CompSlider: Compositional Slider for Disentangled Multiple-Attribute Image Generation

Supplementary Material

In the following section, we first introduce additional details for our foundation model and additional implementation details to provide a deeper understanding of how our slider model operates. We then discuss limitations of our method and explore the impact of threshold selection in structure loss to highlight its role in balancing structure consistency and attribute control. This is followed by a discussion on feature behavior to shed light on how our method effectively guides attribute generation. Finally, we present details of our user studies and showcase additional qualitative results to further validate the effectiveness of our approach.

6. Additional Details on the T2I Foundation Model

The foundation model is a U-Net [45] that predicts the noise ϵ between the noisy image $\boldsymbol{x}_t^{\mathcal{I}}$ at timestep t and the denoised image $\boldsymbol{x}_{t-1}^{\mathcal{I}}$ at timestep $t-1$. The single-step inference process is defined as

$$\boldsymbol{x}_{t-1}^{\mathcal{I}} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t^{\mathcal{I}} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\text{U-Net}(\boldsymbol{x}_t^{\mathcal{I}}, \boldsymbol{c}^{\mathcal{I}}, \boldsymbol{c}^{\mathcal{T}}, t)\right) + \sigma_t \boldsymbol{z}, \tag{12}$$

where α_t and $\bar{\alpha}_t$ are the variance schedule terms, U-Net$(\boldsymbol{x}_t^{\mathcal{I}}, \boldsymbol{c}^{\mathcal{I}}, \boldsymbol{c}^{\mathcal{T}}, t)$ predicts the noise at time t, σ_t is the standard deviation for the stochastic noise term, and $\boldsymbol{z} \sim \mathcal{N}(0, I)$ is sampled from a normal distribution.

7. Additional Implementation Details

We trained CompSlider using 8 A100 GPUs, and the entire training process took about 16 hours for 20000 iterations. The batch size is set to 2048, and the learning rate is initially warmed up to 1×10^{-4} over 500 steps, then gradually decreased to 1×10^{-7} following a cosine annealing schedule. Our 16 sliders include: "Age", "Smile", "Surprise", "Sadness", "Anger", "Brown Hair", "Blond Hair", "Gray Hair", "Black Hair", "Red Hair", "Yaw Rotation", "Pitch Rotation", "Beard", "Tone", "Vector Style", and "Scene Complexity". During inference, our CompSlider generates multiple image conditions based on different slider values, which are then input into the foundation model alongside a text prompt to produce the final image results. To ensure consistency across different image conditions for the same text prompt, we keep both the initial noise and the sampled noise in the denoising process of the foundation model identical. Additionally, we found that adding noise (*e.g.*, at an intensity of around 75%) to the classifier-free guidance [20] image results, and then using these noise-added images as the initial noise in the foundation model, improves the consistency of details across outputs from different slider values.

8. Limitations.

When multiple subjects share the same attribute, our method lacks precise target selection, as it applies attribute changes uniformly across all subjects. This limitation arises because the model does not inherently distinguish which subject should be edited. A potential solution to this issue is enhancing the model's text reasoning capabilities, allowing it to better interpret textual instructions and selectively apply modifications to the intended target.

9. Impact of Threshold Selection in Structure Loss

Table 4. Ablation Study on the threshold in our structure Loss.

Threshold	Continuity%↑	Consistency%↑	Scope%↑
0.5	64.68	**96.44**	25.50
0.3	77.48	92.79	57.15
0.1	**81.07**	90.95	**59.02**

To demonstrate how the threshold in our structure loss affects the slider model's ability to maintain identity, we conducted a search over different threshold values. Tab. 4 shows that increasing the threshold improves structural consistency. However, a larger threshold comes at the cost of reduced continuity and a narrower adjustment scope, as the structure loss regularization causes the model to produce the same image condition over a wider range of slider values, weakening the sliders' control over attributes.

10. Extension to Video Generation

Figure 9 illustrates how CompSlider extends to text-to-video generation using video foundation models. Our method successfully controls the specified attribute consistently across frames, demonstrating its ability to generalize beyond static images.

11. Discussion on Feature Behavior

To further illustrate how our CompSlider generates image conditions and controls the foundational model to produce images with specific attributes, we visualized the cross attention between the generated image conditions and the

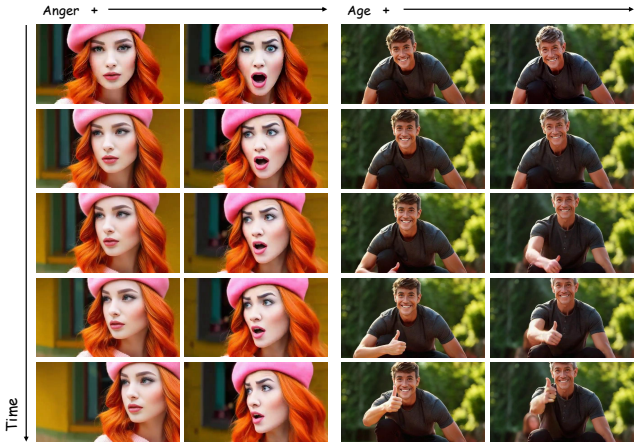

Figure 9. Applying our image slider conditional priors to a video generation model maintains effective control and identity.

