Supplementary Material for SemanticControl: A Training-Free Approach for Handling Loosely Aligned Visual Conditions in ControlNet

Woosung Joung*
mung3477@korea.ac.kr
Daewon Chae*
cdw098@korea.ac.kr
Jinkyu Kim
jinkyukim@korea.ac.kr

Korea University
Seoul, Republic of Korea

1 Dataset

Mild Conflicts. Although Liu et al. [2] didn't release the evaluation dataset, we referred to their qualitative results and collected similar prompt-control pairs. We used control images of 12 subjects and various prompts that have a minor conflict with the image, including animals (dog, cat, horse, pig, deer, wolf, squirrel, hedgehog, eagle, parrot, duck), cartoon characters (hulk, spider-man, jerry, pikachu, micky mouse), fruits (apple, strawberry, pineapple, pear), vehicles (car, pick-up truck, tractor, bike, motorcycle), objects (cup, vase, candle), celebrities (obama) and landmarks (Pyramid, Arch of Triumph, Colosseum in Rome).

Significant Conflicts. We focused on significant conflicts between the subject of the text prompt and the visual condition—human and non-human. We collected 15 unique actions of humans, by gathering images from stock image services and previous works that investigated the problems related to generating human actions [1, 2]. Evaluation was conducted with 11 different non-humans (a dog, a dog plushie, a cat, a panda, a teddy bear, a polar bear, a squirrel, an elephant, a fox, a gorilla, an owl).

Surrogate Prompts. Surrogate prompts should be aligned with the given visual condition. We constructed surrogate prompts for our dataset by substituting the subject of the prompt that we are trying to generate. Another key component of our method is the *non-conflicting tokens*. We selected those tokens from the surrogate prompt with the following rules: (i) remove stopwords, (ii) remove tokens that are unrelated to the desired context, and (iii) exclude conflicting tokens. Examples of images and the corresponding prompts are provided on Figure 1.

* denotes equal contribution.

2 Implementation Details

Excluding Special Tokens. We empirically found that special tokens such as $< sot >$ and $< eot >$ tend to receive disproportionately high attention values. Their dominance can suppress the attention assigned to other tokens, making it difficult to assess the relative importance of meaningful semantic tokens. Thus, before aggregating those maps to construct a control scale mask, we excluded those special tokens and re-normalized the softmax over the remaining tokens to obtain a clearer distribution of token-level influence.

Control Scale for Blocks without a Cross-attention Operator. Since our method is built on the pipeline of Stable diffusion [3], the first upsampling block of the decoder in UNet does not contain cross-attention operators. Since this block also requires an inferred control scale, we empirically chose to use the control scale mask of the middle block of the UNet as the alternative. This was because both blocks have the same input image resolution.

Control Bias λ. We use a control bias of $\lambda = 3.0$ in our main experiments. We also report results with $\lambda = 1.0$ in Table 1, where our method still outperforms SmartControl.

Table 1: Quantitative results with different control bias λ.

Methods	Visual Condition: Depth Map			
	CLIP(↑)	BLIP(↑)	ImageReward(↑)	PickScore(↑)
SmartControl	0.3186	0.4735	0.8574	0.2271
SemanticControl (λ=0.1)	0.3295	0.4856	1.0473	0.2298
SemanticControl (λ=0.3)	0.3322	0.4904	1.1538	0.2304

3 Additional results

We provide additional generated examples for significant conflicts using various condition types—depth maps (Figure 2), canny edges (Figure 3), human skeletons (Figure 4), and normal maps (Figure 5). Since SmartControl does not support normal maps and human skeletons, we only compare the result with ControlNet for those cases.

4 Details on Human evaluation

As mentioned in the main paper, we conduct a human evaluation to compare our method against baseline approaches in terms of overall preference. For each visual condition, we evaluate 1,005 generated image pairs using Amazon Mechanical Turk, collecting responses from five independent raters per image. We aggregate the results through majority voting. Figure 6 presents the actual instruction and a sample question shown to each rater.

References

[1] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation, 2024. URL https://arxiv.org/abs/2311.15841.

[2] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

[3] 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.

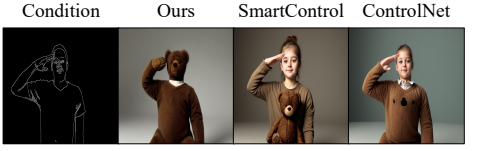
(a) Condition Images and Prompt Formats for Mild Conflicts

(a) Condition Images and Prompt Formats for Significant Conflicts

Figure 1: Condition images and prompts for each level of conflict. We constructed the surrogate prompts for each condition by substituting only the subject of the prompt. **Bold** indicates non-conflicting tokens of the surrogate prompt. {*subject*} tokens are used as conflicting tokens for significant conflicts, and we did not use conflicting tokens for mild conflicts.

