Text Slider: Efficient and Plug-and-Play Continuous Concept Control for Image/Video Synthesis via LoRA Adapters

Supplementary Material

A. Limitation

Text Slider provides a training-efficient approach for continuous attribute modulation, enabling faster custom slider training with lower GPU requirements and scalability to larger concept sets, making it accessible to a broader range of users. However, our method inherits the limitation of low-rank adaptation [20], it remains sensitive to excessively large inference-time scaling factors, which can cause catastrophic forgetting of base knowledge and resulting in unnatural expressions or structural distortions.

B. More Implementation Details

Model Checkpoints. For text-to-image experiments, we use the official SD-XL model checkpoint[1] consistently throughout all evaluations. For SD-1.5, we adopt a high-quality community model[2] to enhance generation fidelity. For text-to-video experiments, we primarily employ SD-1.5 community checkpoints[23], selected for their compatibility and visual quality. For video-to-video experiments, we use the same community checkpoint[3] as MeDM [7] to ensure consistency across our evaluations.

Training Prompt Pairs Generation. We follow the same prompt generation strategy as Concept Slider [13]. Specifically, we use OpenAI GPT-4o with a predefined system prompt[4], which generates contrastive prompt pairs for the target concept. More detailed examples of the training prompt pairs are provided in Table A1.

C. Qualitative Comparison

Text-to-Image Generation. Figure A1-A3 present more diverse results on SD-XL using Text Slider. Figure A4 and A5 provides additional qualitative comparisons with baseline methods on SD-XL and SD-1, respectively. Notably, unlike Concept Slider, which requires model-specific training, Text Slider generalizes across different architectures that share the same text encoder without the need for retraining and achieves comparable results.

Text-to-Video Generation. Figure A6 showcase a detailed comparison with baseline methods. To ensure a fair comparison, we primarily focus on person-related attributes, as baseline methods like Attribute Control have limited ability to manipulate global properties such as style or weather.

Video-to-Video Generation. As shown in the baseline comparison in Figure A7, our approach delivers competitive visual quality with significantly lower computational overhead. In contrast, Video-P2P struggles with subtle facial edits and often introduces artifacts, while also requiring per-video model tuning. Text Slider, by comparison, offers a plug-and-play solution that generalizes across videos without additional fine-tuning.

D. User Study

In Figure A9, we illustrate some sample questions of the questionnaire in our user study. An instruction and an example question are provided on the left to let evaluators familiar with the criteria and question format before starting the actual evaluation on the right. Each question consists of three rows corresponding to different methods, with their order randomized to prevent bias and ensure fairness in assessment.

E. Ablation Study

Diffusion Noise Prediction. In Figure A10, we compare the diffusion noise prediction setting by reporting ΔCLIP and LPIPS across five attribute intensities for four attributes: age, smile, curly hair, and chubby. Figure A11 further provides qualitative results, confirming that our method achieves performance comparable to the setting that back-propagates through the diffusion model.

CLIP Text Encoder. We present a qualitative comparison of three settings in Figure A12: our default (CLIP+OpenCLIP), CLIP-only, and OpenCLIP-only. All settings effectively manipulate attributes, but the default setting offers stronger control over certain attributes (*e.g.*, curly hair, chubby), enabling a broader and more diverse range of concepts.

F. Societal Impact

Text Slider offers an efficient approach for continuous attribute control in image and video synthesis, enabling creative applications in design and entertainment. Its efficiency makes advanced generative tools more accessible to users with limited computational resources. However, the ability to manipulate visual attributes raises risks of misuse in misinformation, deepfakes, and identity spoofing. Therefore, responsible deployment should include safeguards such as content provenance tracking, user consent mechanisms, and bias audits to ensure ethical and fair use.

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[2] https://huggingface.co/SG161222/Realistic_Vision_V6.0_B1_noVAE

