UNLOCKING TRAINING-FREE CONTINUOUS ATTRIBUTE CONTROL IN TEXT-TO-IMAGE DIFFUSION TRANSFORMERS VIA M-SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion transformers (DiTs) have shown impressive capabilities in generating high-quality images, yet lack straightforward mechanisms to continuously adjust the strength of individual semantic attributes during image synthesis. To achieve such control, existing approaches often rely on training additional low-rank control adapters or optimizing semantic directions in textual embedding space, incurring extra computational costs. In this paper, we investigate whether the internal conditioning representations of pre-trained text-to-image DiTs can be directly used for continuous attribute control. Our analysis identifies a compact global modulation conditioning space, termed M-space, in which a shared conditioning representation is used to produce layer-wise modulation parameters in text-to-image DiTs. Attribute directions in M-space can be obtained by simply averaging representation differences across diverse pairs of text prompts that differ only in the target attribute. Linear traversal along these directions enables smooth and bidirectional attribute control, while their composition enables multiple attributes to be manipulated simultaneously. Furthermore, we introduce Self-Attention Map Blending (SAMB) and Attention-Guided Modulation Routing (AGMR) to retain attribute-irrelevant content during editing and enable instance-level attribute control, respectively. Extensive qualitative and quantitative experiments on text-to-image DiTs demonstrate that, without requiring any additional training, our method achieves performance comparable to existing state-of-the-art approaches in continuous attribute control.

1 INTRODUCTION

In recent years, diffusion transformers (Peebles & Xie, 2023; Chen et al., 2024) (DiTs) have revolutionized the network architecture for text-to-image diffusion models, substantially improving image synthesis quality and text prompt adherence (Esser et al., 2024). Despite the flexibility of textual conditioning, it remains challenging for existing text-to-image DiT models (Baumann et al., 2025; Chiu et al., 2026) to continuously control semantic attributes in the synthesized image. For instance, a user may wish to smoothly adjust a person's apparent age or smile intensity, or modulate global visual characteristics such as stylization degrees, yet such control cannot be readily achieved by merely modifying the input text prompt.

To address this limitation, existing methods typically resort to additional training to learn control directions, either in the model parameter space or in the textual embedding space. For example, Concept Sliders (Gandikota et al., 2024) and Text Slider (Chiu et al., 2026) learn per-attribute LoRA (Hu et al., 2021) adapters in the diffusion backbone and in the text encoder, respectively. Continuous Attribute Control (Baumann et al., 2025), by contrast, optimizes attribute directions in the textual embedding space and enables subject-level attribute control, yet it still requires optimizing a separate semantic direction for each attribute. Another line, represented by Kontinuous Kontext (Parihar et al., 2026) and SliderEdit (Zarei et al., 2026), trains control adapters based on pre-trained image editing models to modulate edit strength, achieving strong generalization across diverse edits. However, these methods cannot be directly transferred to native text-to-image DiT models for continuous semantic attribute control. Notably, Continuous Attribute Control (Baumann et al., 2025) also offers a training-free approach that identifies semantic directions in the textual embedding space by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

contrasting prompts with and without an attribute prefix. While the method is effective in many scenarios, the experiments report unintended side effects due to semantic entanglement in the textual embedding space (Radford et al., 2021).

In this paper, we investigate whether the internal conditioning representations of text-to-image DiTs can be directly exploited for training-free continuous semantic attribute control. In modern DiTs, text conditions influence generation through both attention and modulation layers. Inspired by the rich semantic structure of StyleGAN's modulation space (Wu et al., 2021) and recent DiT personalization methods (Garibi et al., 2025; Zhong et al., 2026; Chen et al., 2025) that utilize modulation as a condition interface, we systematically study the text-conditioning modulation spaces of pre-trained DiTs. We first examine three candidate modulation spaces along the text-conditioning pathway: the pooled text embedding space (P-space), the global modulation space (M-space), and the per-layer modulation space (M^+-space). Our analysis reveals that M-space is both effective and efficient for continuous attribute control: it is compact yet semantically expressive, and semantic directions in M-space can be simply obtained by averaging representation differences across prompt pairs that differ only in the target attribute. However, linear traversal along these semantic directions in M-space can sometimes perturb attribute-irrelevant content in the source image. To address this, we further introduce Self-Attention Map Blending (SAMB), which preserves attribute-irrelevant content by blending self-attention maps from the source image. Moreover, to enable fine-grained instance-level control, we propose Attention-Guided Modulation Routing (AGMR), which routes distinct modulations to different subject regions identified by cross-modal attention responses. We conduct extensive experiments on representative text-to-image DiTs, showing the effectiveness of our method to achieve smooth and continuous attribute control yet without any additional training.

Our contributions can be summarized as follows:

- To the best of our knowledge, we present the first training-free approach for continuous semantic attribute control in text-to-image DiTs. Through a systematic study of the text-conditioning modulation spaces in pre-trained DiTs, we identify the global modulation space (M-space) as best suited for attribute control, owing to its compactness and strong semantic expressiveness.
- We further introduce SAMB to preserve attribute-irrelevant content during editing, and AGMR to support fine-grained instance-level attribute control.
- Extensive qualitative and quantitative experiments on representative text-to-image DiTs demonstrate that our training-free method achieves performance comparable to existing state-of-the-art approaches for continuous attribute control.

2 Related Works

2.1 Text-to-Image Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2020; Nichol & Dhariwal, 2021) generate images by iteratively denoising from a Gaussian noise to produce a clean image sample. Early text-to-image diffusion models (Rombach et al., 2022; Nichol et al., 2021; Saharia et al., 2022) were primarily built on U-Net architectures (Ronneberger et al., 2015), but recently Diffusion Transformer (DiT) (Peebles & Xie, 2023) architecture trained with flow matching objectives (Lipman et al., 2022) has become the dominant paradigm, offering significant improved generation quality and text-image alignment. Prior U-Net-based models typically inject text conditions into models only as keys and values in the cross-attention layers. In contrast, DiT-based models employ multi-modal attention that processes text and image tokens jointly, while global text conditions are further used to modulate generation through their modulation layers.

2.2 Continuous Attribute Control in Diffusion Models

Existing methods for continuous semantic attribute control primarily learn control directions in either the model parameter space or the textual embedding space. Concept Sliders (Gandikota et al., 2024) distill the score differences from pairs of opposing prompts or sample images into a low-rank parameter direction in the diffusion backbone. After training, linear scaling the LoRA outputs modulate the concept while minimizing interference with other attributes. Text Slider (Chiu et al., 2026)

follows a similar recipe but operates in the parameter space of the text encoder, which learns LoRA adapters optimized by attribute differences in the textual embedding space. Unlike two previous mentioned methods, Continuous Attribute Control (Baumann et al., 2025) shows that token-level semantic directions exist within CLIP text embeddings, and introduces two variants: an optimization-free variant and a optimization-based variant. The optimization-free variant obtains semantic directions by contrasting subject token embedding with and without the desired prefix attribute. While the method needs no learning and is effective in most cases, it often introduces unintended changes to attribute-irrelevant content, and is limited to attributes expressible as prefixes to the subject noun. On the other hand, the optimization-based variant identifies sample-specific directions corresponding to modulations of the target attribute in the noise prediction space and backpropagate them through the diffusion model to discover generalizable, fine-grained local modulation directions within the tokenwise CLIP embedding space. Kontinuous Kontext (Parihar et al., 2026) and SliderEdit (Zarei et al., 2026) build on pretrained instruction-based image editing models to modulate edit strength. Kontinuous Kontext extends an editing model to accept a scalar edit strength, training a lightweight projector that maps the scalar and instruction to coefficients in the model's modulation space, along with a rank-4 LoRA, using synthesized image-edit-instruction-strength quadruplets. SliderEdit disentangles individual instructions from a multi-part prompt and exposes each as a globally trained slider, learning a single set of low-rank adaptation matrices that generalize across diverse edits, attributes, and compositional instructions without per-attribute training. These methods generalize well across diverse and even unseen edits, but they depend on a pretrained editing model and cannot be directly transferred to native text-to-image DiTs for attribute-level control. Unlike these methods, we aim to exploit the modulation space of pretrained text-to-image DiTs for achieving training-free continuous semantic attribute control.

