Supplementary Materials
All-in-One Slider for Attribute Manipulation in Diffusion Models

Weixin Ye[1,2*] Hongguang Zhu[3*] Wei Wang[1,2†] Yahui Liu[4] Mengyu Wang[1,2] Xuecheng Nie[5]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intelligence + X International Cooperation Joint Laboratory of the Ministry of Education
[3]City University of Macau [4]Kuaishou [5]Meitu

1. Implementation Details

Training Details. We use SDXL as the base generative model and extract intermediate text embeddings from its 11th (of 12) and 29th (of 32) transformer layers blocks (`layer.10` and `layer.28`) across the dual encoders to train the Attribute Sparse Autoencoder. The activations are computed over prompts covering 52 attributes, each augmented with diverse identity-aware prefixes to encourage broad semantic coverage. The dimension of text embedding is 2048 (concatenation of 768 and 1280). The autoencoder has a latent dimension of 32,768 with an expansion factor of 16. It is trained using the Adam optimizer with a learning rate of 4×10^{-4} and a batch size of 4096 latent tokens per step. The total training budget is 400 million tokens, corresponding to approximately 97,656 training steps. During attribute selection, we retain the top-$k = 128$ most activated dimensions. To mitigate overfitting and encourage broader attribute coverage, we introduce an auxiliary selection mechanism with $k_{\text{aux}} = 256$ and apply a regularization coefficient of $\alpha = 0.1$ to increase its influence during training. All experiments were performed using a single NVIDIA GeForce RTX 4090 GPU.

Training Prompt Construction. To expose the Attribute Sparse Autoencoder to a wide variety of visual concepts, we construct a diverse set of text prompts associated with over 52 semantic attributes (*e.g.*, *age*, *arched eyebrows*, *wearing necklace*). For each attribute, we manually define several textual expressions that describe different attribute states (*e.g.*, "old man", "very old person", "young man", "very young person").

To enrich visual diversity and reduce identity bias, we prepend each prompt with descriptive prefixes drawn from a fixed pool. These include phrases like "a photo of a", "a photo of a beautiful", as well as identity-related descriptors such as "a photo of a black / white / asian / arab / caucasian", and their compositional forms (*e.g.*, "a photo of a beautiful asian").

We construct the full prompt set by exhaustively combining all prefixes and prompt templates for each attribute. If the total number of combinations is less than 1000, we retain all; otherwise, we randomly sample 1000 combinations using different random seeds to ensure diversity across prompt sets. This results in a balanced and diverse prompt distribution per attribute.

Attribute List. We consider a total of 52 facial attributes and use their corresponding text embeddings to train the Attribute Sparse Autoencoder. These attributes span appearance, expression, hairstyle, and accessory-related features. The full list is as follows: *age, angry, arched-eyebrows, bags-under-eyes, bald, bangs, big-lips, big-nose, birthmarks, black-hair, blond-hair, brown-hair, bushy-eyebrows, chubby, colorful-outfit, double-chin, elegant, eye-size, eyeglasses, fitness, freckled, frowning, gray-hair, groomed, high-cheekbones, makeup, moles, mouth-slightly-open, multicolored-hair, mustache, narrow-eyes, oval-face, pale-skin, pierced, ponytail, pouting, receding-hairline, rosy-cheeks, sad, scarred, sideburns, smile, surprised, tattooed, tiredness, wavy-hair, wearing-earrings, wearing-hat, wearing-lipstick, wearing-necklace, wearing-necktie, width.*

Image Editing Evaluation Criteria. In the evaluation of image edits, we introduce Qwen Score (QS) to assess semantic attribute changes or facial expression. This is done using a 5-point scale across three dimensions: Action / Expression Fidelity, Identity Preservation, and Visual & Anatomical Coherence. This system is based on the scoring framework provided by the Qwen2.5-VL-Instruct model. The criteria are as follows:

Expression Fidelity captures how faithfully the edited image reflects the target expression described in the instruction. A high score reflects a precise match in terms of intensity and orientation. In contrast, lower scores are triggered by absent or incorrectly rendered expressions, partial

*Equal contribution: weixinye@bjtu.edu.cn, hgzhu@cityu.edu.mo
†Corresponding author: wei.wang@bjtu.edu.cn

changes (e.g., only some facial muscles involved), or misaligned pose dynamics.