[3] https://civitai.com/models/43331/majicmix-realistic

[4] https://github.com/rohitgandikota/sliders/GPT_prompt_helper.ipynb

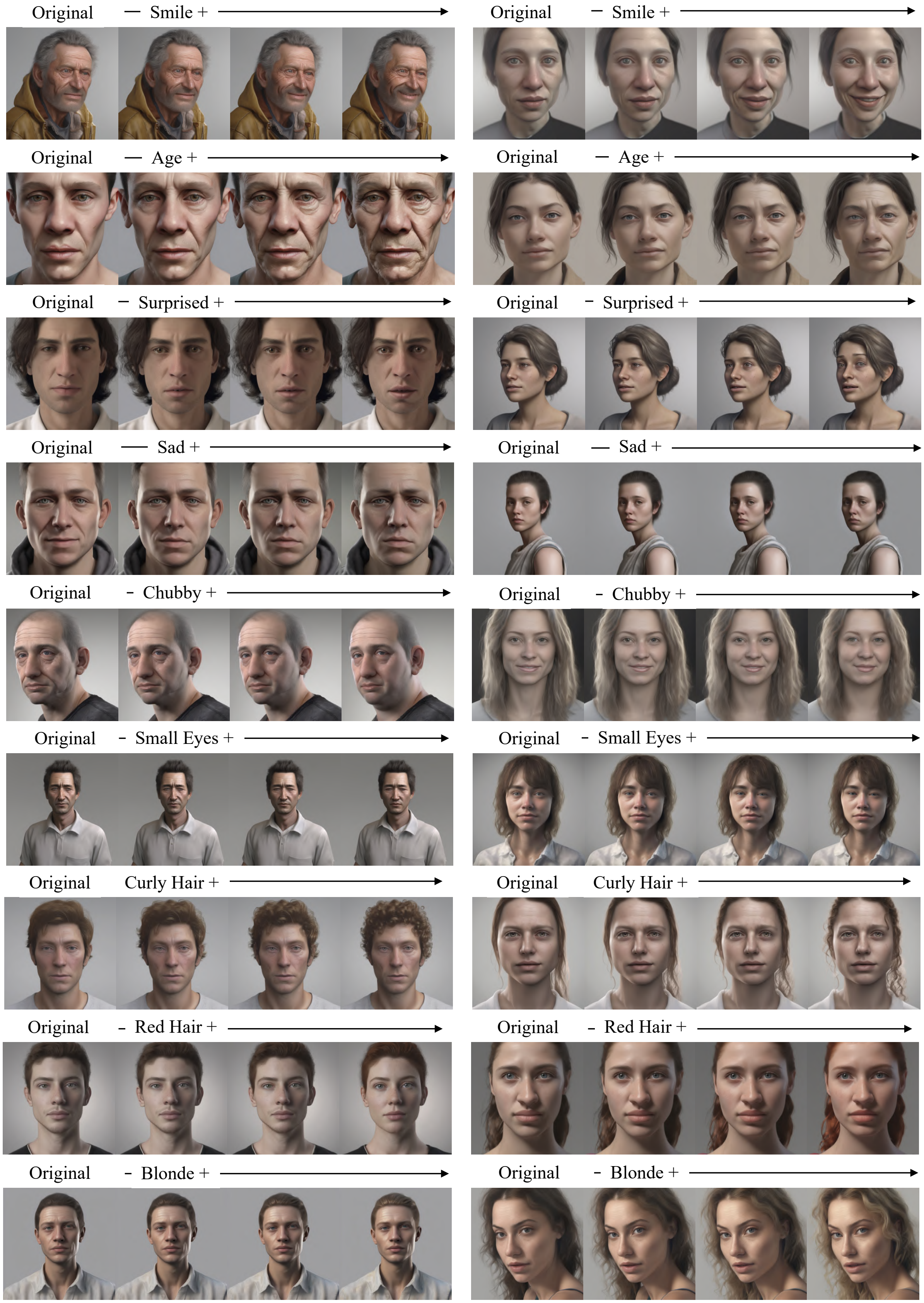


Figure A1. **More Qualitative Results on SD-XL.** We present more results using Text Slider across face, eyes and hair-related attributes.

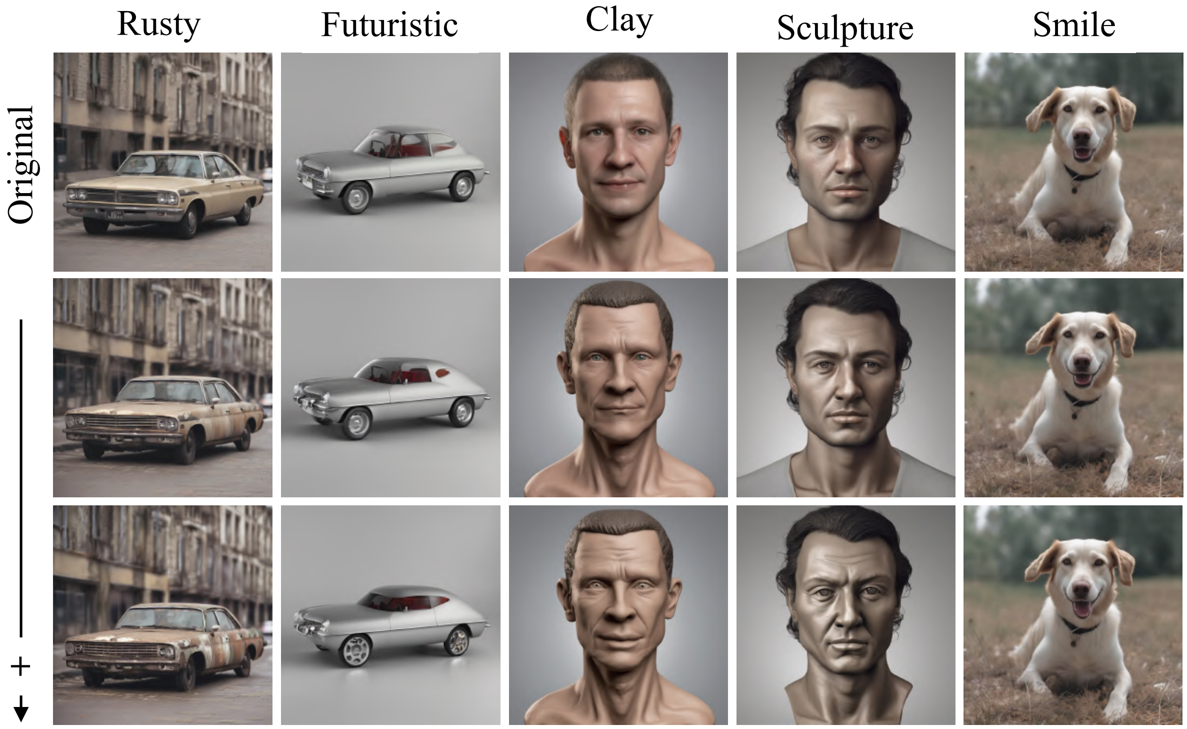


Figure A2. **More Qualitative Results on SD-XL.** Our method is also effective for attributes related to cars, styles, and dogs.