2.3 Modulation Spaces for Image Synthesis in DiTs

Inspired by the observation that StyleGAN's modulation space contains rich, structured, and linearly accessible semantic information (Wu et al., 2021), recent customized image generation works (Garibi et al., 2025; Zhong et al., 2026; Chen et al., 2025) have begun to explore the modulation space of DiT-based text-to-image models as a controllable interface. TokenVerse (Garibi et al., 2025) observes that the modulation space in DiTs is semantic and enables localized control over complex concepts, and proposes an optimization-based framework that finds a distinct direction in the modulation space for each concept token. XVerse (Chen et al., 2025) transforms reference images into offsets for token-specific text-stream modulation, achieving consistent multi-subject control over identity and semantic attributes. Mod-Adapter (Zhong et al., 2026) introduces a modulation adapter that predicts concept-specific modulation directions in a tuning-free manner, leveraging the localized and semantically meaningful properties of the DiT modulation space for both object and abstract concepts. While these works confirm that the DiT modulation space is semantically structured and supports localized control, they focus on concept personalization rather than continuous attribute strength control. Motivated by these observations, we investigate whether the modulation space of a pretrained text-to-image DiT can be directly exploited for training-free continuous semantic attribute control in this paper.

3 Method

As illustrated in Figure 1, our method enables training-free continuous semantic attribute control by operating directly in the modulation space (*M*-space) of a pretrained text-to-image DiT. To further maintain attribute-irrelevant content and enable instance-level control, we introduce Self-Attention Map Blending (SAMB) and Attention-Guided Modulation Routing (AGMR), respectively. We elaborate on the technical details in the following subsections.

3.1 Preliminaries

Modern text-to-image DiTs are typically composed of a stack of multimodal diffusion transformer (MMDiT) blocks. Given a text prompt c, a pre-trained text encoder T produces text tokens $T(c)$, which interact with image tokens through joint attention within each MMDiT block. In parallel, a global pooled text embedding $p(c)$ is transformed by a MLP into a global modulation conditioning vector $m(p)$. $m(p)$ is further combined with the timestep embedding $\tau(t)$ and then is sequentially

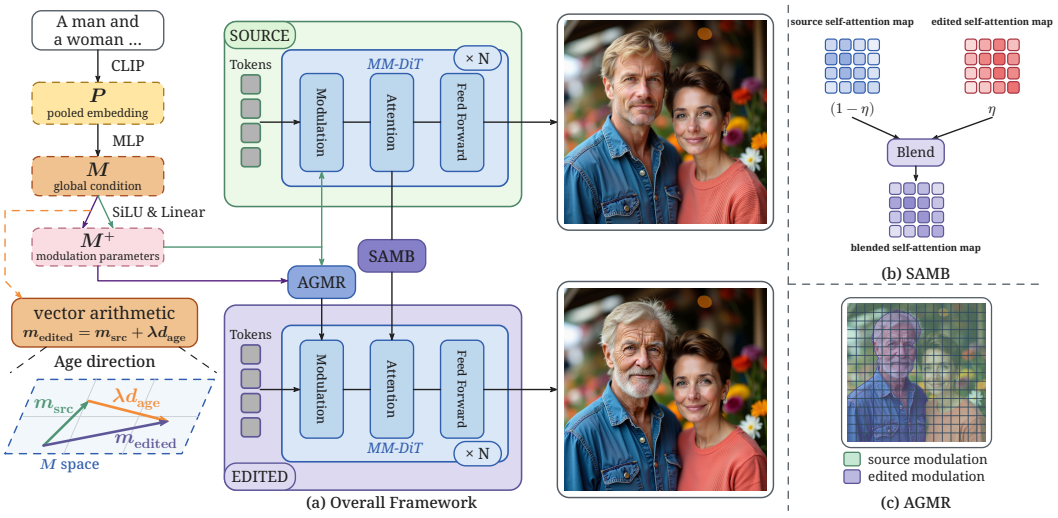


Figure 1: Overview of our method. For simplicity, the timestep input is omitted. (a) A vector addition in M-space changes the global conditioning modulation input to the DiT blocks in the edited branch. A source branch guides the edited branch during generation using Self-Attention Map Blending (SAMB) and Attention-Guided Modulation Routing (AGMR) to preserve source content and enable instance-level attribute control, respectively. (b) SAMB blends source and edited image self-attention maps to preserve attribute-irrelevant content. (c) AGMR routes image tokens between source and the edited modulations according to the cross-modal attention responses of the target subject token.

fed into a non-linear activation $\sigma(\cdot)$ and a linear layer to produce modulation parameters in MMDiT blocks. The process can be defined as follows:

$$\begin{aligned} p(c) = pool(T(c)),\ m(c) = \mathrm{MLP}(e), \\ u(c,t) = m(c) + \tau(t),\ m_\ell^+(c,t) = W_\ell \sigma\left(u\left(c,t\right)\right) + b_\ell, \end{aligned} \tag{1}$$

where $pool(\cdot)$ denotes the global pooling operation, and $m_l^+(c,t)$ is consisted of shift, scale, and gate parameters α_l, β_l, and γ_l for modulating token features h_l in lth MMDiT block:

$$\begin{aligned} \widetilde{h}_l &= (1+\beta_l) \odot \mathrm{LN}(h_l) + \alpha_l, \\ h_l' &= h_l + \gamma_l \odot F(\widetilde{h}_l, \cdot), \end{aligned} \tag{2}$$

where $LN(\cdot)$ and $F(\cdot)$ denote layer normalization (Ba et al., 2016) and a residual operation, respectively. Note that the modulation is applied on both text and image tokens in MMDiT blocks.

The above textual modulation pathway exposes three possible modulation conditioning spaces. For the set of text prompts $\mathcal{Q}$ and the set of MMDiT layers $\mathcal{L}$, we define:

$$\begin{aligned} \mathcal{P} &= \{p(q); q \in \mathcal{Q}\}, \\ \mathcal{M} &= \{m(q); q \in \mathcal{Q}\}, \\ \mathcal{M}^+ &= \{m_\ell^+(q,t); q \in \mathcal{Q}, \ell \in \mathcal{L}\}, \end{aligned} \tag{3}$$

where $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{M}^+$ are termed as the P-space, M-space, and M^+-space in this paper.

