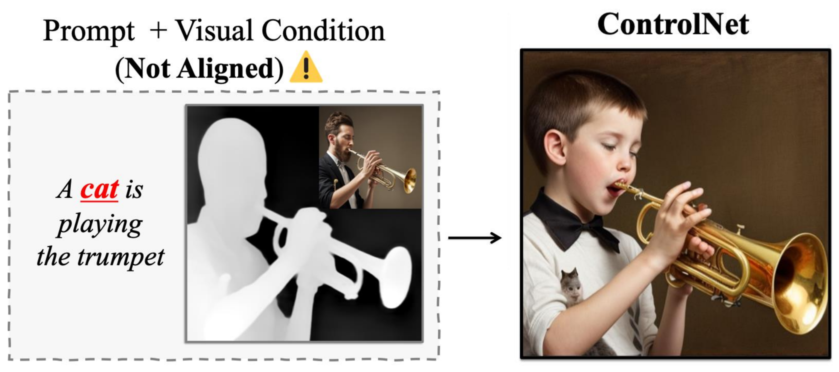
SemanticControl: A Training-Free Approach for Handling Loosely Aligned Visual Conditions in ControlNet

Woosung Joung* , Daewon Chae*, Jinkyu Kim
Korea University

BMVC 2025

Motivation

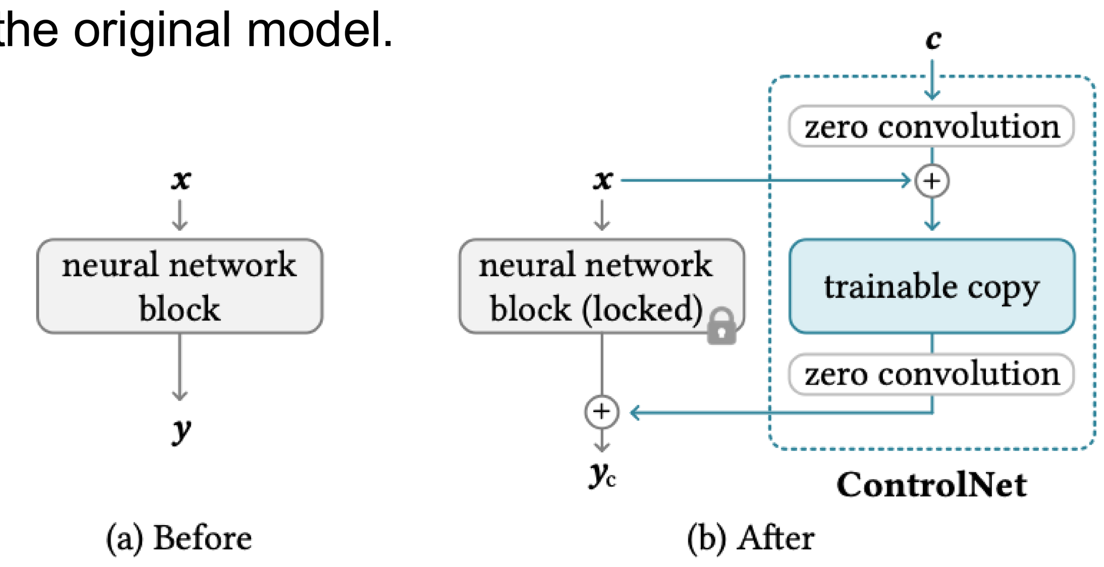
ControlNet's effectiveness is constrained by the need for visual conditions that are precisely aligned with the semantics of the target prompt.



Obtaining such aligned conditions is often infeasible; particularly on uncommon scenarios.