Identity Preservation evaluates whether the subject remains recognisable and consistent across the original and edited images. Perfect preservation includes consistent facial structure, hairstyle, and clothing. Score degradation occurs when noticeable drifts—such as altered facial features, skin tone, or missing accessories—make the subject appear different or less identifiable.

Visual & Anatomical Coherence focuses on the realism and structural consistency of the rendered image. High scores reflect anatomically correct bodies, natural lighting and shadows, and coherent textures. Lower scores are given in the presence of visual artifacts, anatomical distortions, or rendering inconsistencies such as cut-out edges, lighting mismatch, or unrealistic joint positions.

Each image pair (before and after steering) is evaluated in terms of three dimensions: Expression Fidelity, Identity Preservation, and Visual & Anatomical Coherence, each scored on a 5-point scale. In addition to the numerical scores, a brief reasoning (within 20 words) is provided to justify the evaluation results. As an example, we present the evaluation of an manipulated image, for which the Qwen model assigned scores of 4 for Action Fidelity, 5 for Identity Preservation, and 4 for Visual & Anatomical Coherence (see Figure 1). The manipulated image shows a successful transition from a neutral expression to a clear, natural-looking smile, while preserving consistency in the head pose and facial features. Minor inaccuracies in the smile's angle or intensity lead to an Action Fidelity score of 4. The subject's identity is perfectly preserved, justifying a score of 5 for Identity Preservation. The image also maintains visual and anatomical coherence, with only slight imperfections, resulting in a score of 4 for Visual & Anatomical Coherence. According to Qwen's analysis, the brief explanation for this result is: *"The subject's expression has changed to a smile, but the overall pose remains the same"*.

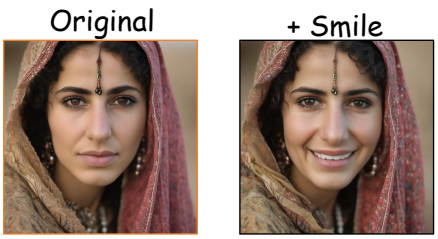

Figure 1. Attribute manipulation results for *Smile*. The image shows the subject before and after the expression modification.

Multi-Subject Manipulation. The goal of multi-subject manipulation is to steer the facial attributes of a specific target subject (*e.g.*, a man or a woman) within a multi-subject scene, while preserving the other subjects and the rest of the scene consistency. As shown in Sec.4.5 in the main paper, we employ the fine-tuning objective $\mathcal{L}_{\text{multi}} = \mathcal{L}_{\text{sae}} + \eta\mathcal{L}_{\text{cons}}$, where η is set to 0.1. During inference, we first encode the paired sentences hidden states H^+ and H^- (with and without the target attribute) into their corresponding sparse latent representations, denoted as z^+ and z^-, using the pre-trained Attribute Sparse Autoencoder. These representations are then passed through the Attention Pooling (AttPooling) Aggregator (AAg) module, which aggregates the attribute-related semantics into compact token-level embeddings. The aggregated directional vector, $\Delta z = \text{AAg}(z^+) - \text{AAg}(z^-)$, captures the pure attribute semantics. This vector is then decoded via the SAE decoder and added to the target subject's embedding e_{target}, yielding the updated subject embedding: $e'_{\text{target}} = e_{\text{target}} + \text{DEC}(\lambda \cdot \Delta z)$. This process ensures that the attribute manipulation is precisely localized to the target subject, while preserving the identity and context of other elements in the scene.

Photography Style List. We train our model on a diverse set of photography styles, including: *architectural, astrophotography, black-and-white, blue hour, bokeh, cinematic, documentary, dreamy, fashion, film grain, fisheye, food, golden hour, HDR, high contrast, infrared, landscape, long exposure, low light, macro, minimalist composition, monochrome, muted colors, natural lighting, neon lighting, overexposed, pastel tone, portrait, product, sepia tone, soft focus, street photography, studio lighting, surreal, telephoto, underexposed, vintage, wide-angle*.

2. Additional Experimental Results

2.1. Additional attribute Manipulation

We present additional qualitative results demonstrating our method's ability to manipulate a wide range of attributes in an identity-preserving manner, as illustrated in Figures 5 to 8. Each figure below shows a series of images in rows, where the leftmost column is the original input image, and the subsequent columns represent progressive edits along a specific semantic attribute.

2.2. Training SAE with FLUX

To evaluate the generalization of our method, we train the SAE on the FLUX.1-dev [2] model. Text embeddings are extracted from the 23rd layer of the T5 text encoder (out of 24 layers in total). Qualitative results of this experiment are presented in Figure 2.