3.2 Identifying Attribute Directions in Modulation Spaces

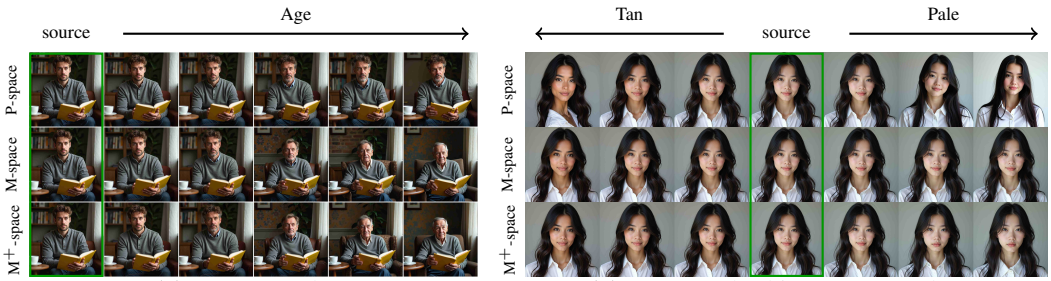
Following prior works that assume semantic attributes are encoded as linear directions in high-dimensional representation spaces (Härkönen et al., 2020; Baumann et al., 2025), we compute the attribute directions in the modulation spaces by averaging the differences across prompt pairs that differ only in the target attribute. Specifically, for a selected attribute A_i, we construct N text prompt pairs $\{(q_i^-, q_i^+)\}_{i=1}^N$ by only changing the attribute phrase while keeping the remaining description fixed within each pair (*e.g.,* "a young woman" as q_i^- versus "a elderly woman" as q_i^+). Given

a source image generated by a text prompt q, the attribute direction and edited feature vectors in *P*-space, *M*-space, and M^+-space then can be obtained as follows:

$$d^P_{A_i} = \frac{1}{N}\sum_{i=1}^{N}\left[p(q_i^+) - p(q_i^-)\right], p_{edited} = p_{src} + \lambda d^P_{A_i}$$
$$d^M_{A_i} = \frac{1}{N}\sum_{i=1}^{N}\left[m(q_i^+) - m(q_i^-)\right], m_{edited} = m_{src} + \lambda d^M_{A_i} \quad (4)$$
$$d^{M^+}_{l,A_i} = \frac{1}{N}\sum_{i=1}^{N}\left[m^+_\ell(q_i^+, t) - m^+_\ell(q_i^-, t)\right], m^+_{l,edited} = m^+_{l,src} + \lambda d^{M^+}_{l,A_i},$$

where p_{edited}, m_{edited}, $m^+_{l,edited}$ and p_{src}, m_{src}, $m^+_{l,src}$ are the edited and the source feature vectors produced by the input text prompt q in the modulation spaces, respective, and λ denotes the edited attribute strength, where positive values strengthen the attribute while negative values weaken it. We only modulate the image tokens in MMDiT blocks as we find modulating text tokens incurs no attribute changes (refer to Figure 14 in the appendix). For modulating multiple attributes simultaneously, users can combine their attribute directions: $d_{all} = \sum_j \lambda_j d^M_j$.

Figure 2 compares continuous attribute control performance in the predefined three modulation spaces. As depicted, on the one hand, editing in *M*-space and M^+-space enables much stronger attribute changes than in *P*-space, indicating more effective attribute control. On the other hand, across the displayed examples, *M*-space and M^+-space produce nearly identical visual results. However, a single attribute direction suffices for editing in *M*-space, whereas M^+-space requires computing attribute directions for every MMDiT block. Therefore, we adopt *M*-space for continuous attribute control in this paper, as it offers comparable control performance to M^+-space with a more compact direction representation.



(a) Age control **(b)** Tan-to-pale skin-tone control

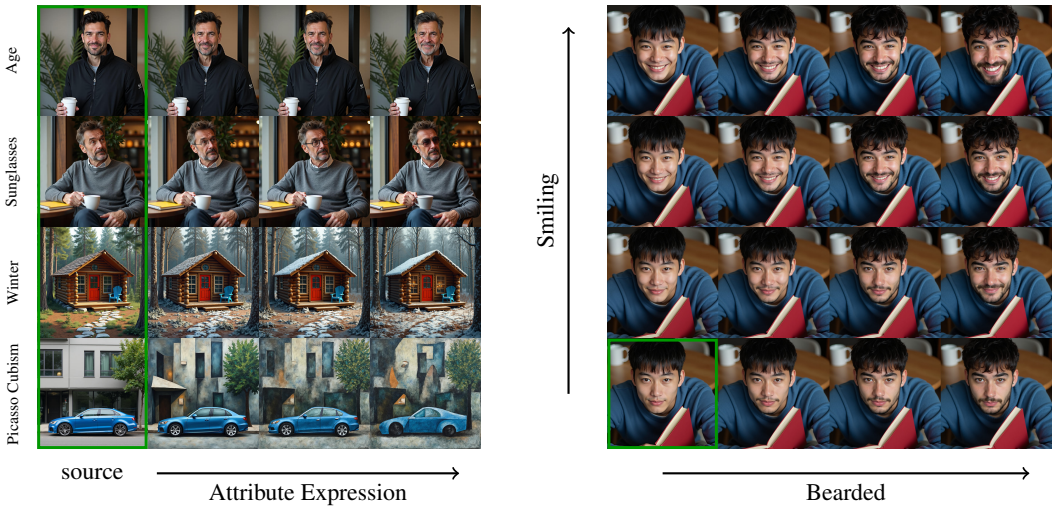
Figure 2: Comparison of visual results by editing in *P*-space, *M*-space, and M^+-space (top to bottom). (a) The rightmost end represents the maximum effective control strength. (b) Both the leftmost and rightmost ends correspond to the maximum effective control strength. In summary, editing in *M*-space and M^+-space produce nearly identical synthesized images, while exhibiting much strong attribute changes than in *P*-space.

3.3 Self-Attention Map Blending

As observed in Figure 2 (a), editing in *M*-space can sometimes cause attribute-irrelevant content changes such as the changes of the poses. To preserve the source image's content, we introduce a Self-Attention Map Blending. Specifically, we run an edit generation branch in parallel to a source generation branch as shown in Figure 1, where both branches use the same prompt and initial latent. At each selected timestep, we replace the edit branch's image self-attention map with a blended version between those from the edit and source branches. Let $A^{\text{src}}_{\ell,t}$ and $A^{\text{edit}}_{\ell,t}$ denote the image self-attention map of the source and the edit branch, respectively, the blended image self-attention map $\widetilde{A}^{\text{edit}}_{\ell,t}$ can be obtained as follows:

$$\widetilde{A}^{\text{edit}}_{\ell,t} = \eta A^{\text{edit}}_{\ell,t} + (1-\eta)A^{\text{src}}_{\ell,t}, \quad (5)$$

where $\eta \in [0, 1]$ controls the blending strength.



(a) Single-attribute control **(b)** Multi-attribute control

Figure 3: Single- and multi-attribute control. (a) Single-attribute control regarding apparent age, wearing sunglasses, summer2winter, and Picasso cubism style (top to bottom), with increasing strength from left to right. (b) Multi-attribute control. The green box marks the source image, while moving rightward increases beard strength and moving upward increases smile intensity.

3.4 Attention-Guided Modulation Routing

To enable instance-level attribute control, we further introduce Attention-Guided Modulation Routing (AGMR), which derives a target-subject mask from the source branch's cross-modal attention map activations and routes the attribute-controlled modulation exclusively to that region.

Given a text token T_{target} identifying the target subject, we aggregate the attention responses of all image tokens to it across multiple attention heads to obtain a spatial mask M (please refer to the appendix for more details). AGMR then routes the i-th image token between the source modulation and the edited modulation according to M:

$$m(i) = \begin{cases} m_{\text{edited}}, & M(i) = 1, \\ m_{\text{src}}, & M(i) = 0, \end{cases} \tag{6}$$

where $m(i)$ denotes the global conditioning modulation feature vector used for i-th image token. More details on AGMR are provided in Appendix A.

