Text Slider: Efficient and Precise Concept Control for Video Generation and Editing via LoRA Adapters

Pin-Yen Chiu I-Sheng Fang Jun-Cheng Chen
Research Center for Information Technology Innovation, Academia Sinica
{nickchiu, ishengfang, pullpull}@citi.sinica.edu.tw

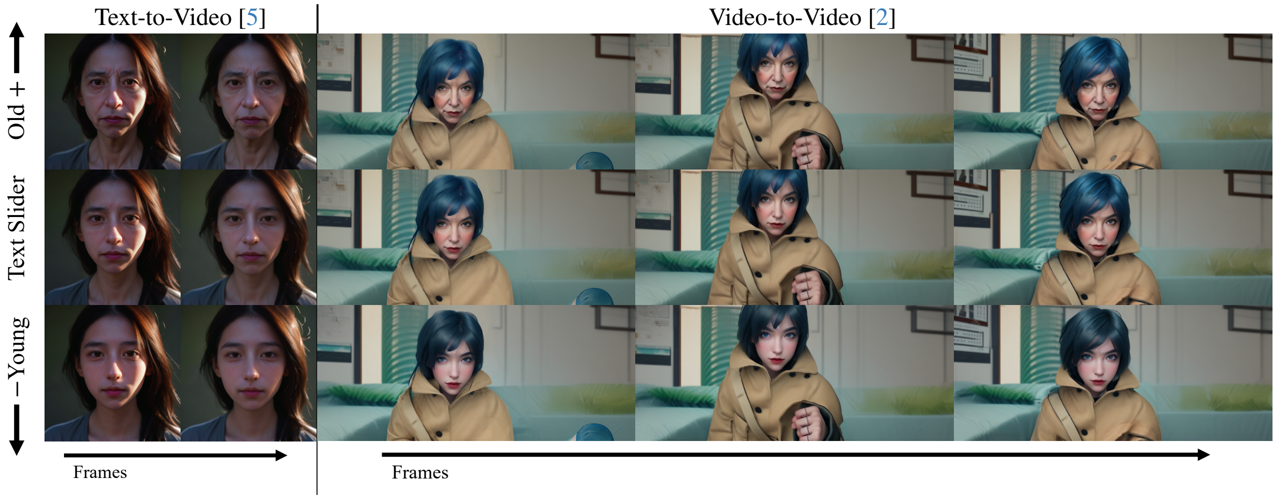

Figure 1. **The Results of Text Slider with Text-to-Video (AnimateDiff [5]) and Video-to-Video (MeDM [2]) Model.** Text Slider has strong adaptability for various video synthesis models and achieves precise control of visual concepts. In these results, we control visual concept (age) and generate videos with different strengths (from young to old) of the concept (age).

Abstract

Video generation and editing using diffusion models have made significant progress in recent years. While free-form text prompts provide flexible control over generation and attribute manipulation, existing methods still struggle to achieve fine-grained control over specific attributes. Moreover, expressing varying degrees of attribute intensity through text alone is often challenging. For example, describing subtle variations in a person's smile can be ambiguous and imprecise. Furthermore, the existing method suffers from limited adaptability and inefficient training. To address these limitations, we introduce Text Slider, a lightweight, efficient and highly adaptable framework that identifies low-rank directions within a pre-trained text encoder, enabling precise control of visual concepts while significantly reducing training time and the number of parameters. Text Slider is plug-and-play, easily composable, and continuously modulated, providing enhanced controllability and fine-grained manipulation for video generation and editing. We demonstrate that Text Slider effectively attenuates or strengthens specific attributes while preserving the original input layout and structure, surpassing current state-of-the-art methods in controllable video synthesis.

1. Introduction

Diffusion models [3, 7, 14] have recently achieved significant progress in text-guided image and video synthesis [1, 8]. While text prompts offer flexible control and allow users to express creative intent, they are often insufficient for achieving continuous and fine-grained manipulation of specific visual concept, especially when subtle variations or intensity levels are needed. For example, showcasing nuanced changes in a person's smile using text alone can be inherently ambiguous. This limitation makes it difficult for creators to perform precise image and video editing.