2.3. Comparison with Raw Embeddings

We compare our sparse direction, $W_{\text{dec}}(\lambda \times \text{ENC}(x_{\text{A}}))$, against raw embedding addition, $\lambda \times x_{\text{A}}$, in Table 1.

Figure 2. Smile manipulation on Flux using our method. The generated results maintain the original identity and scene context while achieving realistic smile steering.

Our method consistently outperforms Emb_{raw} across all attributes and metrics, yielding an average improvement of 0.212 in QS and 0.196 in IS. This confirms that SAE-derived directions capture purer semantic signals than the raw embeddings from the pretrained text encoder. As noted in [1], original embedding spaces are often dense and semantically entangled, which can cause unintended identity shifts (see Fig.1 in the main paper). By mapping these into a high-dimensional sparse space, our approach achieves semantic disentanglement by decomposing representations into their underlying, interpretable components. This alignment with SpLICE [1] further demonstrates the effectiveness of decomposing dense representations into sparse, clean units.

Table 1. Comparison with raw attribute embedding.

Method	Old		Smile		Makeup		Avg	
	QS	IS	QS	IS	QS	IS	QS ↑	IS ↑
Emb_{raw}	3.791	0.497	4.220	0.520	3.960	0.489	3.990	0.502
Ours	**4.049**	**0.716**	**4.265**	**0.637**	**4.291**	**0.742**	**4.202**	**0.698**

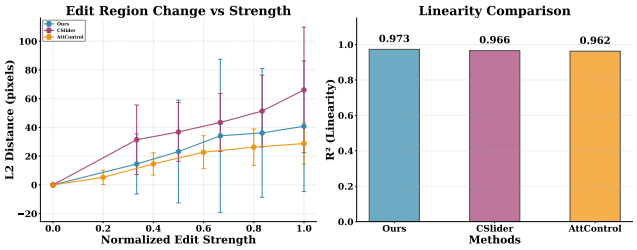

Figure 3. Geometric edit continuity and linearity comparison.

2.4. Geometric metrics for continuity

We adopt dlib landmarks to track facial geometry changes, computing L2 pixel distances within the edit region across steering strengths. As shown in Figure 3, all methods achieve monotonic geometric progression, with our method exhibiting a moderate edit magnitude (neither over-editing like *CSlider* nor under-editing like *AttControl*). More importantly, our method achieves the highest linearity (R^2 of 0.973), compared to *CSlider* (0.966) and *AttControl* (0.962), confirming its superior precision in continuous control.

2.5. Visualize sparsity of latent space

We compare raw activations (*i.e.*, CLIP text embeddings) with those in our SAE latent space. As shown in Figure 4, raw embeddings are semantically entangled, whereas our SAE yields a sparse, axis-aligned representation. In this space, distinct dimensions are selectively triggered for specific attributes (*Smile* vs.*Unsmile*) samples.

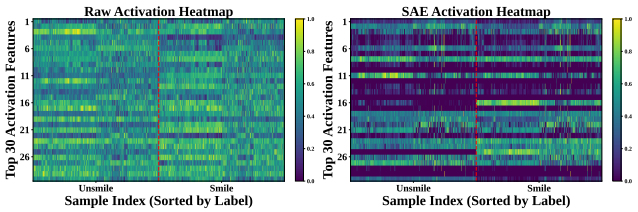

Figure 4. Raw vs. SAE latent space sparsity.

3. Prompt Information

For reproducibility, we provide the text prompts used to generate the images in the appendix figures.

3.1. Prompts for Main Figures

The text prompts used to generate the images for the main figures are as follows.

Figure 1: (1) *A photo of a gentle Thai man with layered hair, in a warm-toned studio.*
Figure 1: (2) *A photo of an honest Dutch man in a white button-up shirt, with a blurred background.*
Figure 1: (3) *A high-quality portrait of a man in a tuxedo, with clear, detailed facial features.*
Figure 4: Prompt: *A close-up of a woman wearing a red scarf, looking thoughtful.*
Figure 5: *A photo of an honest Latin American talented woman.*
Figure 6: (1) *A woman with medium-length blonde hair, pale skin, and soft blue eyes.*
Figure 6: (2) *A photo of a peaceful Middle Eastern woman.*
Figure 7: *A photo of a woman wearing a red scarf.*
Figure 8: (1) *A man wearing glasses while holding a banana.*
Figure 8: (2) *A man in front of a Christmas tree with his dog.*
Figure 9: (1) *A photo of a forest.*
Figure 9: (2) *A path through a dense forest.*
Figure 10: (1) *A portrait of a woman and a man, standing under cherry blossoms.*
Figure 10: (2) *A portrait of a woman and a man holding cameras in a motorcycle garage with subtle tones.*
Figure 10: (3) *A portrait of a woman and a man, wearing scarves in a winter city, in gentle morning sunshine.*