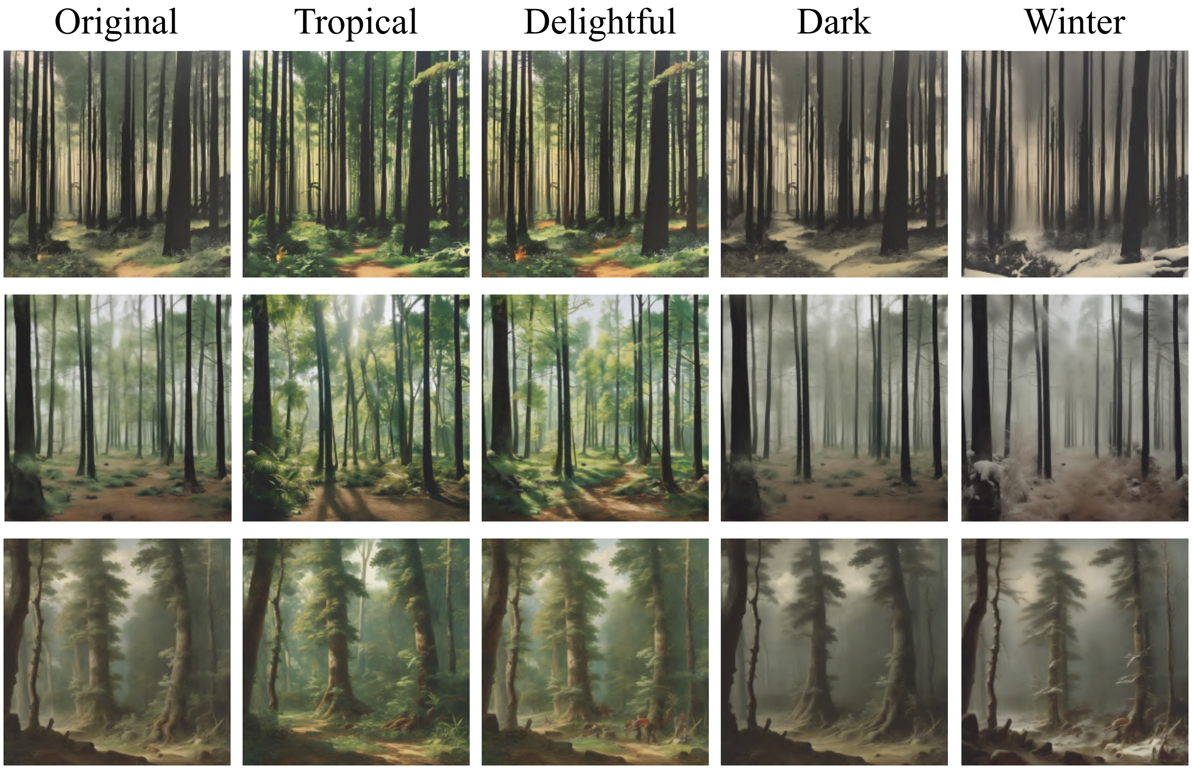


Figure A3. **Qualitative Results for scene-related attributes on SD-XL.**

Slider	Target	Positive	Negative	Preserved
Age	person	person, elderly, wrinkles, gray hair, aged skin, signs of maturity, wise expression	person, young, smooth skin, youthful appearance, no wrinkles, energetic expression	white race, black race, indian race, asian race, hispanic race ; male, female
Smile	person	person, smiling, happy face, big smile, joyful expression	person, frowning, grumpy, sad, neutral expression	white race, black race, indian race, asian race, hispanic race ; male, female
Chubby	person	person, chubby, round face, soft features, fuller cheeks, plump body shape	person, skinny, thin face, sharp features, slim body shape	white race, black race, indian race, asian race, hispanic race ; male, female
Beard	person	person, full beard, thick facial hair, well-groomed beard, masculine appearance	person, clean-shaven, no facial hair, smooth chin, youthful look	white race, black race, indian race, asian race, hispanic race ; male, female
Makeup	person	person, wearing makeup, well-applied foundation, eye shadow, lipstick, glamorous look, enhanced facial features	person, no makeup, natural skin, bare face, unaltered appearance	white race, black race, indian race, asian race, hispanic race ; male, female
Small eyes	person	person, small eyes, narrow eye shape, subtle eyelids, low eye-to-face ratio	person, large eyes, wide eye shape, prominent eyelids, high eye-to-face ratio	white race, black race, indian race, asian race, hispanic race ; male, female
Surprised	person	person, surprised expression, wide eyes, raised eyebrows, open mouth, shocked face, startled posture, expressive emotion	person, neutral expression, relaxed face, calm demeanor, closed mouth, steady gaze, no visible emotion	white race, black race, indian race, asian race, hispanic race ; male, female
Curly hair	person	person, curly hair, defined curls, voluminous texture, spiral or wavy patterns, natural bounce, well-defined ringlets	person, straight hair, smooth texture, no curls or waves, flat appearance	white race, black race, indian race, asian race, hispanic race ; male, female ; short hair, long hair, medium length hair
Red hair	person	person, red hair, vibrant copper tones, fiery orange red shades, natural auburn highlights, bold striking hair color	person, blond hair, black hair, brown hair, grey hair, non-red shades	white race, black race, indian race, asian race, hispanic race ; male, female ; short hair, medium hair, long hair ; straight hair, wavy hair, curly hair
Blonde	person	person, blonde hair, golden tones, light yellow shades, bright and radiant hair color	person, dark hair, black hair, brown hair, red hair, non-blonde shades	white race, black race, indian race, asian race, hispanic race ; male, female ; short hair, medium hair, long hair ; straight hair, wavy hair, curly hair
Cartoon		cartoon style, exaggerated features, bold outlines, flat shading, vibrant colors, stylized characters, playful proportions, simplified textures, hand-drawn appearance	realistic style, natural proportions, detailed textures, realistic lighting, lifelike shading, photographic accuracy	white race, black race, indian race, asian race, hispanic race ; male, female ; urban background, nature background, indoor scene
Pixar		pixar style, 3D animation, smooth and rounded features, expressive faces, high-quality rendering, vibrant and clean visuals, family-friendly aesthetic	realistic style, detailed textures, natural proportions, lifelike appearance, photographic realism	white race, black race, indian race, asian race, hispanic race ; male, female ; child, adult, elderly
Clay		clay style, claymation look, sculpted textures, hand-molded appearance, matte surfaces, visible fingerprints, soft rounded edges, stop-motion aesthetic, playful handcrafted feel	realistic style, smooth digital textures, lifelike proportions, clean lines, high detail realism, photographic surfaces	white race, black race, indian race, asian race, hispanic race ; male, female ; indoor setting, outdoor setting, neutral background, colorful background
Sculpture		sculpture, carved appearance, stone or marble texture, rigid posture, chiseled features, statue-like, solid material, classical sculpture aesthetic, matte surface	lifelike, soft skin, natural textures, fluid posture, organic materials, realistic surface detail	white race, black race, indian race, asian race, hispanic race ; male, female
Tropical		tropical scene, lush green palm trees, warm sandy beaches, turquoise ocean, humid atmosphere, exotic plants, bright sunlight, vibrant flowers, tropical wildlife	non-tropical scene, dry landscape, temperate forest, rocky terrain, cool climate, muted colors, overcast sky	white race, black race, indian race, asian race, hispanic race ; male, female ; urban setting, nature background, indoor setting
Winter		winter scene, snow-covered landscape, icy ground, frosty trees, frozen lakes, snowflakes falling, cloudy sky, cold atmosphere, winter lights, visible snow piles	summer scene, dry ground, green grass, leafy trees, clear sky, warm lighting, no snow, sunlit atmosphere	white race, black race, indian race, asian race, hispanic race ; male, female ; mountain setting, urban setting, countryside
Delightful		delightful atmosphere, joyful expressions, cheerful colors, heartwarming scene, positive energy, vibrant and lively mood	gloomy atmosphere, sad expressions, dull colors, depressing scene, negative emotion, dark and heavy mood	white race, black race, indian race, asian race, hispanic race ; male, female ; indoor setting, outdoor setting, urban background, natural background
Rusty	car	car, rusty, corroded metal, peeling paint, oxidized surface, aged appearance, weathered condition	car, clean, polished, shiny surface, new paint, well-maintained, pristine condition	sedan, SUV, truck, convertible ; red, blue, black, white, silver
Futuristic	car	car, futuristic, sleek aerodynamic design, glowing neon lights, advanced technology features, metallic surfaces, sci fi style, high-tech appearance	car, traditional, old-fashioned, classic design, rustic, vintage, minimal technology	sedan, SUV, truck, convertible ; red, blue, black, white, silver