Existing methods for precise concept control primarily focus on image synthesis, while video synthesis has become an emerging research direction. For example, Prompt-to-Prompt [6] achieves localized control by modifying cross-attention weights, allowing user to edit an image with latent diffusion models (LDMs) [14]. To extend this to the video domain, Video-P2P [10] applies a similar cross-attention reweighting strategy. However, its effectiveness is limited, particularly for fine-grained facial attributes such as age or smile. Another approach is Concept Slider [4], which enables concept-specific generation by learning Low-Rank Adapters (LoRA) and modulating a scaling factor during inference. However, Concept Slider suffers from limited adaptability and inefficiency. Specifically, a separate slider must be trained for each diffusion model architecture. For example, sliders trained for Stable Diffusion 1.5 (SD-1.5) [14] and Stable Diffusion XL (SD-XL) [12] are not interchangeable. Moreover, training a single slider requires approximately 30 minutes and nearly 3 million parameters for SD-1.5, with even longer training time and larger model sizes required for SD-XL, as shown in Table 1. These scalability issues hinder its practical application across diverse model architectures, tasks, and large sets of concepts. To fully unlock the potential of creative and expressive generation, it is essential to develop more adaptable and efficient methods that support continuous and precise attribute modulation, particularly in the context of video synthesis.

In this paper, we introduce Text Slider, an approach inspired by Concept Slider [4], with the added advantage of being seamlessly extendable to various pre-trained image and video diffusion models that share the same text encoder. Unlike Concept Slider, which learns low-rank directions within the diffusion model using contrastive text prompts based on concepts extracted from a pre-trained text encoder, we find the similar concept representations can be learned directly by injecting low-rank parameters into the text encoder itself. Specifically, we fine-tune LoRA adapters exclusively on the text encoder without requiring backpropagation through the diffusion model, enabling efficient and effective control over specific concepts. This design significantly reduces computational requirements, using only $\approx$ 23% of the parameters and $\approx$ 10% of training time required by Concept Slider with SD-1.5, and $\approx$ 15% of the parameters and $\approx$ 7% of training time on SD-XL. Moreover, Text Slider can adapt to different model architectures that share the same text encoder (*e.g.* SD-1.5 and SD-XL), supporting precise and continuous concept control.

Our main contributions are summarized as follows:

- We propose Text Slider, a method that injects and fine-tunes LoRA modules within the pre-trained text encoder, allowing precise concept control without modifying and backpropagating through the diffusion model, reducing $\approx$ 77% of parameters and $\approx$ 90% of training time in SD-1.5, and $\approx$ 85% of parameters and $\approx$ 93% of training time in SD-XL.

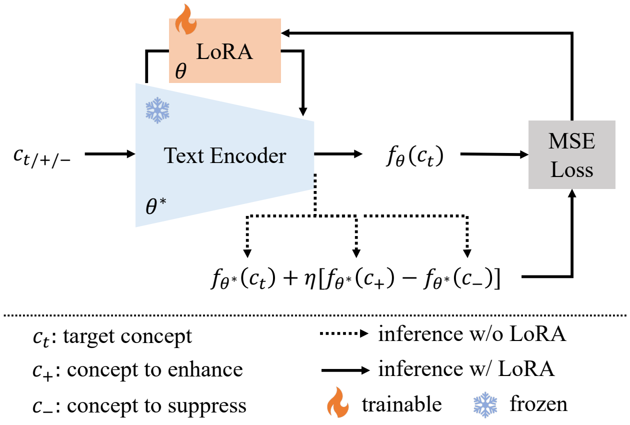

Figure 2. **Overview of Text Slider.** Text Slider injects and learns low-rank parameters within the text encoder using contrastive prompts derived from concept representations extracted by a pre-trained text encoder, enabling precise control over visual concepts.