4 Experiments

4.1 Experimental Settings.

Evaluation Details. We conduct the main experiments on FLUX.1-dev (Labs, 2024) and additionally apply our method to Stable Diffusion 3.5 Large (SD3.5) (Stability AI, 2024) to examine whether *M*-space control generalizes across DiT architectures. Unless otherwise specified, in our experiments, the images are generated at 1024×1024 resolution with 28 sampling steps and a guidance scale of 3.5. We use *M*-space modulation edit between sampling step 3-27, while SAMB is applied at sampling step 3-27. We evaluate on a diverse set of attributes, including age, expression, hairstyle, accessories, and style. Because control coefficients can not reflect the same control strength for different methods, for fair qualitative and quantitative comparison, we first determine the maximum effective control strength for each method and then sample synthesized images using control strengths uniformly sampled between zero and this maximum. For quantitative metrics, we test each attributes in both directions using 8 discrete nonzero control strengths. We evaluate a diverse set of attributes, including age, expression, hairstyle, accessories, style, and lighting. The evaluation prompts cover diverse subject and scene characteristics, including gender, ethnicity, and age. For each method, quantitative metrics are computed on more than 1,500 edited images.

Baselines and Metrics. We compare against three continuous attribute-control methods, Concept Sliders (CS) (Gandikota et al., 2024), Continuous Attribute Control (CAC) (Baumann et al., 2025), and Text Slider (TS) (Chiu et al., 2026), as well as two continuous image-editing methods, Kontinuous Kontext (KK) (Parihar et al., 2026) and SliderEdit (SE) (Zarei et al., 2026). For CS and CAC, which were originally developed on earlier Stable Diffusion backbones, we re-implement their methods on FLUX.1-dev for a fair comparison. TS transfers zero-shot to FLUX.1-dev through the shared CLIP text encoder. KK and SE are evaluated using their default settings.

Following SliderEdit (Zarei et al., 2026), we evaluate continuous control along three dimensions: extrapolation, continuity, and disentanglement. Extrapolation measures the strongest target-attribute response along a control trajectory, continuity measures how smoothly the target attribute changes with control strength, and disentanglement measures how well attribute-irrelevant content is preserved. For extrapolation and continuity, higher scores are better, whereas for disentanglement, lower scores indicate fewer unintended changes. Note that disentanglement includes ID, a face identity metric computed using ArcFace (Deng et al., 2019). Further details on the metrics and their computation are provided in Appendix B. Quantitative results are summarized in Table 1.

4.2 Qualitative Results

Figure 3 shows that our method enables smooth and continuous control over diverse high-level semantic attributes, while multiple directions can be composed for multi-attribute control.

Figure 4 demonstrates subject-specific control enabled by AGMR. The beard of the man and the apparent age of the woman can be adjusted independently within the same generation.

Figure 5 provides a qualitative comparison with existing methods. Our method produces smoother attribute transitions while largely preserving non-target subject characteristics. In contrast, several baselines introduce noticeable changes in pose, facial shape, and other unrelated characteristics when modifying the target attribute.

More qualitative results can be found in the Appendix C. Our method also generalizes to SD3.5; detailed results are provided in Appendix D.

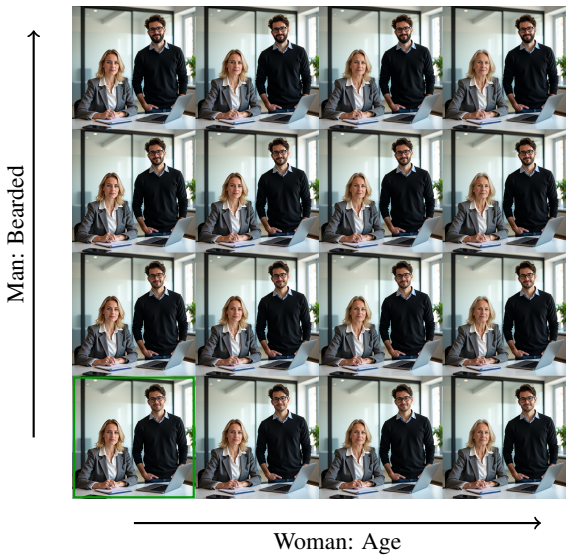


Figure 4: Instance-level attribute control. Moving upward strengthens the man's beard, while moving rightward increases the woman's apparent age.

4.3 Quantitative Results

Our method achieves the best extrapolation scores across all three vision–language models, indicating the strongest target-attribute response, while remaining competitive in continuity. It also substantially improves disentanglement, reducing $\text{LPIPS}_{\text{Alex}}$, $\text{LPIPS}_{\text{VGG}}$, and DINO distances by 53.5%, 41.6%, and 11.2% relative to the strongest baselines, respectively. This improvement does not come at the cost of attribute response, since our method simultaneously attains the best extrapolation scores. We attribute this favorable trade-off to M-space control, which operates in a disentangled latent space and thus allows target attributes to be edited while preserving attribute-irrelevant content.

4.4 Real-Image Attribute Editing

Our method extends to real-image attribute editing when combined with inversion. We adopt RF-inversion (Rout et al., 2025) and apply our method during reconstruction trajectories to adjust the target attribute, while the editing performance depends on the inversion fidelity. Figure 6 presents our method's real-image editing results.

Table 1: Quantitative comparison of continuous attribute control. Arrows indicate whether higher (↑) or lower (↓) values are better. Values are averaged over attributes and evaluation prompts. Best and second-best results are shown in bold and underlined, respectively.

	Extrapolation ↑			Continuity ↑			Disentanglement ↓			
Method	CLIP	SigLIP	BLIP	CLIP	SigLIP	BLIP	$\text{LPIPS}_{\text{Alex}}$	$\text{LPIPS}_{\text{VGG}}$	DINO	ID
CS	0.2446	0.0649	0.3628	0.2178	0.2123	0.2197	<u>0.2087</u>	<u>0.2411</u>	0.1046	0.4115
CAC	0.2474	0.0658	<u>0.4389</u>	0.2497	0.2195	0.2368	0.3403	0.3605	0.1928	0.5069
KK	0.2411	0.0709	0.3495	0.1779	0.1781	0.1785	0.2858	0.3324	<u>0.0565</u>	<u>0.2844</u>
SE	<u>0.2490</u>	<u>0.0727</u>	0.4343	**0.2906**	**0.2929**	**0.2418**	0.2594	0.3081	0.0597	**0.2843**
TS	0.2465	0.0666	0.4112	0.2484	<u>0.2264</u>	0.2228	0.2760	0.3034	0.1418	0.4155
Ours	**0.2506**	**0.0758**	**0.4577**	<u>0.2736</u>	0.2117	<u>0.2371</u>	**0.0971**	**0.1407**	**0.0502**	0.3322