Table A1. **Detailed Prompts for Training Sliders.**

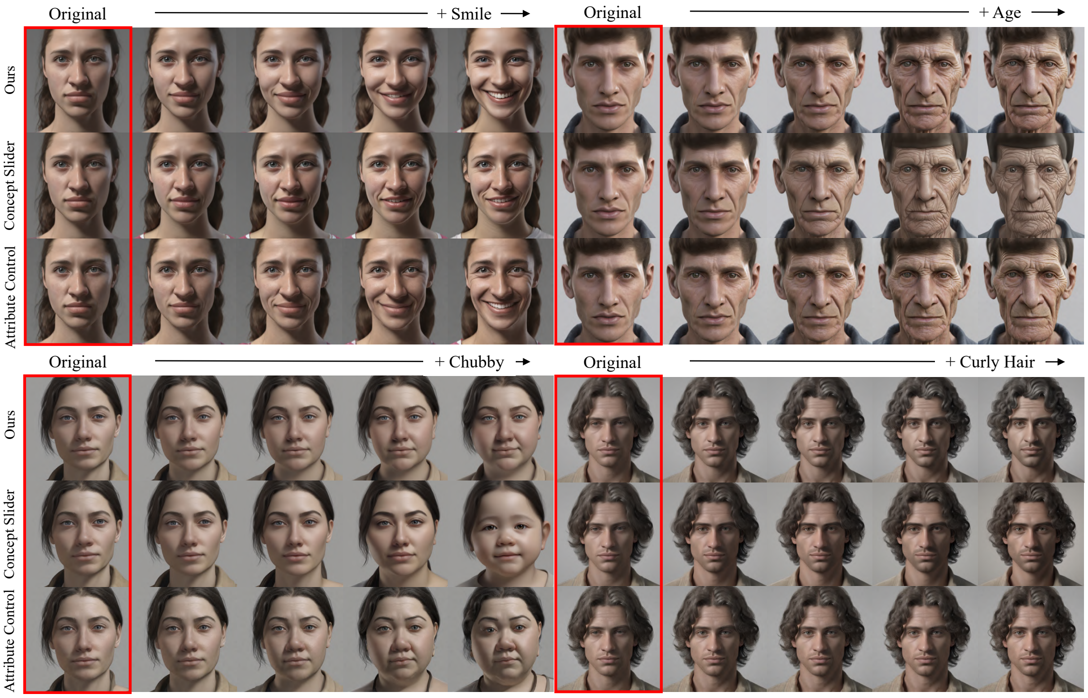


Figure A4. **Qualitative Comparison of Text-to-Image Results on SD-XL.** We qualitatively compare Text Slider with Concept Slider [13] and Attribute Control [1] on SD-XL [27] across four attributes, smile, age, chubby and curly hair. Each attribute evaluated at four levels of intensity. Red boxes highlight the original generated images for reference.

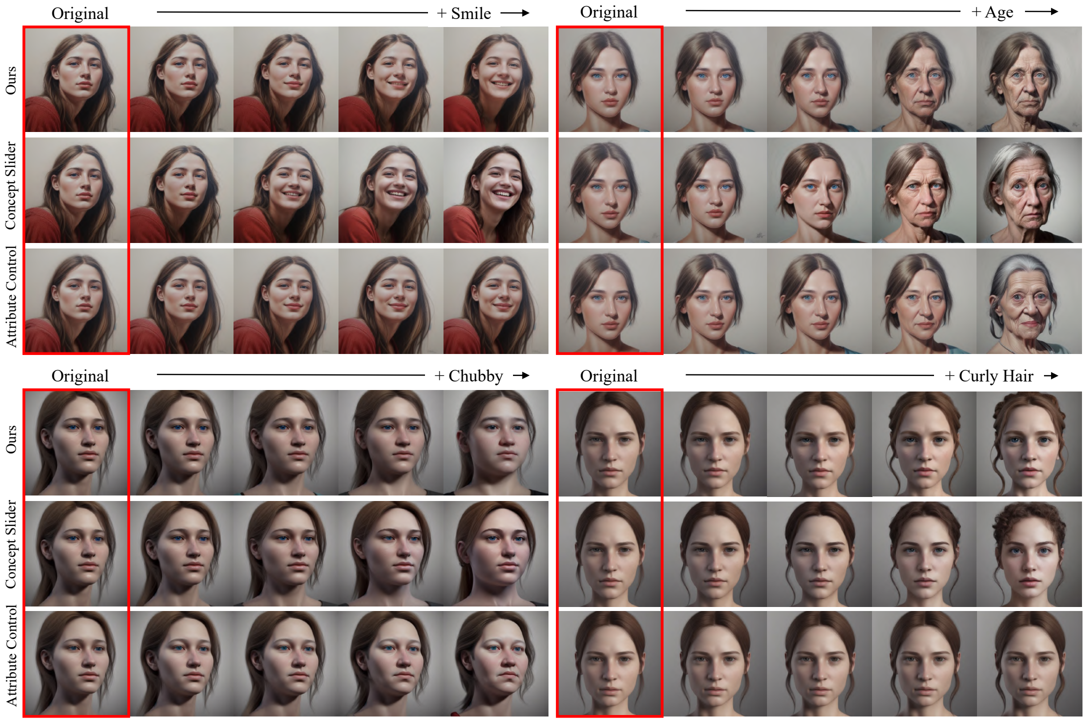


Figure A5. **Qualitative Comparison of Text-to-Image Results on SD-1.** We qualitatively compare Text Slider with Concept Slider [13] and Attribute Control [1] on SD-1 [30] across four attributes, smile, age, chubby and curly hair. Each attribute evaluated at four levels of intensity. Red boxes highlight the original generated images for reference.

