Supplementary Material: Leveraging Semantic Attribute Binding for Free-Lunch Color Control in Diffusion Models

1. Color binding regions as a function of steps

Figure 1 displays color binding regions based on $\delta_\lambda(\tilde{\mathbf{A}}_o)$ every 10-th denoising step,

Figure 1. Color binding regions based on $\delta_\lambda(\tilde{\mathbf{A}}_o)$. Depicted every 10-th denoising step.

2. Multi-object

ColorWave successfully applies different colors to multiple objects within the same scene by leveraging separate attention maps for each target object token, as demonstrated in Figure 2. As seen, the method achieves a robust color control. Potential failure cases may arise when the object has faulty or misrepresented attention maps, therefore rendering the color binding difficult or unfaithful.

3. Experimental details

Evaluation metric details For each generated image, we use the Segment-Anything model [3] to automatically extract masks for the target objects, allowing precise measurement of color attributes only within relevant regions. We report these metrics under multiple conditions: considering all pixels within the object mask (100%), and also considering the top 10% and 50% of pixels closest to the target color, which helps account for natural variations in object appearance such as highlights and shadows.

Implementation details We implement *ColorWave* using a pretrained Stable Diffusion XL (SDXL) [4] model as our base architecture. For the image conditioning component, we utilize the IP-Adapter framework [7] with an encoder based on OpenCLIP-ViT-H-14 [2].We only inject the color embedding into *the first decoder layer* to compute cross-attention maps, which shows better *stylization* performance as previous works proved [1, 5, 6].

Figure 2. Multi-object color control examples. (first) "*A green rose growing from a red vase on a table*", (second) "*A woman wearing blue pants sitting on a white sofa*", (third) "*A blue teddy bear in Times Square on top of a pink car*", (fourth) "*An orange parrot with a purple hat perched on a tree*", (fifth) "*A photo of a blue dog on a yellow surfboard surfing a wave wearing a orange lifevest*".

For the adapter masking, we keep the top 20% largest values on each object map. During inference, we process the user-specified RGB values that creates temporary color reference images. These references are encoded by the IP-Adapter and strategically injected into the model's cross-attention layers.

4. Color Leakage Metric

To quantify unintended color attribution in generated images, we introduce a color leakage metric that measures the percentage of pixels in non-target regions that exhibit colors similar to the specified target color. This metric is particularly important for evaluating whether color control methods successfully confine the desired color to the intended object without affecting other parts of the image.

The color leakage metric operates by first extracting the reference RGB color from the image filename, then converting both the reference color and the generated image to HSV color space. For each pixel in the image, we calculate the circular hue difference between the pixel's hue and the reference hue, accounting for the wraparound nature of the hue channel at $180°$. Pixels with hue differences within a threshold of 10 degrees are considered to match the target color.

5. Mechanism of Semantic Attribute Binding

As depicted in Figure 3, the semantic attribute binding phenomenon arises from IP-Adapter's decoupled cross-

attention design, where shared query projection matrices create implicit alignment between visual and textual feature spaces. We quantify this binding by computing the inner product similarity $\langle \mathbf{K}_{\text{text}}, \mathbf{K}_{\text{image}} \rangle$ between text prompt tokens and image feature tokens.

The heatmap reveals systematic correspondence patterns across color combinations, with diagonal peaks indicating that the model implicitly learns to associate linguistic color terms with their visual counterparts. The token-level analysis demonstrates that color words (e.g., "pink") exhibit peak similarity with corresponding visual color features, while non-color tokens maintain lower similarities. This mechanism generalizes across the continuous color space, with perceptually similar colors showing elevated cross-similarities that reflect the model's understanding of color relationships.

6. Extra results of *ColorWave*

Extra results are showcased in Figures 4 to 11.

References

[1] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *ICML*, 2024. 1

[2] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open clip, 2021. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1

[4] 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. In *ICLR*, 2024. 1

[5] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 1

[6] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 1

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1

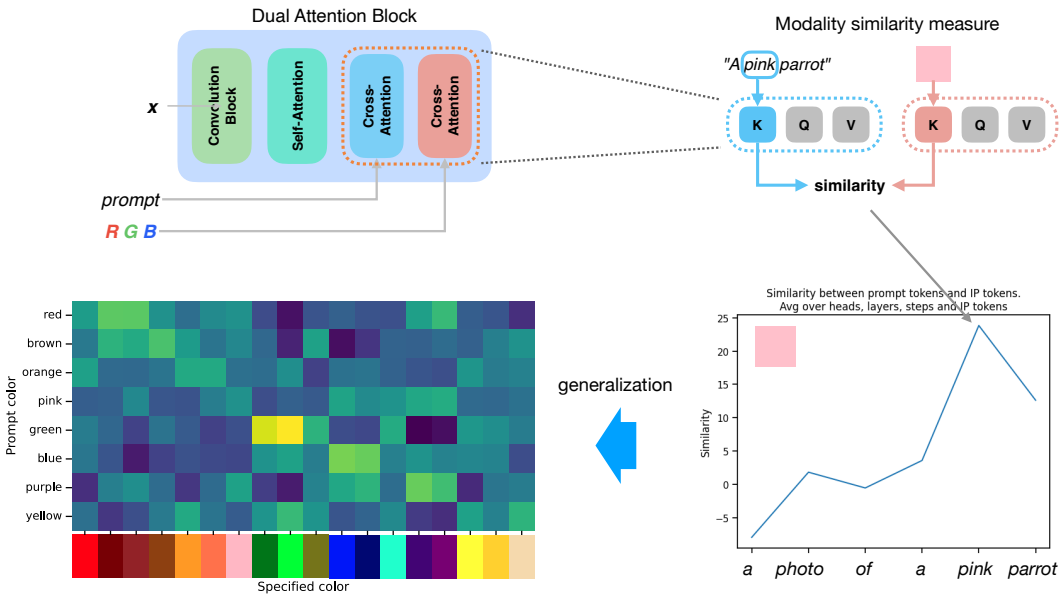

Figure 3. The dual cross-attention architecture enables implicit alignment between text and image features through shared query projections. The similarity heatmap (bottom left) shows correspondence patterns between prompt colors and target colors, with diagonal peaks indicating successful color-semantic binding. The token-level similarity plot (bottom right) demonstrates how individual prompt tokens relate to image features, with color words showing peak correspondence to matching visual color information.

Figure 4. ”A photo of a red -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 5. ”A photo of a red -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 6. ”A photo of a pink -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 7. ”A photo of a orange -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 8. ”A photo of a purple -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 9. ”A photo of a green -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 10. ”A photo of a blue -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)

Figure 11. ”A photo of a blue -object-” - Reference color is depicted in the external frame of the grid - Objects same as in ColorPeel method (see qualitative evaluation)