Figure 10: (4) *A portrait of a woman and a man wearing photography gear at a music studio in a cool color palette.*

3.2. Prompts for Appendix Figures

The text prompts used to generate the images in the appendix are listed below.
Figure 5 prompts:
Smile: (1) *A photo of a charming Singaporean man with clean-cut hair, with studio backdrop.*
(2) *A photo of a composed Moroccan man wearing a woolen scarf, with a blurred background.*
(3) *A photo of an honest Dutch man in a white button-up shirt, with a blurred background.*
Old: (1) *A photo of a refined Italian man wearing a checkered blazer, with soft window light.*
(2) *A photo of a serene Danish man with round glasses, near a sunlit window.*
(3) *A photo of a poised Greek person in a linen shirt, in a studio with soft shadows.*
Wearing-Lipstick: (1) *A photo of an honest Latin American talented woman.*
(2) *A woman with short hair, looking at the camera.*
Afro: *A portrait of a woman standing by the beach at sunset, with soft and natural lighting.*
Wavy-Hair: *A photo of a composed Swedish man with short blond hair, with a blurred background.*

Figure 6 prompts:
Eyeglasses: (1) *A photo of an elegant Chinese man wearing a red scarf, with a blurred background.*
Eyeglasses: (2) *A photo of a cheerful Filipino man wearing a red scarf, with a soft gray backdrop.*
Hat: (1) *A photo of a beautiful woman.*
Hat: (2) *A photo of a wise European man wearing a casual sweater, in a sunny environment.*
Bald: (1) *A photo of a gentle Vietnamese man with short cropped hair, with a soft gray backdrop.*
Bald: (2) *A photo of a serene Danish man with round glasses, near a sunlit window.*
Blond-Hair: (1) *A photo of a sophisticated Russian woman with a pearl necklace, in a minimalist room.*
Blond-Hair: (2) *A woman with short hair, looking at the camera.*

Figure 7 photography style prompts:
(1) *A photo of sun rays through tall trees.*
(2) *A peaceful scene of a lake at sunset.*
(3) & (6) *A group of tall trees in a park.*
(4) & (5) *A photo of a lone tree in a vast field.*
(7) *A sunrise over a calm ocean.*
(8) *A photo of a snowy forest path.*

Figure 8 prompts:
(1) *A portrait of a woman and a man framing a shot together in a snowy village in cinematic lighting.*
(2) *A close-up of a woman and a man, front of a cafe, with crisp autumn light.*
(3) *A portrait of a woman and a man, wearing scarves in a winter city, in gentle morning sunshine.*
(4) *A portrait of a woman and a man standing by the beach at sunset, with soft and natural lighting.*

References

[1] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). In *Advances in Neural Information Processing Systems*, pages 84298–84328. Curran Associates, Inc., 2024. 3