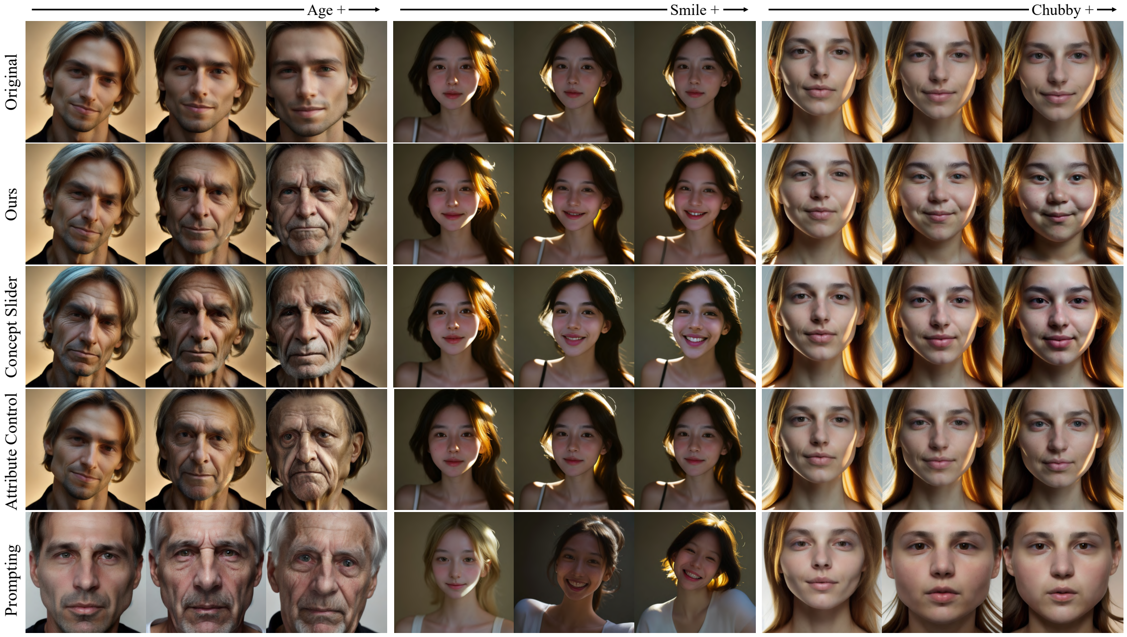


Figure A6. **Qualitative Comparison of Text-to-Video Results.** We compare AnimateDiff [16] integrated with Text Slider, Concept Slider [13], and Attribute Control [1] across three attributes. For each video, three representative frames are sampled to illustrate the gradual progression of attribute intensity over time.

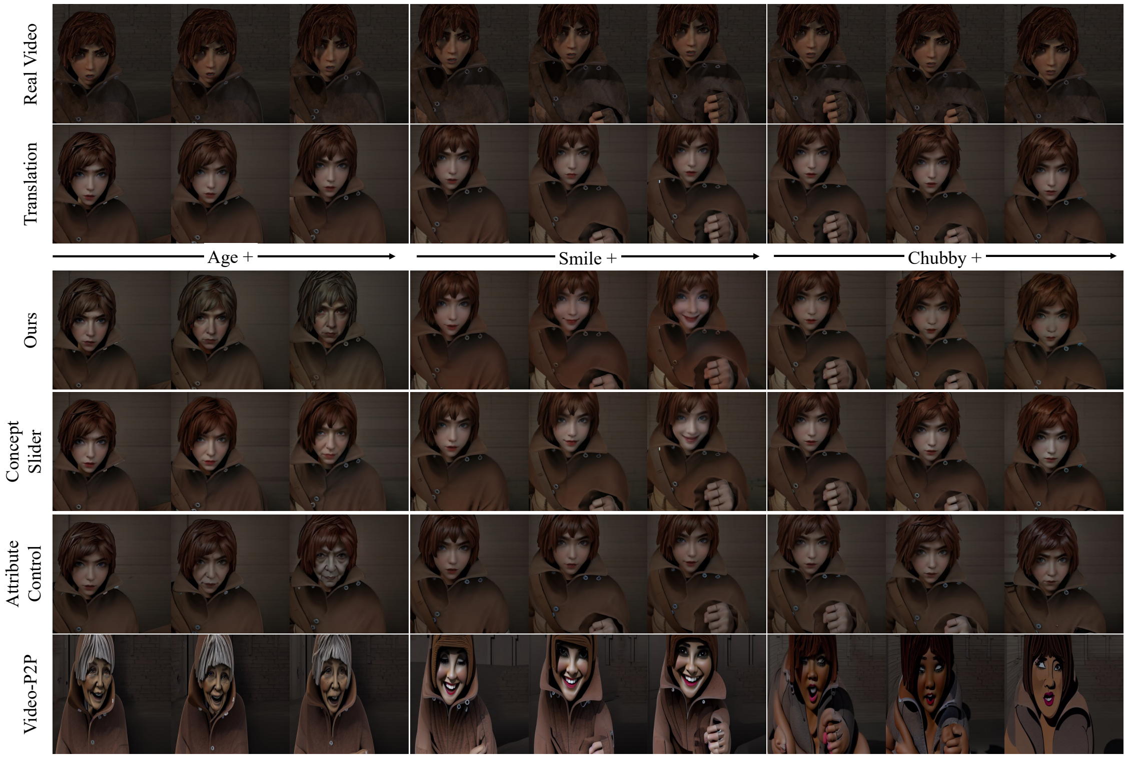


Figure A7. **Qualitative Comparison of Video-to-Video Results.** Real videos are first translated using MeDM [7] with SDEdit [25], followed by editing with Text Slider, Concept Slider [13], and Attribute Control [1] across three attributes. For Video-P2P, real videos are first inverted and then edited using attention map-based control. For each video, three representative frames are sampled to illustrate the gradual progression of attribute intensity over time.