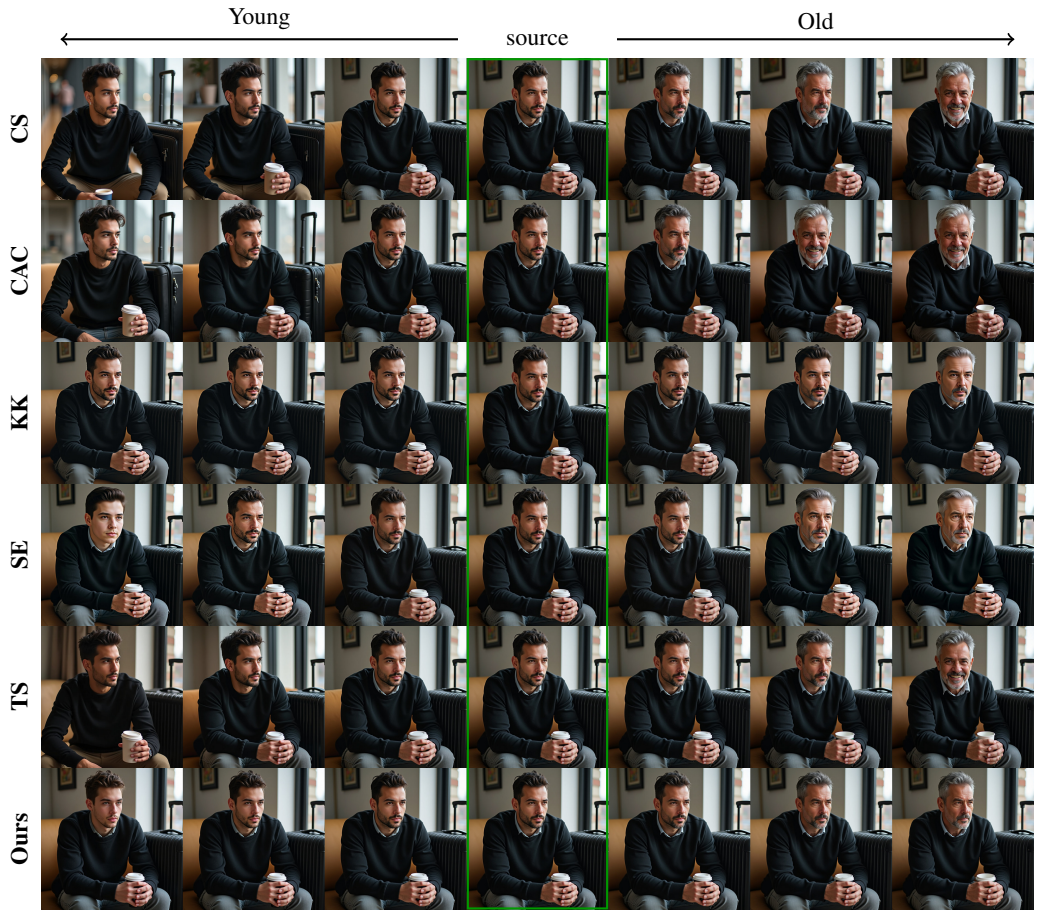


Figure 5: Qualitative comparison of bidirectional apparent age control.

4.5 Ablation Studies

We ablate the SAMB and AGMR from our method to assess their effectiveness for preserving source content and enabling instance-level attribute control, respectively.

Without SAMB. Figure 7 shows that removing SAMB, while slightly improving extrapolation and continuity, causes expression control to induce larger changes in pose, objects, and background and substantially degrades disentanglement, with $\text{LPIPS}_{\text{Alex}}$, $\text{LPIPS}_{\text{VGG}}$, DINO, and identity dis-



Figure 6: Real-image age editing.



Figure 7: Ablation of SAMB .

Table 2: Quantitative ablation of SAMB. Green indicates performance degradation after removing SAMB, while red indicates performance improvement. Best results are in bold.

	Extrapolation ↑			Continuity ↑			Disentanglement ↓			
Variant	CLIP	SigLIP	BLIP	CLIP	SigLIP	BLIP	$\text{LPIPS}_{\text{Alex}}$	$\text{LPIPS}_{\text{VGG}}$	DINO	ID
Ours	0.2506	0.0758	0.4577	**0.2736**	0.2117	0.2371	**0.0971**	**0.1407**	**0.0502**	**0.3322**
w/o SAMB	**0.2622**	**0.0788**	**0.5169**	0.2735	**0.2207**	**0.2524**	0.2706	0.3066	0.1600	0.5035
Δ (%)	+4.63	+3.96	+12.93	-0.04	+4.25	+6.45	+178.68	+117.91	+218.73	+51.57

tances increasing by 178.68%, 117.91%, 218.73%, and 51.57%, respectively (Table 2), indicating that the fusion module mainly helps preserve attribute-irrelevant content.

Without AGMR. Without AGMR, applying the edited attribute modulation to all image tokens reduces the method to a global control setting and therefore cannot isolate a specific subject in multi-subject scenes. We also consider an alternative baseline that applies the edited attribute modulation only to the text tokens corresponding to the target subject; however, this produces negligible attribute changes. Qualitative comparisons of both variants are provided in Appendix E.

5 Conclusion

In this paper, we present a training-free method for continuous semantic attribute control via M-space in text-to-image DiTs. We investigate several candidate modulation spaces and identify *M*-space for attribute control due to its effectiveness and compactness. By constructing reusable attribute directions in *M*-space, our method enables smooth single- and multi-attribute control without any extra training overheads in the pre-trained DiT models. Further, we introduce SAMB and AGMR to improve source content preservation during editing and enable instance-level control in multi-subject image generation, respectively. Extensive experiments show the effectiveness of our method in achieving comparable continuous attribute control performance against existing state-of-the-art approaches.

AI use statement

Generative AI tools were used solely to assist with language editing and polishing, including improving grammar, clarity, and readability. They were not used to formulate the research ideas, develop the methodology, design or conduct experiments, analyze or interpret the results, or derive mathematical claims. All AI-assisted revisions were carefully reviewed and verified by the authors, who take full responsibility for the final content of this paper.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying

semantic directions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13231–13241. IEEE, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://doi.org/10.1109/ICCV48922.2021.00951.

Bowen Chen, Brynn zhao, Haomiao Sun, Li Chen, Xu Wang, Daniel K. Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/0b0153a91f827b14e8bfea4e211362f3-Abstract-Conference.html.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 57611–57640, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fe989bb038b5dcc44181255dd6913e43-Paper-Conference.pdf.

Pin-Yen Chiu, I Fang, Jun-Cheng Chen, et al. Text slider: Efficient and plug-and-play continuous concept control for image/video synthesis via lora adapters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 613–622, 2026.

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

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4690–4699. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00482. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

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.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, 2024. arXiv:2311.12092.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. doi: 10.48550/ARXIV.2304.07193. URL https://doi.org/10.48550/arXiv.2304.07193.

Rishubh Parihar, Or Patashnik, Daniil Ostashev, Venkatesh Babu Radhakrishnan, Daniel Cohen-Or, and Kuan-Chieh Jackson Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37929–37939, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Hu0FSOSEyS.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Stability AI. Stable Diffusion 3.5: Reference implementation. https://github.com/Stability-AI/sd3.5, 2024. GitHub repository.

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 (CVPR)*, pp. 12863–12872, June 2021.

Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan A Rossi, and Soheil Feizi. Slideredit: Continuous image editing with fine-grained instruction control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14430–14439, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. In *International Conference on Learning Representations*, volume 2026, pp. 5922–5940, 2026.

A Implementation Details of Attention-Guided Modulation Routing

This section provides the implementation details of the spatial mask used by Attention-Guided Modulation Routing (AGMR). The mask is estimated online from the source branch's image-to-text attention and requires neither an external segmentation model nor manual annotations.

Target text tokens. Given a prompt, we first locate the core text tokens corresponding to the target subject using tokenizer offsets. Let

$$\mathcal{T}_{\text{target}}$$

denote the resulting set of target text-token indices.

Attention response. At sampling step t and double-stream block ℓ, we use only the source branch to compute the attention response from image tokens to the target text tokens. Let $A^{(h)}_{\ell,t}(i,j)$ denote the attention weight from image-token query i to text-token key j in head h. The target-subject response for image token i is

$$r_{\ell,t}(i) = \frac{1}{H}\sum_{h=1}^{H}\sum_{j\in\mathcal{T}_{\text{target}}} A^{(h)}_{\ell,t}(i,j), \tag{7}$$

where H is the number of attention heads. The resulting responses are normalized over image tokens to obtain a spatial response map.

Seed and candidate regions. We construct the binary subject mask using two response thresholds. Tokens above the 0.95 response quantile are used as high-confidence seeds,

$$\mathcal{S}_{\ell,t} = \{i \mid r_{\ell,t}(i) \geq Q_{0.95}(r_{\ell,t})\}, \tag{8}$$

while tokens above the 0.80 quantile form the candidate region,

$$\mathcal{C}_{\ell,t} = \{i \mid r_{\ell,t}(i) \geq Q_{0.80}(r_{\ell,t})\}. \tag{9}$$

Here $Q_p(\cdot)$ denotes the p-th quantile of the response values.

