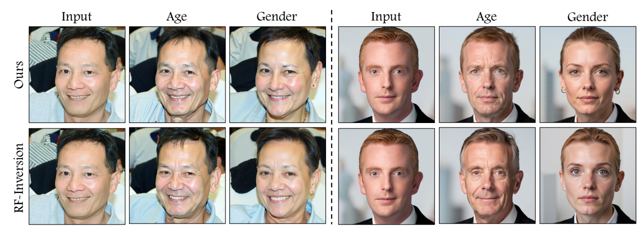

Figure 5. **Real Image Editing.** By integrating *FluxSpace* on the inversion approach proposed by RF-Inversion [38], we extend our method for real image editing task. As we show qualitatively, our method achieves improved disentanglement over the performed edits compared to the baseline approach, where we use identical hyperparameters for the inversion task on both approaches.

The results indicate that our method excels in the CLIP-I and DINO metrics, showcasing its superior ability to retain the original identity while applying disentangled edits. Although methods such as TurboEdit are able to achieve higher CLIP-T scores, they do so at the cost of significant alterations to the subject's original identity. For instance, TurboEdit, despite its high CLIP-T score for eyeglasses, completely changes the individual's identity, as evidenced in Figure 4. This underscores our method's strength in balancing identity preservation with accurate semantic edits.

User Study. We conducted a user study involving 50 participants recruited through the Prolific.com crowd-sourcing platform, in which we provide additional details in Sec. B of the supplementary material. Each participant was presented with an original image alongside its edited version and the corresponding target text. They were then asked to assess how well the edited image captured the intended semantics while maintaining the original facial characteristics of the subject. This was rated on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). The results in Table 1 demonstrate that our method significantly outperforms competing approaches.

5.5. Ablation Studies

We perform ablation studies to demonstrate the effect of hyperparameters introduced by our method. Overall, we present ablations on fine-grained editing scale λ_{fine}, the coarse editing scale λ_{coarse}, the masking coefficient τ_m, and the timesteps that the edit is applied in Fig. 6.

Ablations on coarse editing scale. We introduce the hyperparameter λ_{coarse} to control the text embedding that determines the modulation parameters of the attention layer inputs [15]. In our experiments, we identified that this representation is heavily involved in the overall appearance of the image. In Fig. 6(a), we demonstrate the effect of λ_{coarse} for controlling the pooled embedding for this modulation, where we use "portrait photo of a man" as the generation prompt, and "comics style" as the editing prompt. Upon interpolation within the scale [0.0, 1.0], we see that λ_{coarse} enables controlling the style of the image, given the editing prompt. As we would like to identify the effect of λ_{coarse} in isolation, we set $\lambda_{fine} = 0$ and $\tau_m = 0$ for all generations.

Ablations on fine-grained editing scale. We investigate the flexibility of content editing over the parameter λ_{fine}. Since the text embedding utilized in feature modulation only provides an approximate overview of the image to be generated, we associate λ_{fine} with fine-grained editing, which determines the scale of the edit applied to the attention outputs. We provide data on the effect of λ_{fine} in Fig. 6(b), where we show semantic interpolation results on the attributes of gender and smile, with increasing values of the editing scale.

Ablations on masking coefficient. We introduce an optional attention masking mechanism to improve the disentanglement properties of the editing process. To demonstrate the effect of our masking strategy, we provide ablations in Fig. 6(c), where the edit "sunglasses" is applied starting from the second denoising step, over a 30-step generation process. As can be observed from the presented results, our masking strategy improves the disentanglement of the edit and preserves the semantic properties of the original image better (e.g. facial expression).

Ablations on Editing Timesteps. Since our method allows the selection of editing timesteps, we perform an ablation over the effectiveness of the selection of timesteps where the edit is performed. As can also be observed from Fig. 6(d), starting the editing process in earlier timesteps allows more drastic editing results, by sacrificing consistency with the original image. This is because the overall structure is still considered noisy from the perspective of the denoising model ϵ_θ. However, our method is still able to perform disentangled editing starting from early timesteps (e.g. skipping only 1-2 denoising steps).