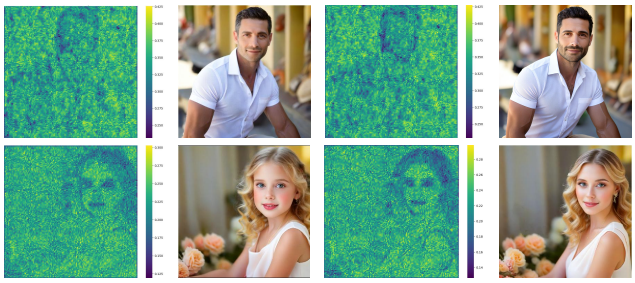

Figure 10. Cross attention in the foundational model between the image conditions generated by CompSlider and the noised image.

noised image within the foundational model. Fig.10 demonstrates how the attention maps from the beard and age sliders guide the model in generating the corresponding output.

12. User Studies Details

To evaluate our slider method against the current state-of-the-art (SOTA), we conducted an A/B test focusing on smoothness and structural consistency in image transitions. For each test instance, participants viewed a specific slider type (indicated above the images as either "vector style" or "number of objects") and two rows of images generated by different methods. The slider intensity increased progressively from left to right, and the order of methods was randomized between the top and bottom rows. Participants selected the row with the smoothest transition and best structural consistency, choosing "First Row," "Second Row," or "Tie." Results were saved to record user preferences across examples. The interface is shown in Fig. 11. Similar to common evaluation settings [22, 31, 55], we conducted a user study with approximately seven participants, all of whom are experts in text-to-image generation research.

13. Additional Qualitative Results

We show additional qualitative examples for sliders in Fig. 12 and Fig. 13.