- Text Slider is plug-and-play, composable and generalizes across various pre-trained diffusion models that share the same text encoder, offering adaptability and reusability.
- Text Slider extends naturally to video synthesis, enabling continuous and fine-grained concept control.

2. Method

2.1. Preliminary

Low-Rank Adaptation (LoRA) [9] is a parameter efficient fine-tuning method that inserts trainable low-rank matrices into pre-trained models while keeping the original weights frozen. Instead of updating the full weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces a low-rank update:

$$W = W_0 + \alpha \cdot BA, \tag{1}$$

where $A \in \mathbb{R}^{r\times k}$, $B \in \mathbb{R}^{d\times r}$, and $r \ll \min(d, k)$. The scaling factor α modulates the strength of the update and can be adjusted at inference time to control the influence of the learned direction.

In our framework, LoRA is applied to the text encoder, enabling efficient and highly adaptable fine-tuning for concept control in both image and video generation.

2.2. Text Slider

Text Slider is a method for fine-tuning LoRA adapters on a text encoder [13] to enable precise image and video control over designated concepts, as shown in Figure 2. Our approach learns low-rank directions that can enhance or suppress the representation of specific attributes when conditioned on a target concept. Unlike previous work [4], Text Slider is not limited to image generation and editing. Since it only fine-tunes LoRA adapters on the text encoder and does not backpropagate through the diffusion model, it can be seamlessly extended to video tasks without any additional effort. The same adapter can be directly applied for