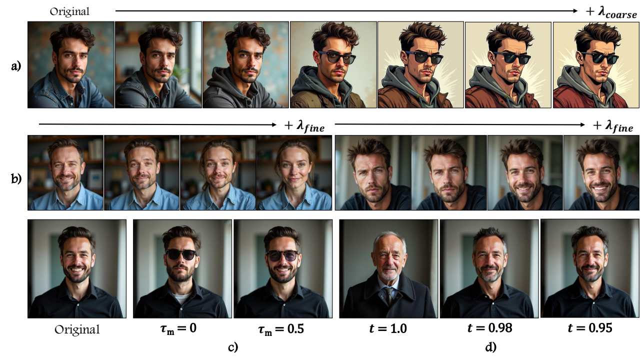

Figure 6. **Ablation Study.** We present ablations over the hyperparameters introduced within the FluxSpace framework. Specifically, we perform ablations on coarse editing scale λ_{coarse}, fine-grained editing scale λ_{fine}, masking coefficient τ_m and timestep t when the editing is initiated. For all ablations, we report qualitative results for changing values of the specified hyperparameters.

6. Limitation

While FluxSpace offers significant benefits in controllable content generation and editing, it also poses ethical risks. Its precise image manipulation can raise privacy concerns when individuals' likenesses are altered without consent [23]. Moreover, the potential for misleading or deceptive content threatens the authenticity of digital media. Realistic edits could enable fake news, impersonation, or public opinion manipulation. Such risks underscore the need for robust ethical guidelines and regulatory frameworks to ensure responsible use.

7. Conclusion

In conclusion, we propose FluxSpace, a novel method that demonstrates a robust ability to perform targeted, disentangled edits across a range of attributes and styles while maintaining the original identity of subjects in images. The qualitative and quantitative results, supplemented by a user study, underscore its effectiveness in preserving the characteristics of the edited image and achieving intended semantic changes, outperforming several state-of-the-art methods. Moreover, FluxSpace excels in performing fine-grained edits across various domains, along with stylistic transformations, showcasing its versatility and wide applicative potential in the field of image editing.

References

[1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. 5

[2] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018. 5

[3] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. *arXiv preprint arXiv:2403.17064*, 2024. 2, 3

[4] David Beniaguev. Synthetic faces high quality (sfhq) dataset, 2022. 13

[5] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (spliCE). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5

[6] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*, 2023. 3

[7] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 3, 6, 7, 11, 12

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 7

[9] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. *arXiv preprint arXiv:2312.05390*, 2023. 2

[10] Yusuf Dalva, Hamza Pehlivan, Oyku Irmak Hatipoglu, Cansu Moran, and Aysegul Dundar. Image-to-image translation with disentangled latent vectors for face editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[11] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Gantastic: Gan-based transfer of interpretable directions for disentangled image editing in text-to-image diffusion models. *arXiv preprint arXiv:2403.19645*, 2024. 3

[12] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models, 2024. 6, 7, 11, 12

[13] 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. *arXiv preprint arXiv:2406.09413*, 2024. 2

[14] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. 2023. 5

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3, 8, 11

[16] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004*, 2015. 5

[17] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 2, 3, 6, 7, 11

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 13, 14

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[22] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 895–904, 2022. 2

[23] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 8

[24] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[26] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[27] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013. 5

[28] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real

images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 13, 14

[29] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 5

[30] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. 5

[31] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023. 2

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 2

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 11

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3, 5, 7, 11

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 7

[36] 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. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 11

[38] L Rout, Y Chen, N Ruiz, C Caramanis, S Shakkottai, and W Chu. Semantic image inversion and editing using rectified stochastic differential equations. 2024. 3, 6, 7, 11, 13, 14

[39] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 7, 12

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 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*, pages 1921–1930, 2023. 3, 13, 14

[41] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[42] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023. 2

[43] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12863–12872, 2021. 1

[44] Hidir Yesiltepe, Yusuf Dalva, and Pinar Yanardag. The curious case of end token: A zero-shot disentangled image editing using clip. *arXiv preprint arXiv:2406.00457*, 2024. 3

[45] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14263–14272, 2021. 1, 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3