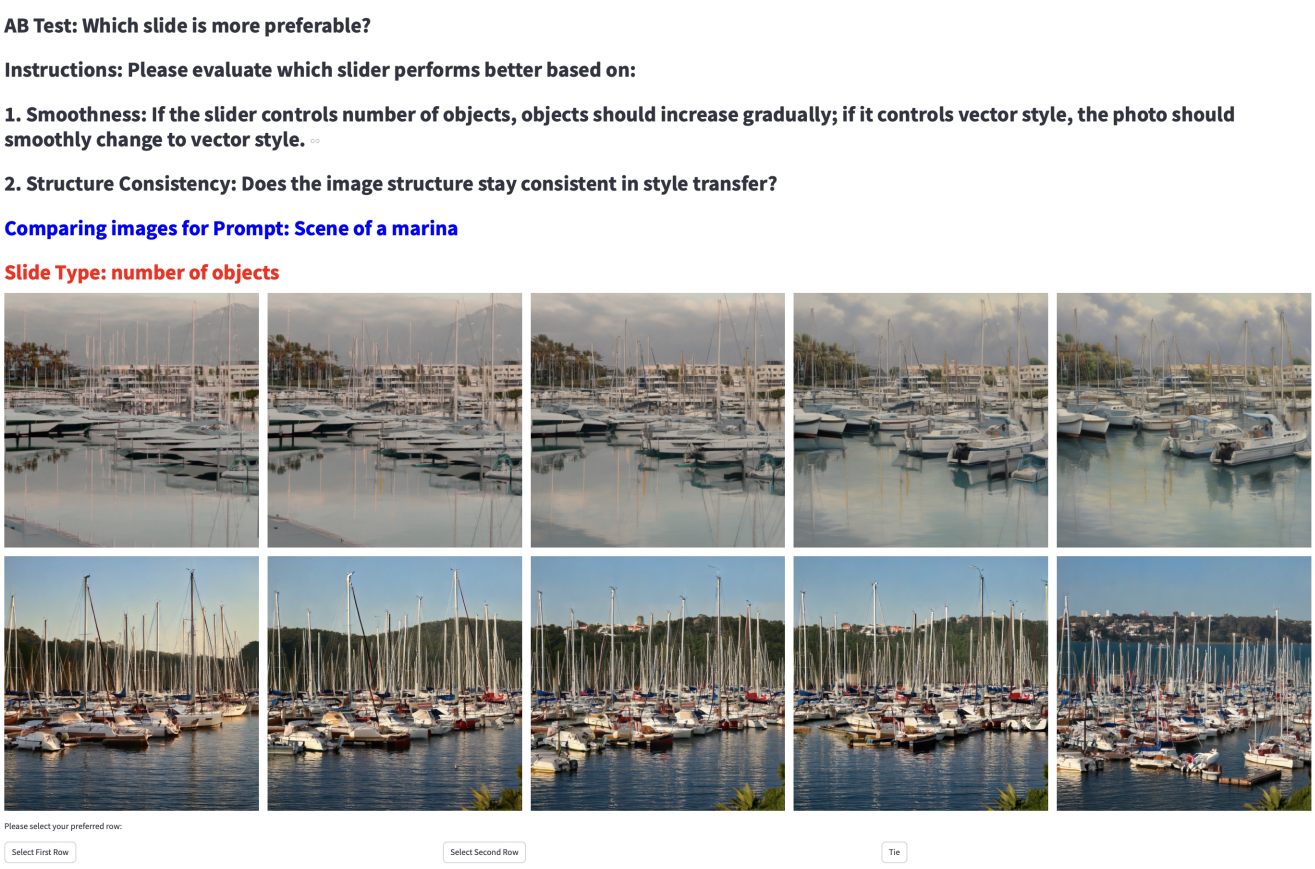

Figure 11. Interface for our A/B test.