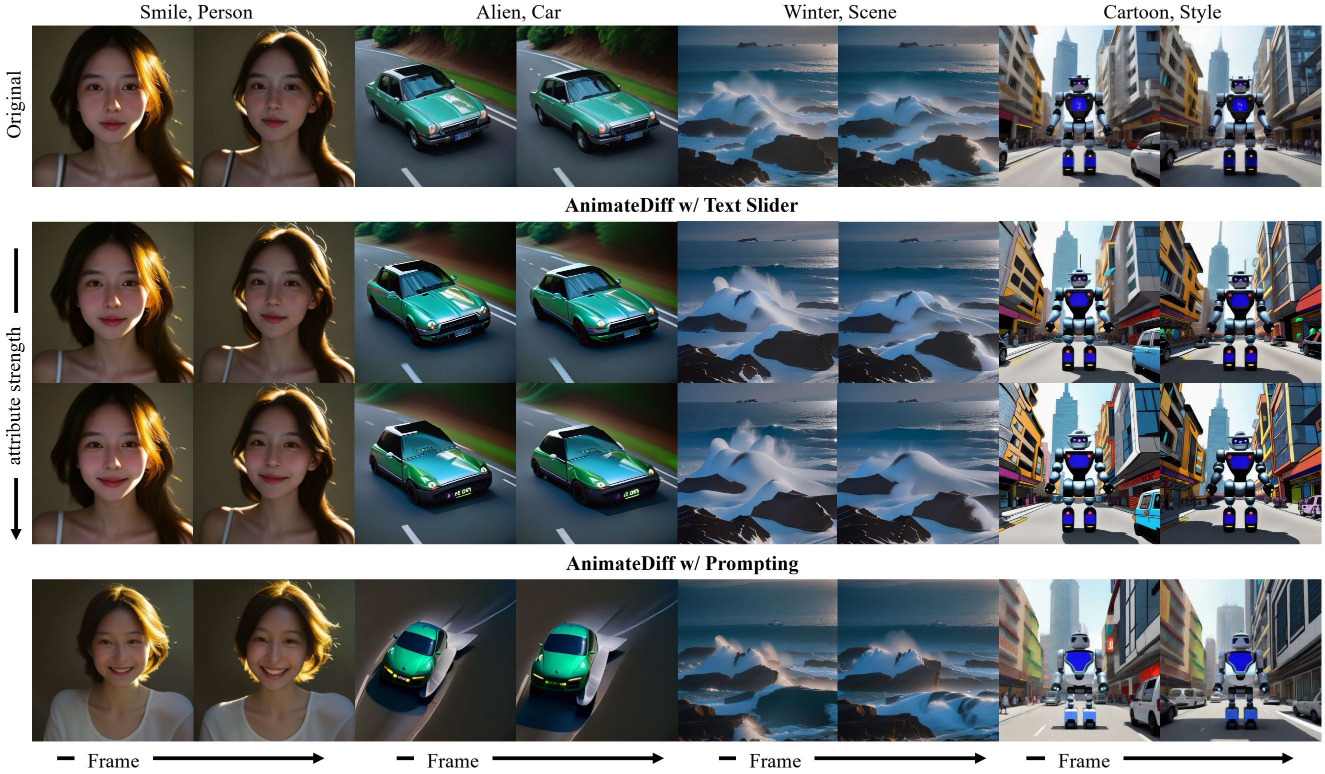

Figure 3. **Comparison of Text-to-Video Results.** AnimateDiff [5] combined with Text Slider enables fine-grained, continuous control over attributes while preserving structure. In contrast, prompt-based control offers limited controllability and often disrupts spatial coherence.

controllable video generation and editing, without any retraining. Moreover, Text Slider requires significantly fewer parameters and less training time.

Given a target concept c_t, we propose to learn a low-rank direction using a model θ that encourages the expression of more positive attributes c^+ while reducing the presence of negative attributes c^-. The model θ is trained by minimizing the mean squared error (MSE) between the prompt embeddings generated by the pre-trained text encoder θ^* and those produced by the adapted model θ. Specifically, f_θ denotes the embedding function that maps a text prompt y into a conditional prompt embedding $f_\theta(y)$. The objective is defined as:

$$\theta^* = \arg\min_\theta \mathbb{E}_y \left\| f_t - f_\theta(y) \right\|_2^2 \tag{2}$$

As illustrated in Figure 2, the target embedding p_t is computed as:

$$f_t = f_{\theta^*}(c_t) + \eta \sum_{q \in \mathcal{Q}} (f_{\theta^*}(c_+, q) - f_{\theta^*}(c_-, q)), \tag{3}$$

where η is a guidance scale and $\mathcal{Q}$ is a set of concepts that should be preserved during attribute manipulation. For example, controlling the "smile" attribute may unintentionally affect other attributes such as age or gender. By incorporating these preserved concepts into the embedding computation, the learned direction becomes more disentangled and less likely to introduce unwanted changes.

To achieve varying degrees of editing strength, we utilize a scaling factor α that can be adjusted at inference time within the LoRA formulation (Equation 1). This parameter controls the intensity of the attribute manipulation, allowing for fine-grained edits, as illustrated in Figure 1.

3. Experiments

3.1. Qualitative Results

We qualitatively evaluate our method on video synthesis tasks using models based on Stable Diffusion 1.5 (SD-1.5) [14]. Specifically, we assess the effectiveness of Text Slider in both text-to-video generation and real video editing. In addition, we demonstrate its ability to compose multiple sliders for controllable video generation. Please refer to Appendix Section A for implementation details.

Text-to-Video Generation. We adopt AnimateDiff [5] as our primary text-to-video framework due to its lightweight design, efficiency, and adaptability to various personalized image diffusion models. To evaluate the effectiveness of combining AnimateDiff with Text Slider, we select four diverse attributes, each applied to a different object category: smile on a person, alien effect on a car, winter effect on a scene, and cartoon style. As shown in Figure 3, Text Slider enables precise and continuous control over attribute inten-