[2] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

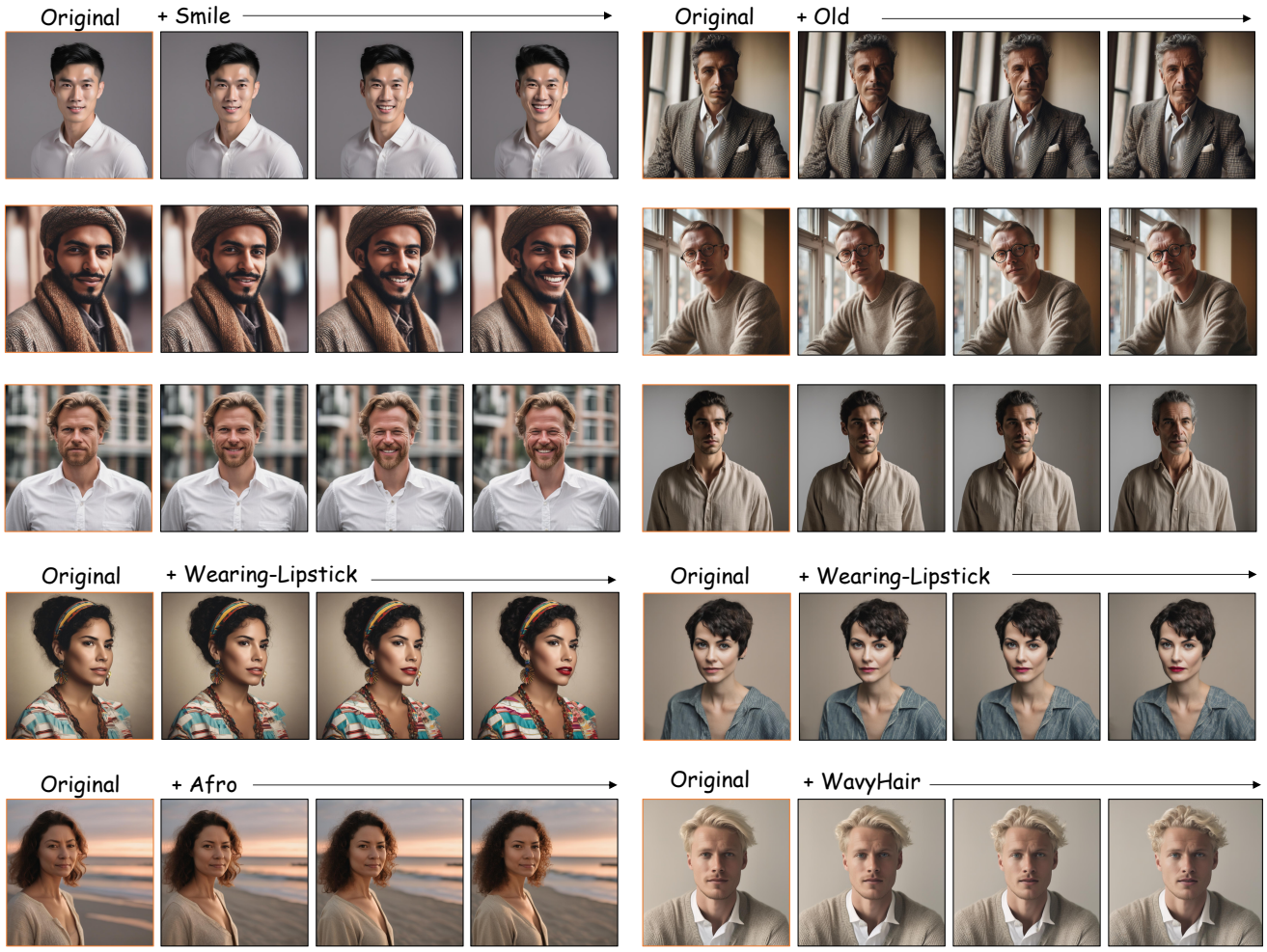

Figure 5. Continuous manipulation of the *Smile*, *Old*, *Wearing-Lipstick*, *Afro* and *WavyHair* attribute. Each row shows a subject transitioning from neutral to the attribute steered, with identity and lighting well preserved.

Figure 6. Attribute manipulation results across various appearance traits. Each pair shows the subject before (left) and after (right) editing, demonstrating the effect of adding attributes, such as *Eyeglasses*, *Hat*, *Bald*, and *BlondHair*. The results illustrate the model's ability to generate realistic and identity-consistent modifications across a diverse set of visual changes.