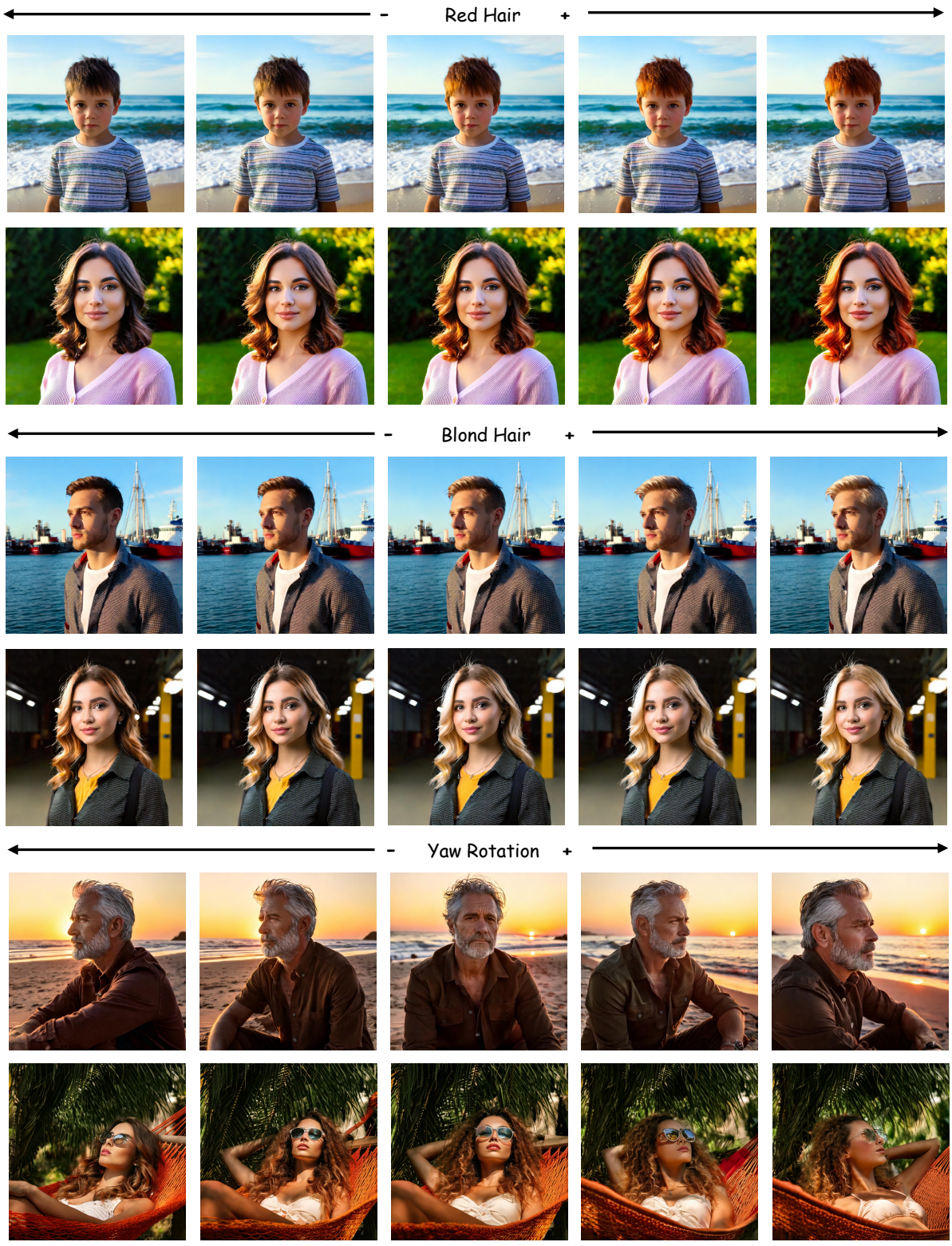

Figure 12. Slider generation results for “Red Hair”, “Blond Hair”, and “Yaw Rotation”.

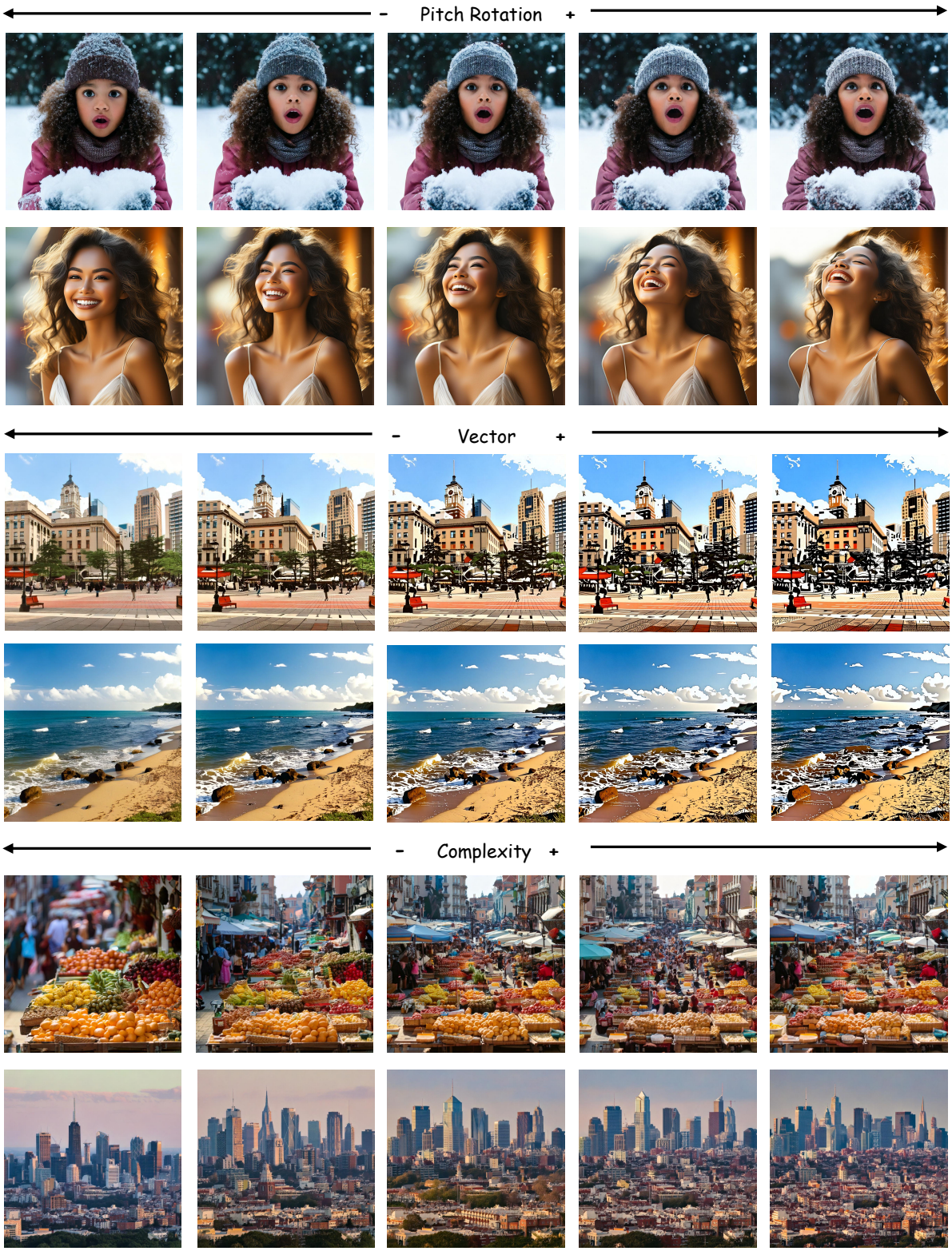

Figure 13. Slider generation results for “Pitch Rotation”, “Vector Style”, and “Complexity”.