Figure 2: Additional examples generated with the depth maps. The same random seed is applied across methods for each prompt.

Prompt: *"A **gorilla** is deadlifting"*



Prompt: *"A **teddy bear** is saluting"*

Prompt: *"A **panda** is cooking vegetables"*

Prompt: *"A **teddy bear** is shooting a handgun"*

Prompt: *"A **panda** is holding a sword"*

Prompt: *"A **teddy bear** is stretching its arms left and right"*

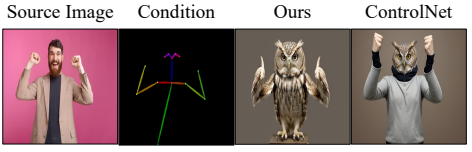
Prompt: *"A **teddy bear** is holding a sword"*

Prompt: *"An **owl** is cooking vegetables"*

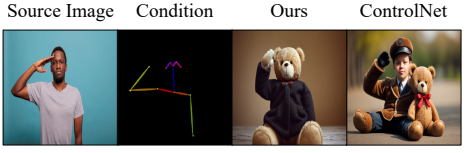
Prompt: *"A **teddy bear** is playing the trumpet"*

Prompt: *"A **teddy bear** with both arms gesture"*

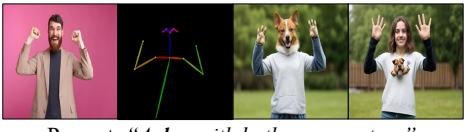
Figure 3: Additional examples generated with the edge maps. The same random seed is applied across methods for each prompt.



Prompt: *"An **owl** with both arms gesture"*



Prompt: *"A **teddy bear** is saluting"*



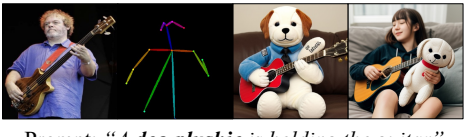
Prompt: *"A **dog** with both arms gesture"*

Prompt: *"A **polar bear** is riding a bike"*

Prompt: *"A **fox** is cooking vegetables"*

Prompt: *"A **dog** is toasting a cocktail to you"*

Prompt: *"A **fox** is playing the trumpet"*



Prompt: *"A **dog plushie** is holding the guitar"*

Prompt: *"A **panda** is playing the trumpet"*

Prompt: *"A **teddy bear** is holding a sword"*

Figure 4: Additional examples generated with the human skeleton of the given image. The same random seed is applied across methods for each prompt. Although generated images followed the visual condition well, detailed structures are different to the given image due to the nature of the human pose.

Figure 5: Additional examples generated with the normal map of the given image. The same random seed is applied across methods for each prompt.

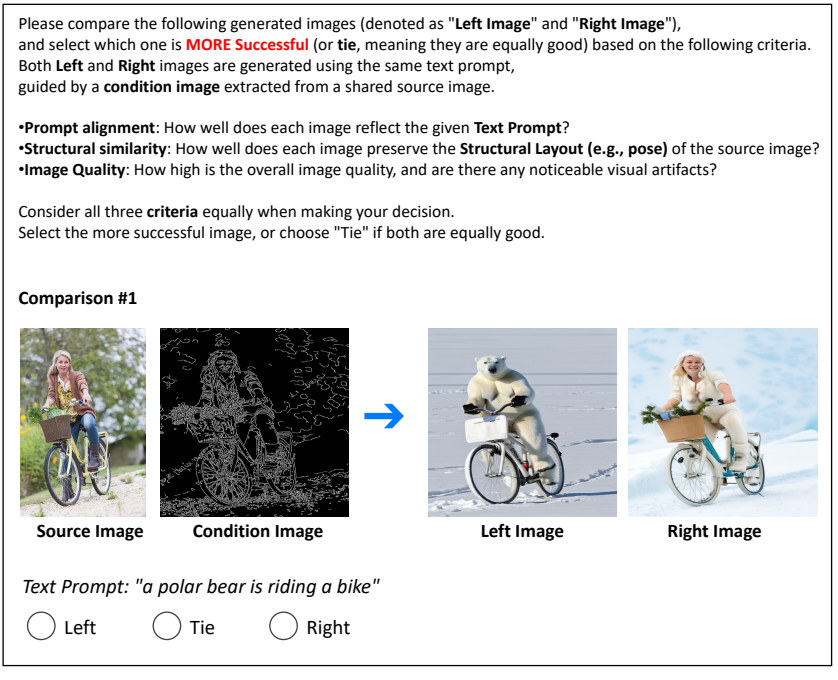


Figure 6: Human survey interface: instructions and an example questionnaire.