Starting from the seed set, we perform 8-neighbor region growing within the candidate region for at most 64 iterations. We then retain the largest connected component as the target-subject region $\mathcal{R}_{\ell,t}$. The final binary mask is

$$M_{\ell,t}(i) = \begin{cases} 1, & i \in \mathcal{R}_{\ell,t}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Masks across blocks. Each double-stream block uses its own mask $M_{\ell,t}$. For the single-stream blocks and the final normalization layer, we average the response maps from the 19 double-stream blocks and apply the same mask-construction procedure to obtain a shared spatial mask.

B Evaluation Metrics

Following SliderEdit (Zarei et al., 2026), we evaluate continuous attribute control using extrapolation, continuity, and disentanglement. For a fixed prompt and initial noise, let I_0 denote the source generation and $\{I_k\}_{k=1}^{K}$ the images generated at different control strengths. For each traversal, we use a target-attribute description r and compute image–text similarity scores $s_k^{(v)}$ with CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and BLIP (Li et al., 2022).

B.1 Extrapolation

Extrapolation measures the strongest target-attribute response reached along the control trajectory:

$$E^{(v)} = \max_{1 \leq k \leq K} s_k^{(v)}. \tag{11}$$

Higher values indicate stronger expression of the target attribute.

B.2 Continuity

Continuity measures whether attribute expression changes smoothly as the control strength varies. Following SliderEdit, the attribute scores are normalized and divided into equal-width bins. Let O_b be the number of scores in bin b and e the expected count under uniform coverage. We compute

$$\chi_v^2 = \sum_{b=1}^{B} \frac{(O_b - e)^2}{e}. \tag{12}$$

The continuity score is based on the inverse reduced chi-squared statistic:

$$C^{(v)} = \left(\frac{\chi_{v,\text{agg}}^2}{\nu_{\text{agg}}} \right)^{-1}. \tag{13}$$

Higher values indicate smoother and more evenly distributed attribute transitions.

B.3 DISENTANGLEMENT

Disentanglement measures how much the controlled images deviate from the source generation I_0. We report LPIPS (Zhang et al., 2018) with AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan & Zisserman, 2015) backbones, DINOv2 (Caron et al., 2021; Oquab et al., 2023) feature distance, and ArcFace (Deng et al., 2019) identity distance:

$$D_d = \frac{1}{|\mathcal{V}_d|} \sum_{k \in \mathcal{V}_d} d(I_k, I_0), \tag{14}$$

where $\mathcal{V}_d$ contains the valid image pairs for metric d. Lower values indicate better preservation of attribute-irrelevant content.

LPIPS measures perceptual differences in pretrained feature space, DINOv2 measures global feature similarity using cosine distance, and ArcFace measures facial-identity preservation. Since intended attribute changes can also increase these distances, we interpret disentanglement together with extrapolation and continuity rather than in isolation.

C ADDITIONAL QUALITATIVE RESULTS

We provide additional single-attribute control results across three prompts for each attribute. For the body-shape and wearing sunglasses results in Fig. 8, the first column is the source generation and the remaining columns increase the control strength toward the indicated attribute. For the remaining bidirectional results, the middle column is the source generation; moving left or right increases the control strength toward the attribute written above the corresponding arrow.



Figure 8: Additional positive-direction traversals for body shape (left) and sunglasses (right).

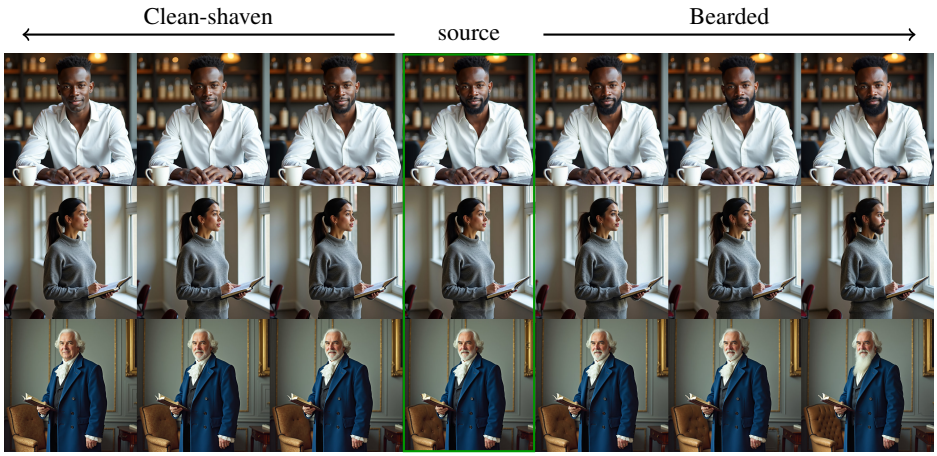


Figure 9: Additional bidirectional control results for facial hair.



Figure 10: Additional bidirectional control results for facial expression.

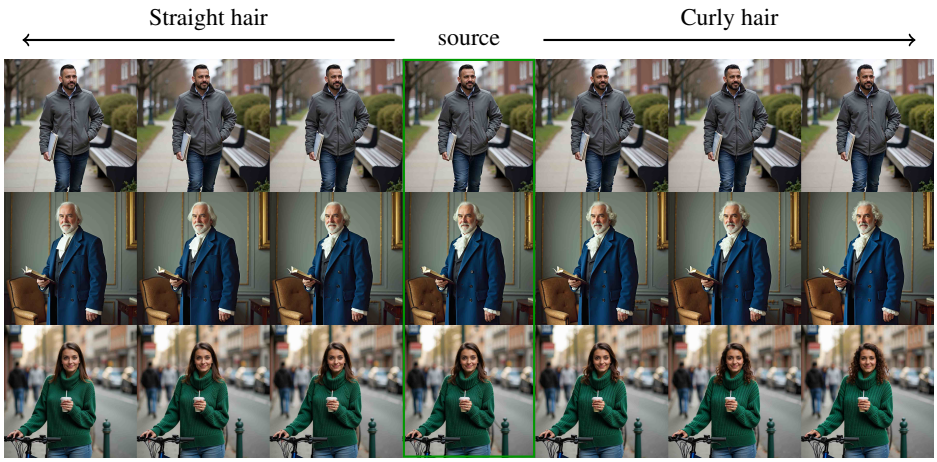


Figure 11: Additional bidirectional control results for hair texture.

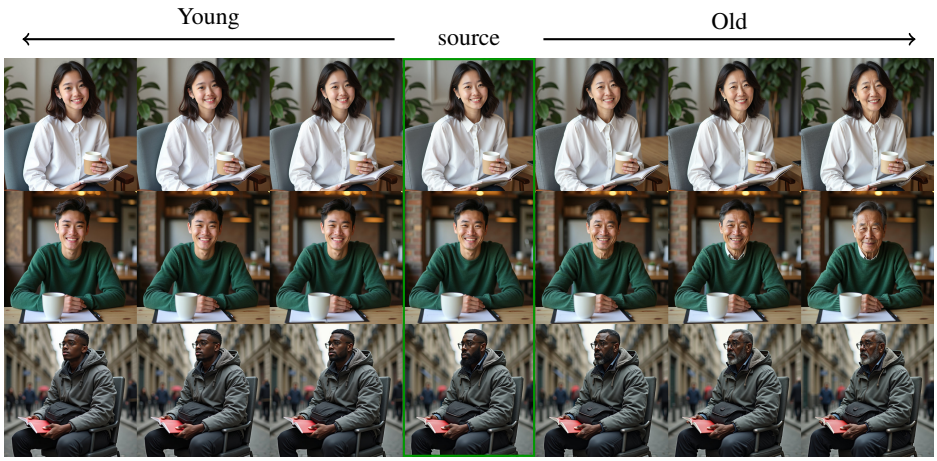


Figure 12: Additional bidirectional control results for age.