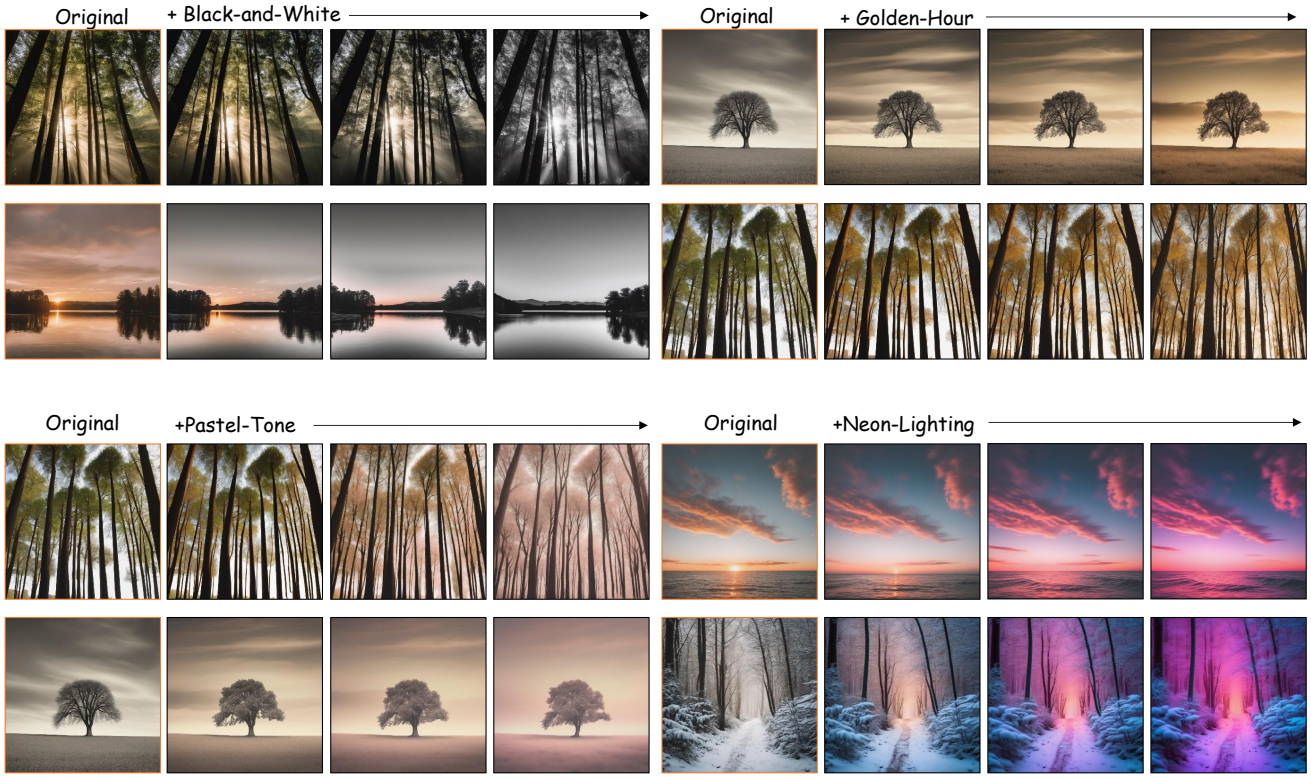

Figure 7. Continuous manipulation of photography style, such as the *Black-and-White*, *Golden-Hour*, *Pastel-Tone* and *Neon-Lighting* attribute. Each row demonstrates the style transition on the original images while maintaining their core structural consistency.

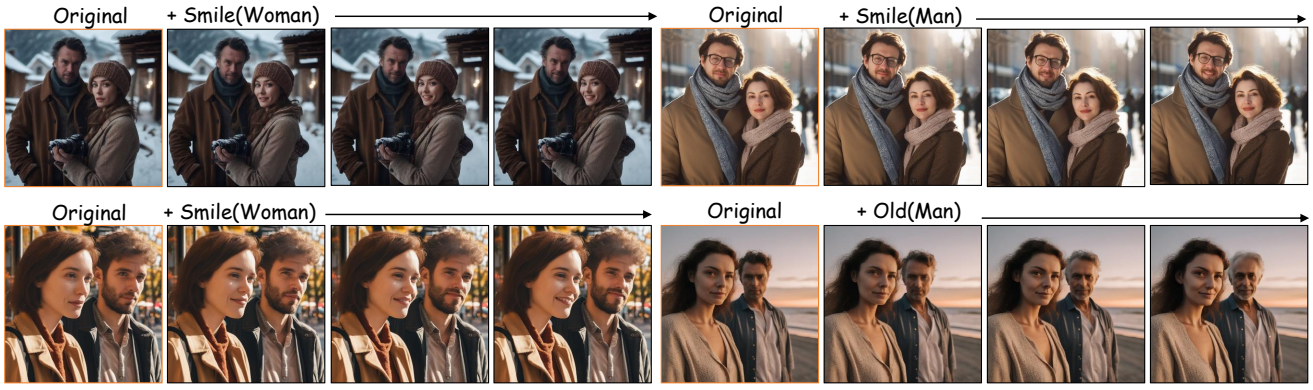

Figure 8. Continuous manipulation of attribute in a multi-subject scene. Each row demonstrates the continuous application of an attribute (*e.g.*, *Smile*, *Old*) to a specific subject (*e.g.*, *man* or *woman*) while keeping the other subject unchanged.