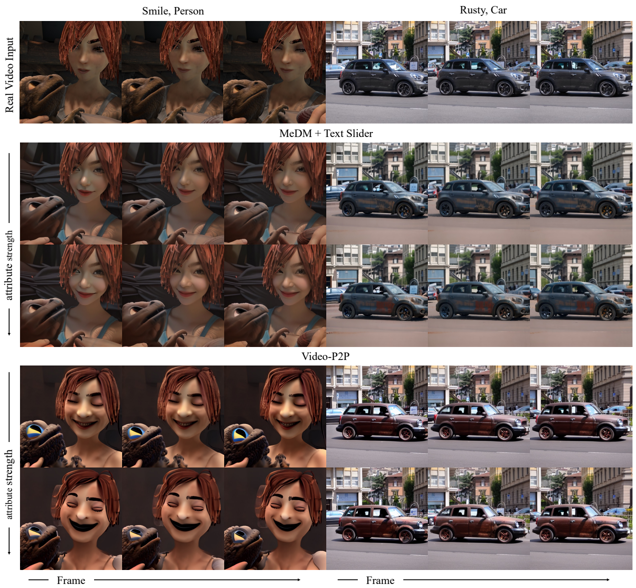

Figure 4. **Comparison of Video-to-Video Results.** MeDM [2] combined with Text Slider enables fine-grained concept manipulation, whereas Video-P2P [10] exhibits limited attribute controllability, particularly for facial attributes such as smile and age.

sity while preserving the spatial layout of the content. In contrast, prompt-based methods often struggle to maintain structural coherence and offer limited controllability.

Video-to-Video Generation. To assess the effectiveness of Text Slider within a video-to-video translation framework, we integrate it with MeDM [2], a zero-shot video editing method based on image diffusion. MeDM perturbs real video frames using SDEdit [11] and applies an image diffusion model in a frame-by-frame manner. As shown in Figure 4, we select two attributes, smile on a person and rusty effect on a car, to examine the performance of combining MeDM with Text Slider, and compare it against Video-P2P [10]. Our method achieves fine-grained attribute modulation while preserving spatial structure and temporal consistency. In contrast, Video-P2P performs poorly on facial attributes and often distorts the original content. Moreover, Video-P2P requires time-consuming, per-video model tuning, whereas Text Slider offers a plug-and-play solution without any additional fine-tuning.

Composing Sliders. In Figure 5, we demonstrate the qualitative results of composing multiple Text Sliders in text-to-video generation using AnimateDiff [5]. By sequentially applying the surprised, glasses, and old attributes, our method preserves structural consistency at each stage while effectively modulating the intended concepts.

3.2. Comparison with Concept Slider

As shown in Table 1, compared to Concept Slider, our method significantly reduces both the number of parameters and the training time while achieving comparable performance in CLIP scores. Moreover, unlike Concept Slider, which requires separate training for each base model (*e.g.*, SD-1.5 and SD-XL), Text Slider generalizes naturally across different architectures that share the same text encoder, such as SD-1.5 and SD-XL, without the need for retraining. This highlights the superior adaptability and efficiency of our approach. We also present qualitative results in Figure S.1 and S.2 of supplementary material to further validate the generalization ability of Text Slider on SD-XL.

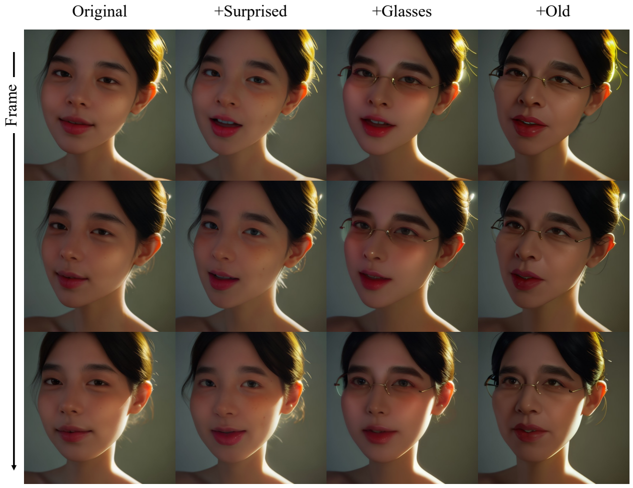

Figure 5. **Slider Composition.** The figure demonstrates the composability of Text Slider in text-to-video generation, enabling sequential attribute modifications while preserving structural consistency and effectively controlling target concepts.