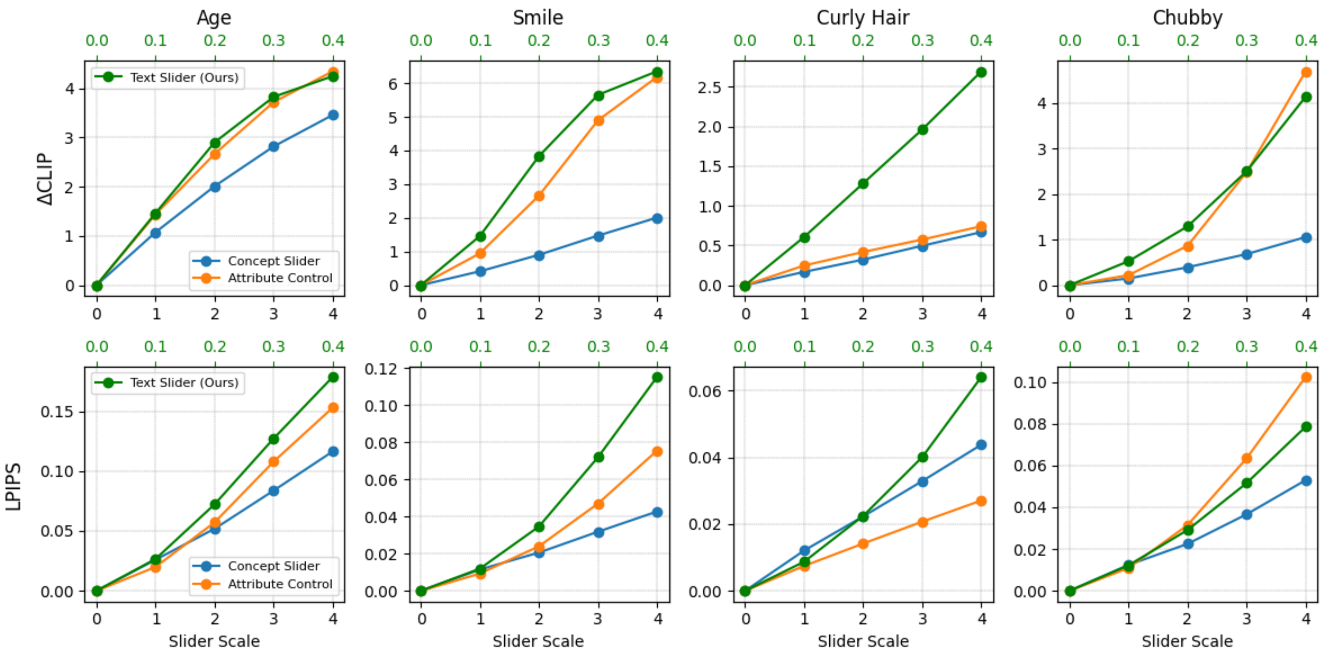
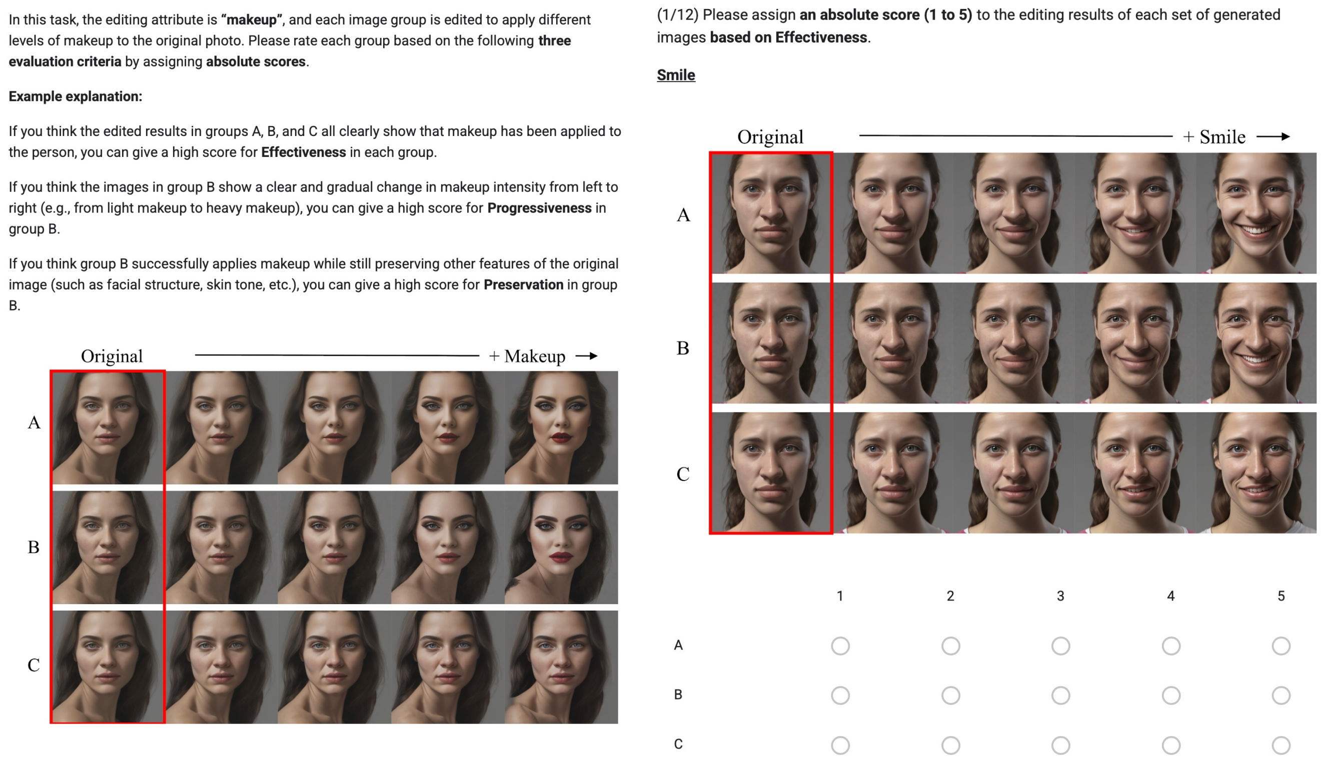