D GENERALIZATION TO SD3.5

Our main experiments use FLUX.1-dev, but the proposed control mechanism is not specific to the FLUX architecture. We apply the same training-free direction construction and inference procedure to SD3.5, another DiT-based text-to-image model. As shown in Fig. 13, traversing the resulting direction in opposite directions produces continuous age changes toward younger and older appearances across different prompts, while largely preserving the main subject and scene content. These results suggest that the proposed control mechanism is not limited to FLUX and can also be applied to another DiT architecture with a similar shared conditioning pathway for block modulation.



Figure 13: Continuous age control on SD3.5. Each row uses a different prompt. Moving left or right increases the control strength toward younger or older appearances, respectively.

E TEXT-TOKEN VS. IMAGE-TOKEN MODULATION FOR INSTANCE-LEVEL CONTROL

For instance-level attribute control, the text tokens referring to the target subject provide a natural selection mask within the text sequence. Applying the modulation induced by an attribute direction only to these tokens would avoid estimating a spatial mask from attention maps. We compare this alternative with global control over all tokens and control restricted to the target subject's image tokens selected by our mask, keeping the pretrained model frozen in all three settings.

Figure 14 compares these settings for age control of the man. Applying control to all tokens makes both subjects appear older, showing that global modulation does not isolate the intended subject. Controlling only the text tokens corresponding to `man` produces almost no visible age change. Restricting control to the selected image tokens progressively ages the man while largely preserving the woman's appearance. Among these three settings, only the last achieves effective instance-level control, supporting our use of AGMR to the target subject's image tokens.

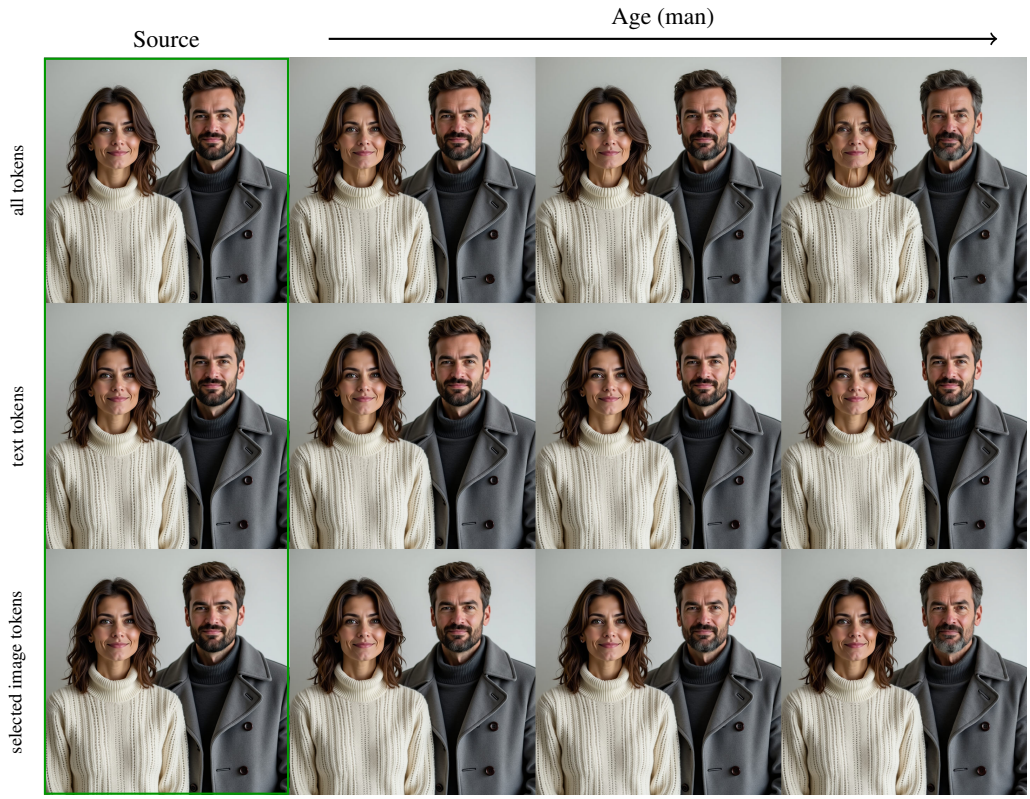


Figure 14: Comparison of token selection for age control of the man. From top to bottom: all tokens, target-subject text tokens, and selected target-subject image tokens. The first column shows the original generation; control strength increases from left to right. All model parameters remain frozen.

F Sampling-Step Ablation for Attribute Control

We vary the sampling-step interval over which the M-space attribute direction is applied to study the effect of intervention timing. The intervals in Table 3 indicate the sampling steps during which the direction is active, with steps 3–27 used as our default setting.

Applying control over all steps (0–27) yields the strongest extrapolation and continuity, but also causes the largest changes to attribute-irrelevant content. Restricting the intervention to steps 3–15 improves disentanglement at the cost of substantially weaker control. The default interval, steps 3–27, retains strong extrapolation while reducing unintended changes compared with full-step control, providing a better balance between attribute strength and content preservation.

Table 3: Ablation of the sampling steps at which attribute control is active. Both endpoints are included; steps 3–27 are our default. Higher extrapolation and continuity and lower disentanglement distances are better. Best and second-best results are shown in bold and underlined, respectively.

	Extrapolation ↑			Continuity ↑			Disentanglement ↓			
Steps	CLIP	SigLIP	BLIP	CLIP	SigLIP	BLIP	$\text{LPIPS}_{\text{Alex}}$	$\text{LPIPS}_{\text{VGG}}$	DINO	ID
0–27	**0.2517**	**0.0782**	**0.5105**	**0.2832**	**0.2324**	**0.3030**	0.1861	0.2256	0.0802	0.4383
3–15	0.2403	0.0568	0.2538	0.2191	0.1695	0.1994	**0.0669**	**0.1035**	**0.0269**	**0.2688**
3–27	<u>0.2506</u>	<u>0.0758</u>	<u>0.4577</u>	<u>0.2736</u>	<u>0.2117</u>	0.2371	0.0971	0.1407	0.0502	0.3322
5–27	0.2443	0.0607	0.3152	0.2602	0.1768	<u>0.2426</u>	<u>0.0765</u>	<u>0.1215</u>	<u>0.0301</u>	<u>0.3118</u>

G Layer Sensitivity to Attribute Directions

We examine which FLUX.1-dev DiT blocks respond most strongly to our M-space attribute directions. SAMB is disabled throughout this analysis to make the attribute-induced changes more visible and facilitate comparison across blocks.

Comparison of block groups. We first apply the attribute direction to all blocks, only the double-stream blocks, or only the single-stream blocks. All-block control generally yields the largest visible target-attribute change, followed by double-stream-only control, while single-stream-only control produces the weakest response. This comparison focuses only on attribute-control strength and does not assess preservation of unrelated content.

Individual-block sensitivity. To examine block-level sensitivity more closely, we evaluate multiple attributes at relatively large control strengths. For each attribute, we apply the direction to one block at a time while keeping the remaining blocks under the source condition. The blocks producing pronounced changes differ across attributes, and no single block is consistently highly sensitive across all tested attributes. We therefore use all-block intervention as a simple and robust default setting in our experiments.