SD-1.5	CLIP-s (↑)	#Params(M) (↓)	Time (min) (↓)
Concept Slider [4]	**26.50**	2.91	30
Text Slider (Ours)	25.97	**0.66**	**3**
SD-XL	CLIP-s (↑)	#Params(M) (↓)	Time (min) (↓)
Concept Slider [4]	26.44	4.32	45
Text Slider (Ours)	**26.88**	**0.66**	**3**

Table 1. **Comparison with Concept Slider.** Text Slider significantly reduce $\approx$ 77% of the parameters and $\approx$ 90% of training time in SD-1.5, and $\approx$ 85% of parameters and $\approx$ 93% of training time in SD-XL while achieving comparable performance.

4. Conclusion

We propose Text Slider, a precise concept control method that is efficient, highly adaptable, plug-and-play, and composable. Text Slider significantly reduces more than 77% of both the number of parameters and the training time required to learn a slider. Moreover, it generalizes across different Stable Diffusion architectures without retraining, whereas Concept Slider requires separate training for each model. Additionally, our approach naturally extends to text-to-video and video-to-video generation, enabling precise and continuous concept control.

Acknowledgement

This research is supported by National Science and Technology Council, Taiwan (R.O.C) under the grant numbers NSTC-113-2634-F-002-007, NSTC-112-2222-E-001-001-MY2, 113-2634-F-001-002-MBK, 113-2634-F-002-008 and Academia Sinica under the grant number of AS-CDA-110-M09, AS-IAIA-114-M08, and AS-IAIA-114-M10.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 1

[2] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1353–1361, 2024. 1, 4

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[4] 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*, pages 172–188, 2024. 2, 4, 1

[5] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 1, 3, 4

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 1

[9] 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. In *International Conference on Learning Representations*, 2022. 2

[10] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 2, 4, 1

[11] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 4, 1

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

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

Text Slider: Efficient and Precise Concept Control for Video Generation and Editing via LoRA Adapters

Supplementary Material

Slider	**Input Prompt**	**Edited Prompt**	**Checkpoint**
Smile	“face photo of a person, moving, light tone, delightful, bright”	“face photo of a **smiling** person, moving, light tone, delightful, bright”	majicmixRealisticV2V25
Alien	“photo of a car, moving”	“photo of an **alien** car, moving”	realistic-vision-V5.1-noVAE
Winter	“photo of coastline, rocks, sunny weather, wind, waves, lightning, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3”	“photo of a **winter** coastline, rocks, sunny weather, wind, waves, lightning, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3”	majicmixRealistic-V2V25
Cartoon	“A funny and charming robot exploring a futuristic city”	“A funny and charming **cartoon** robot exploring a futuristic city”	realistic-vision-V5.1-noVAE

Table S.1. Sliders, prompts, and model checkpoints used in qualitative experiments for text-to-video generation.

A. Implementation Details

All Text Sliders are trained for 500 epochs using the AdamW optimizer with a learning rate of 2×10^{-4} and bfloat16 precision. The LoRA rank is set to $r = 4$, and the guidance scale is $\eta = 4$. In all experiments, LoRA modules are applied to every layer of the clip-vit-large-patch14 text encoder, including the projection layers within the self-attention and MLP blocks. For video generation tasks, we follow the structure-preserving strategy of SDEdit [11]. Specifically, LoRA adapters are disabled during the initial denoising steps by setting their multipliers to 0 until timestep $t = 700$, after which the adapters are activated for the remaining steps.

A.1. Text-to-Video Generation

For the qualitative comparison in Figure 3 of the main paper, we generate videos using AnimateDiff [5] in combination with the prompts and pre-trained checkpoints listed in Table S.1. These prompts span a diverse set of subjects, human, vehicle, scene, and stylized character, highlighting the generality and fine-grained controllability of our method in various video generation scenarios.

To apply Text Slider, we integrate it into the text encoder of AnimateDiff and perform inference using a scaling factor $\alpha = 0 \sim 0.5$ for each target concept. For prompt-based baseline comparisons, concept control is attempted by directly inserting attribute keywords before the subject noun, as shown in the “Edited Prompt” column of Table S.1.