Figure A8. **Detailed Performance of Text-to-Image Generation on SD-XL.** We report performance metrics using ΔCLIP and LPIPS across four attributes—*age*, *smile*, *curly hair*, and *chubby*—evaluated at five levels of attribute intensity (slider scales). For Concept Slider [13] and Attribute Control [1], we assess scales from 0 to 4, while for Text Slider, we use a range of 0 to 0.4 due to its more compact scaling. Our method achieves comparable performance to the baselines while significantly reducing computational costs.



In this task, the editing attribute is **"makeup"**, and each image group is edited to apply different levels of makeup to the original photo. Please rate each group based on the following **three evaluation criteria** by assigning **absolute scores**.

Example explanation:

If you think the edited results in groups A, B, and C all clearly show that makeup has been applied to the person, you can give a high score for **Effectiveness** in each group.

If you think the images in group B show a clear and gradual change in makeup intensity from left to right (e.g., from light makeup to heavy makeup), you can give a high score for **Progressiveness** in group B.

If you think group B successfully applies makeup while still preserving other features of the original image (such as facial structure, skin tone, etc.), you can give a high score for **Preservation** in group B.

(1/12) Please assign **an absolute score (1 to 5)** to the editing results of each set of generated images **based on Effectiveness**.

Smile

	1	2	3	4	5
A	○	○	○	○	○
B	○	○	○	○	○
C	○	○	○	○	○

Figure A9. **User Study Example.** Instructions and an example question on the left are provided to make evaluators familiar with the criteria, while the actual evaluation on the right presents three rows each denotes different methods, with their order randomized to prevent bias and ensure a fair assessment.

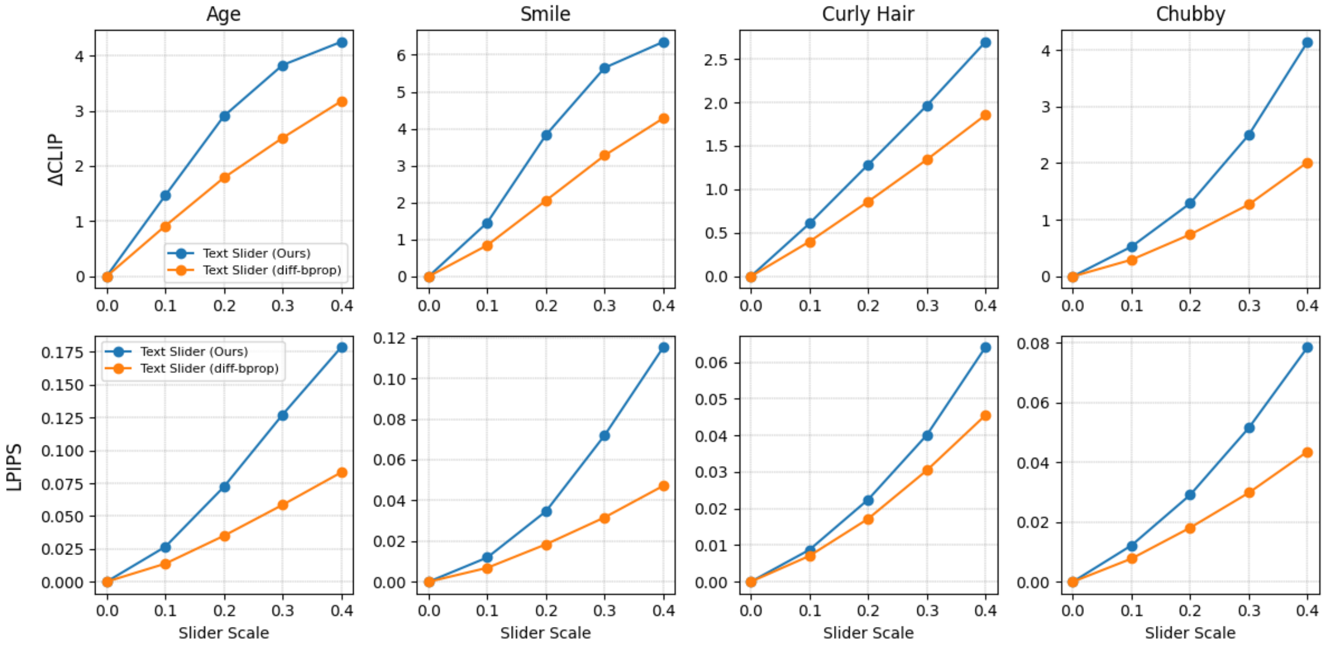


Figure A10. **Comparison on Performance with Diffusion Noise Prediction.** `diff-bprop` denotes the setting where the same LoRA modules are injected into the text encoder, but backpropagation is performed through the diffusion model. Our method achieves comparable performance in both ΔCLIP and LPIPS.