Figure 15 illustrates the comparison of block groups for age control.



Figure 15: Age control using different block groups, without SAMB. From top to bottom, the direction is applied to all blocks, only double-stream blocks, or only single-stream blocks. Control strength toward older appearances increases from left to right.

H TRAINING-FREE CAC ON FLUX.1-DEV

The training-free variant of CAC (Baumann et al., 2025) constructs a control direction directly from text embeddings. It encodes the text with and without an attribute prefix and takes the difference between the resulting representations as the attribute direction, without learning additional parameters. We adapt this construction to FLUX.1-dev to examine its effectiveness for attribute control.

Figure 16 shows results for curly hair, heavyset build, bearded, and sunglasses. The first column contains the source generations, and the remaining columns show the corresponding controlled outputs.

The desired attribute changes are weak or absent in these examples. Hair remains largely straight in the curly-hair traversal, and the body-shape change is limited. In the beard traversal, changes to the hairstyle and the appearance of glasses are more pronounced than the intended beard change. The traversal of the sunglasses changes the subject's appearance and pose without producing sunglasses. These results show that our adaptation of training-free CAC does not provide effective, selective attribute control on FLUX.1-dev in the tested cases: the target attribute responds weakly, while unrelated content can change.

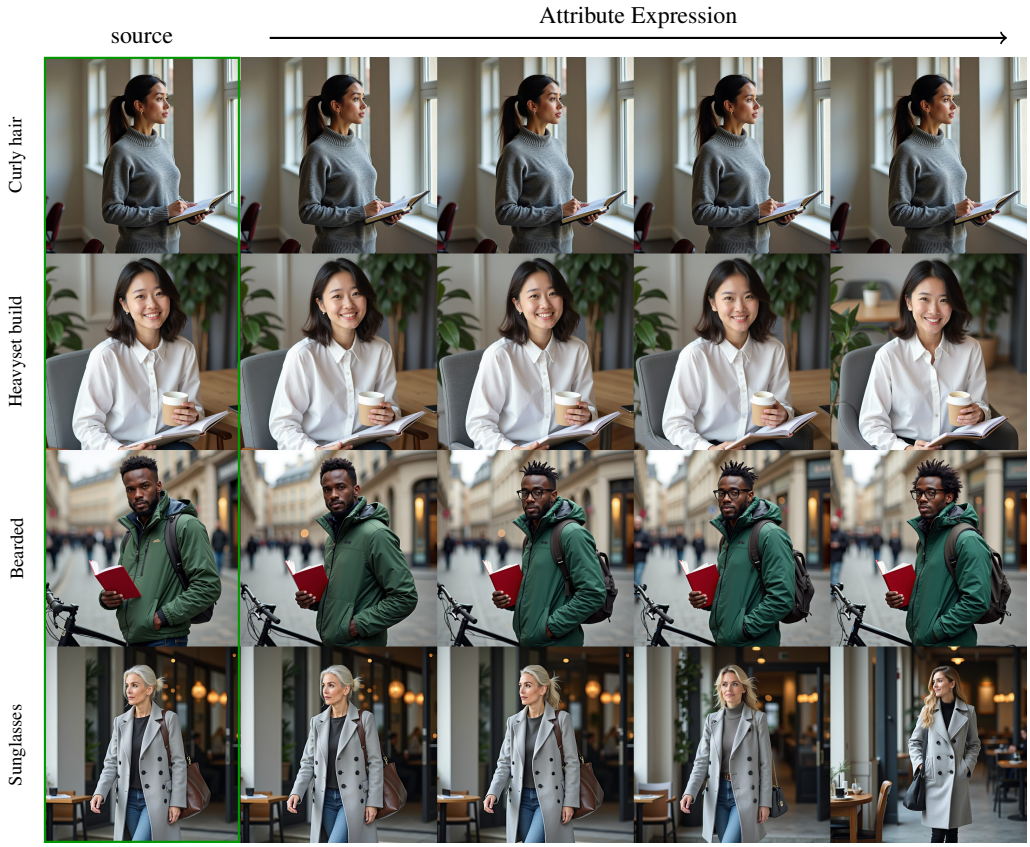


Figure 16: Training-free CAC adapted to FLUX.1-dev. Rows from top to bottom target curly hair, heavyset build, a beard, and sunglasses. The first column shows the source generations. The controlled outputs exhibit weak target-attribute changes and, in some cases, changes to unrelated content.

I ASSESSING CONTINUOUS ATTRIBUTE CONTROL WITH FLUXSPACE

FluxSpace (Dalva et al., 2024) provides separate strength parameters for fine-grained editing of attention outputs and coarse editing of pooled text representations, and demonstrates semantic interpolation through qualitative examples. Using its publicly released implementation on FLUX.1-dev,

we evaluate whether these controls provide gradual attribute changes across a range of attributes, prompts, and hyperparameter settings. We vary either the fine-grained or coarse control strength while keeping the other fixed.

Across most tested settings, the target attribute changes rapidly over a narrow range of control strengths and then shows little further variation. The resulting trajectories therefore tend to exhibit abrupt transitions followed by saturation, with limited coverage of intermediate attribute states.

For this reason, we do not include FluxSpace as a primary baseline in our continuous attribute-control evaluation.

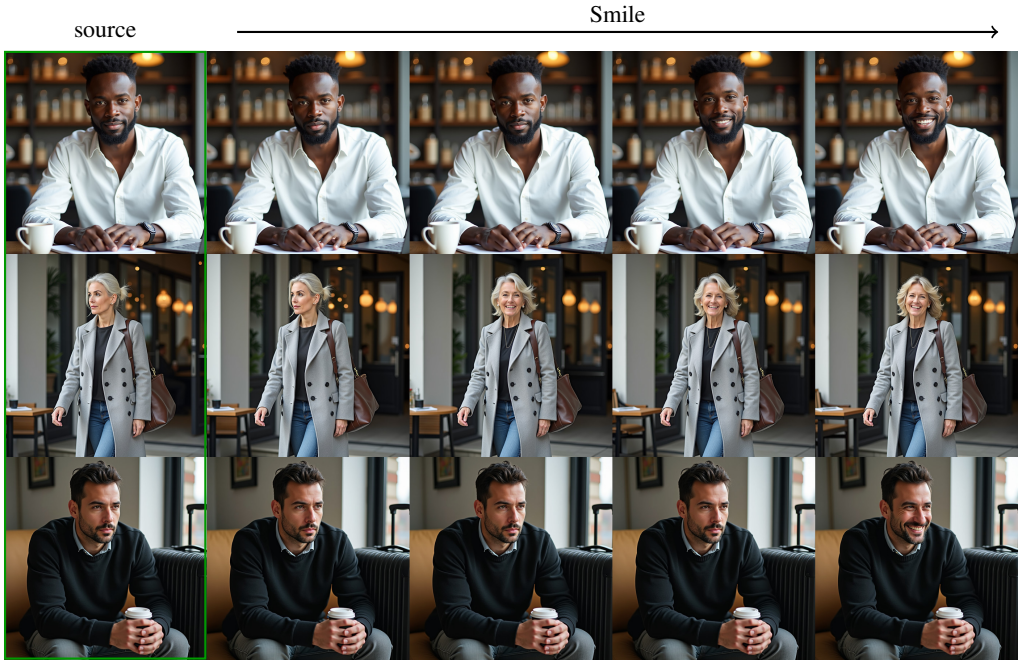


Figure 17: Smile control using the publicly released FluxSpace implementation. The first column is the source generation; the remaining columns show increasing control strengths toward a smile. The results exhibit an abrupt transition toward the target expression over a narrow range of control strengths, after which further increases produce little additional change, indicating a saturation effect.