A.2. Video-to-Video Generation

For qualitative evaluation of real video editing, we generate results using MeDM [2] combined with Text Slider, and compare them against the baseline method, Video-P2P [10]. To integrate MeDM with Text Slider, we insert the slider into the text encoder and perform inference with a scaling factor $\alpha = 0 \sim 0.5$ for the corresponding concept. For Video-P2P, we use the ”Input Prompt” and “Edited Prompt” entries from Table S.2 to perform attribute-specific video editing.

Slider	**Input Prompt**	**Edited Prompt**
Smile	”a person cuddle a little creature”	”a **smiling** person cuddle a little creature”
Rusty	”a car”	”a **rusty** car”

Table S.2. Sliders and prompts used in qualitative experiments for video-to-video generation.

SD-1.5	Age	Smile	Muscular	Curly hair	Winter weather
Concept Slider [4]	27.80	**29.38**	20.03	**28.55**	**26.73**
Text Slider (Ours)	**28.76**	29.06	**20.74**	25.35	25.96
SD-XL	Age	Smile	Muscular	Curly hair	Winter weather
Concept Slider [4]	29.85	**29.18**	**25.46**	26.42	21.31
Text Slider (Ours)	**30.51**	27.66	23.80	**27.70**	**24.71**

Table S.3. Comparisons of CLIP scores for individual concept.

A.3. Comparison with Concept Slider

For the quantitative comparison, we compute the CLIP score over 5,000 image-prompt pairs using the clip-vit-large-patch14 encoder to embed both images and prompts. Specifically, we select five distinct attributes, as listed in Table S.4, and generate 1,000 images per attribute to evaluate the alignment between visual output and textual descriptions. We also report the CLIP scores for individual attributes in Table S.3.

B. Limitation

We observe that Text Slider exhibits limited controllability when targeting specific objects within a scene. For instance, in the presence of multiple objects, it often struggles to apply attribute manipulation to a designated object. This highlights a critical direction for future work: enhancing the precision of concept control to allow attribute manipulation localized to specific objects or regions.

Slider	**Input Prompt**	**Evaluation Prompt**
Age	”image of a person, photorealistic”	”image of an **old** person, photorealistic”
Smile	”image of a person, photorealistic”	”image of a person with **smile**, photorealistic”
Muscular	”image of a person, photorealistic”	”image of a **muscular** person, photorealistic”
Curly hair	”image of a person, photorealistic”	”image of a person with **curly hair**, photorealistic”
Winter weather	A bustling city street, with people walking	A bustling **winter** city street, with people walking

Table S.4. Sliders and prompts used in quantitative comparison with Concept Slider.