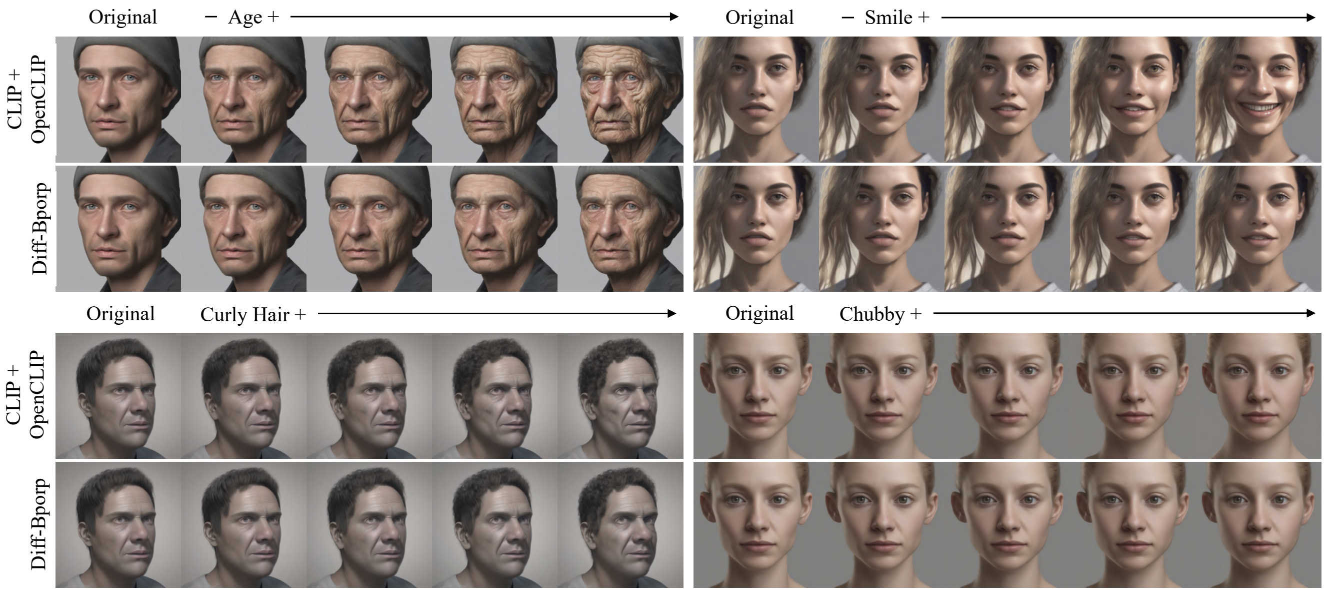


Figure A11. **Qualitative Comparison on Diffusion Noise Prediction.** We evaluate four attributes—age, smile, curly hair, and chubby—using five levels on a 0–0.4 scale with 0.1 intervals. Our method achieves comparable qualitative performance to the variant that backpropagates through the diffusion model, while significantly reducing computational costs.

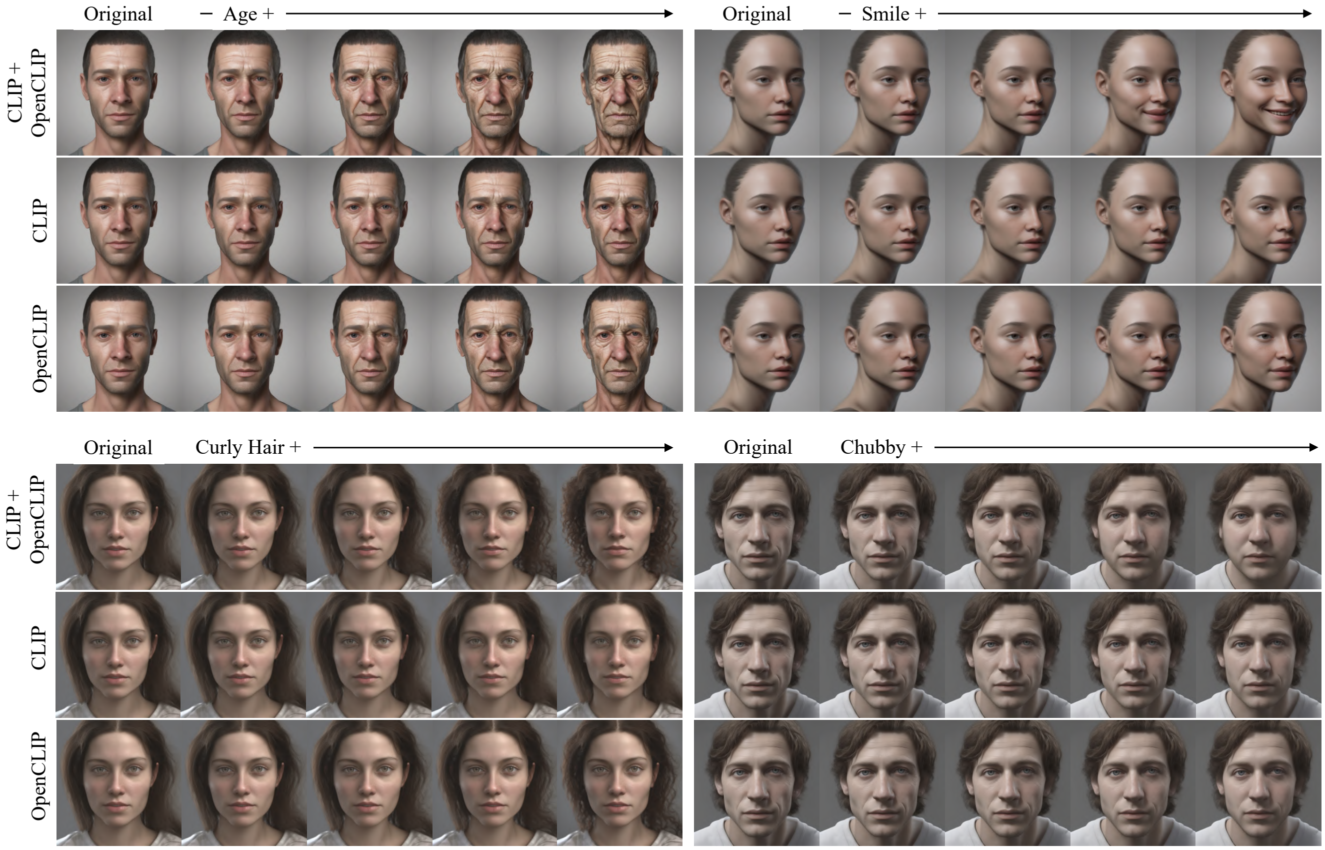


Figure A12. **Qualitative Comparison across CLIP Text Encoders.** All settings enable effective attribute manipulation, with single text encoder configurations offering a more training-efficient alternative. However, the default configuration provides stronger control over certain attributes (*e.g*., curly hair, chubby), enabling a broader and more diverse range of concepts.