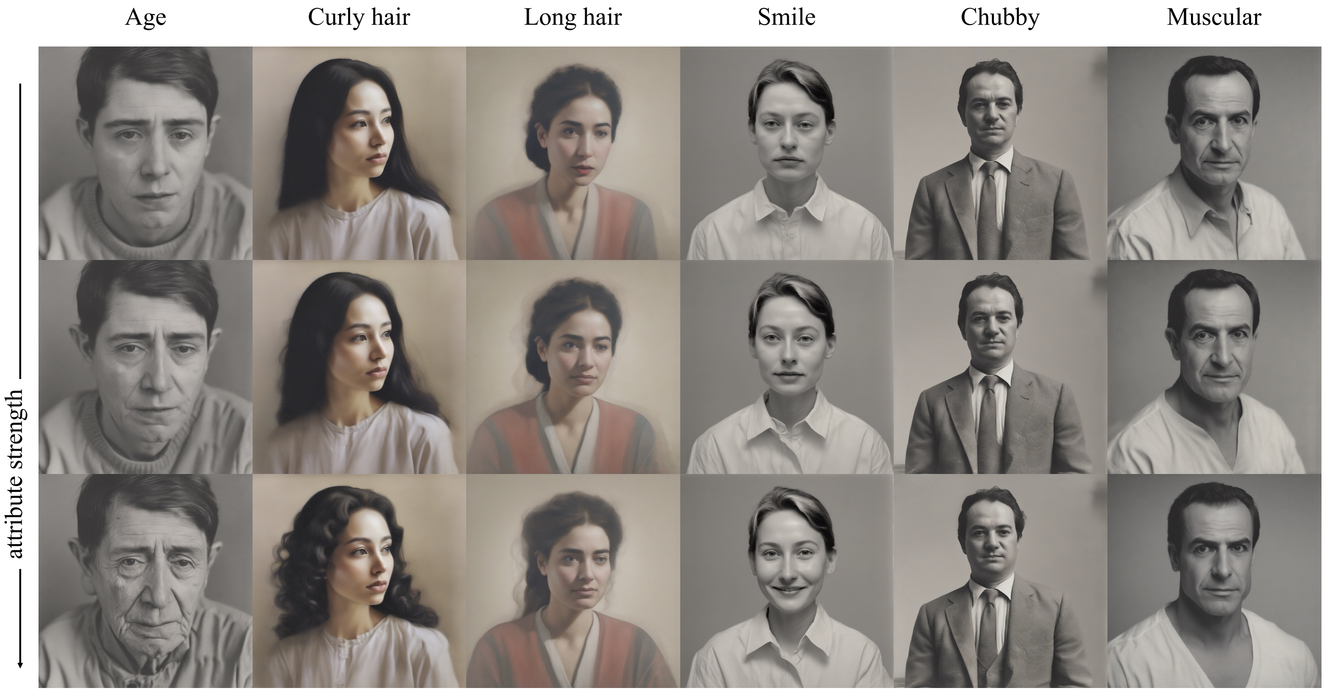

Figure S.1. **Qualitative Results of Text Slider on SD-XL.** The figure showcases Text Slider’s ability to modulate human-related attributes with varying strengths using SD-XL.

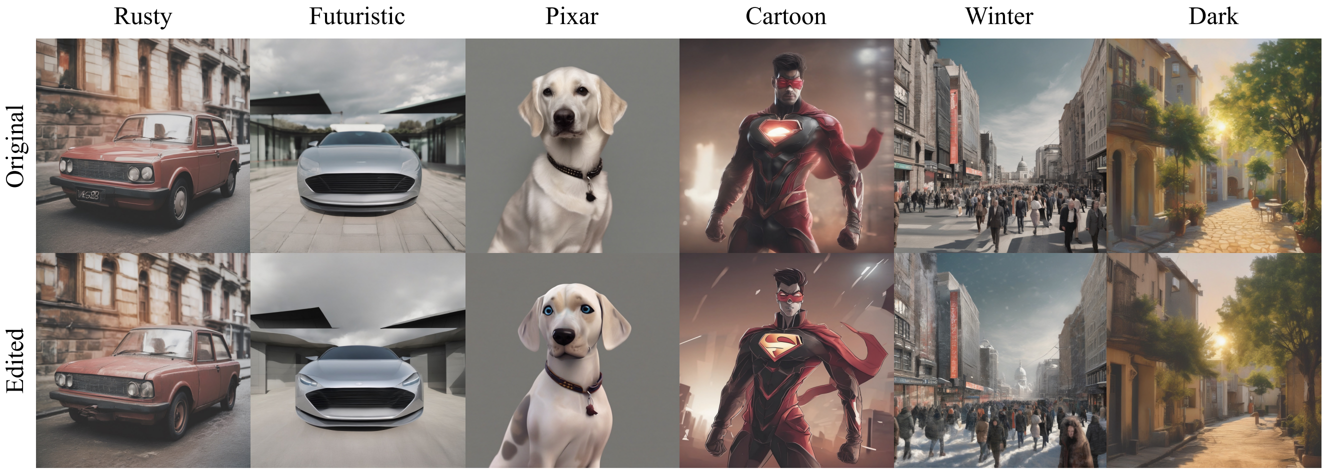

Figure S.2. **Qualitative results of Text Slider on SD-XL.** The figure showcases Text Slider’s capability with SD-XL in modulating car, style, and scene attributes across varying strengths.