ControlNet

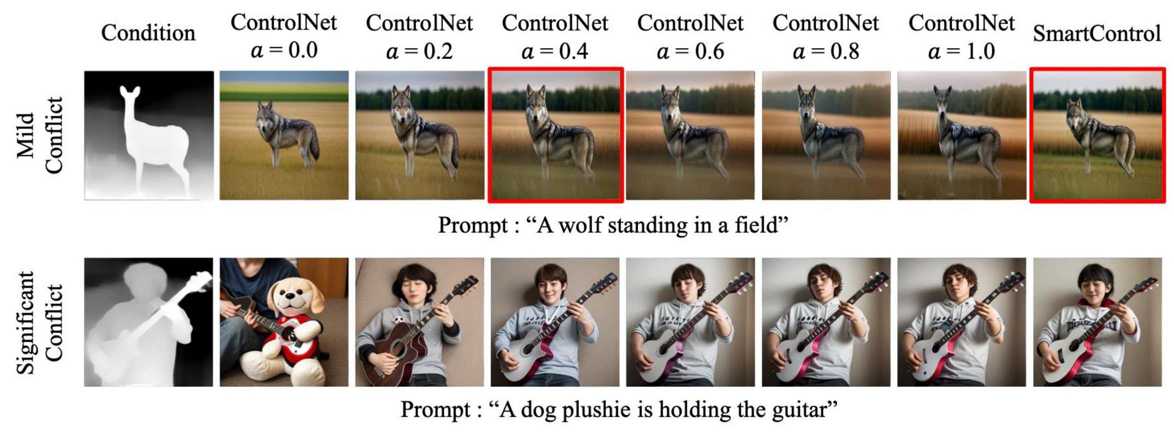
ControlNet [1] injects additional conditions into the blocks of pretrained text-to-image model using the trainable copy of the original model.



(a) Before (b) After

Relaxing conflicting visual condition

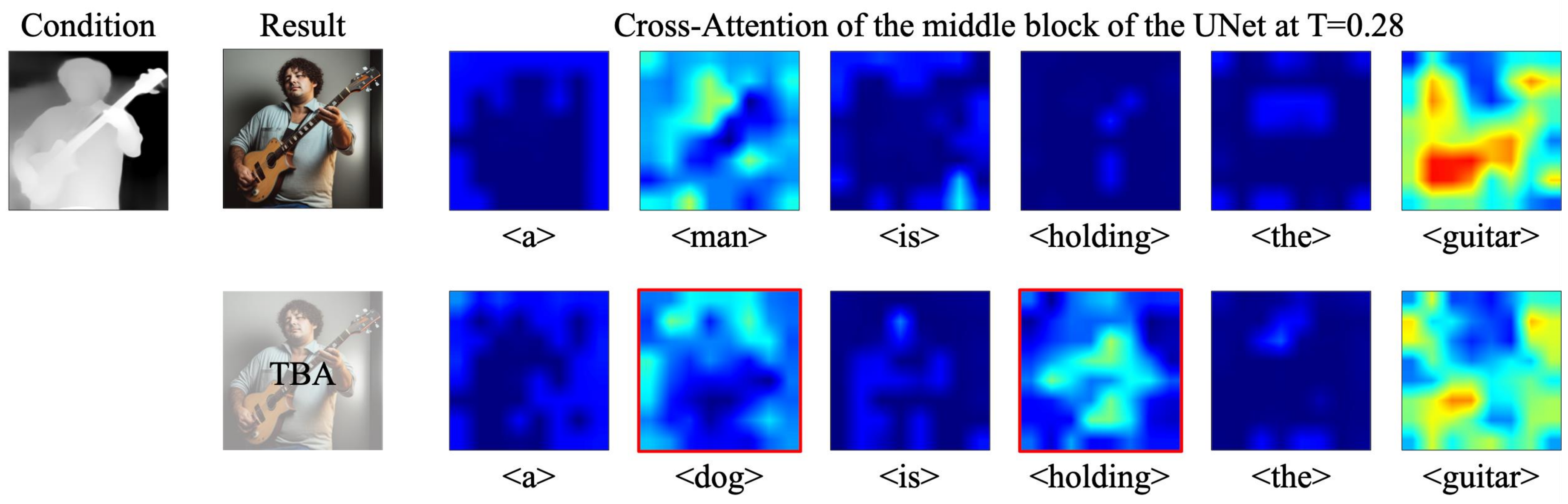
SmartControl [2] trains a network to identify the conflict regions and predict the local control scales.



However, under severe conflict, there is no control scale to learn from—training-free approach is needed.

Cross-Attention Maps from ControlNet

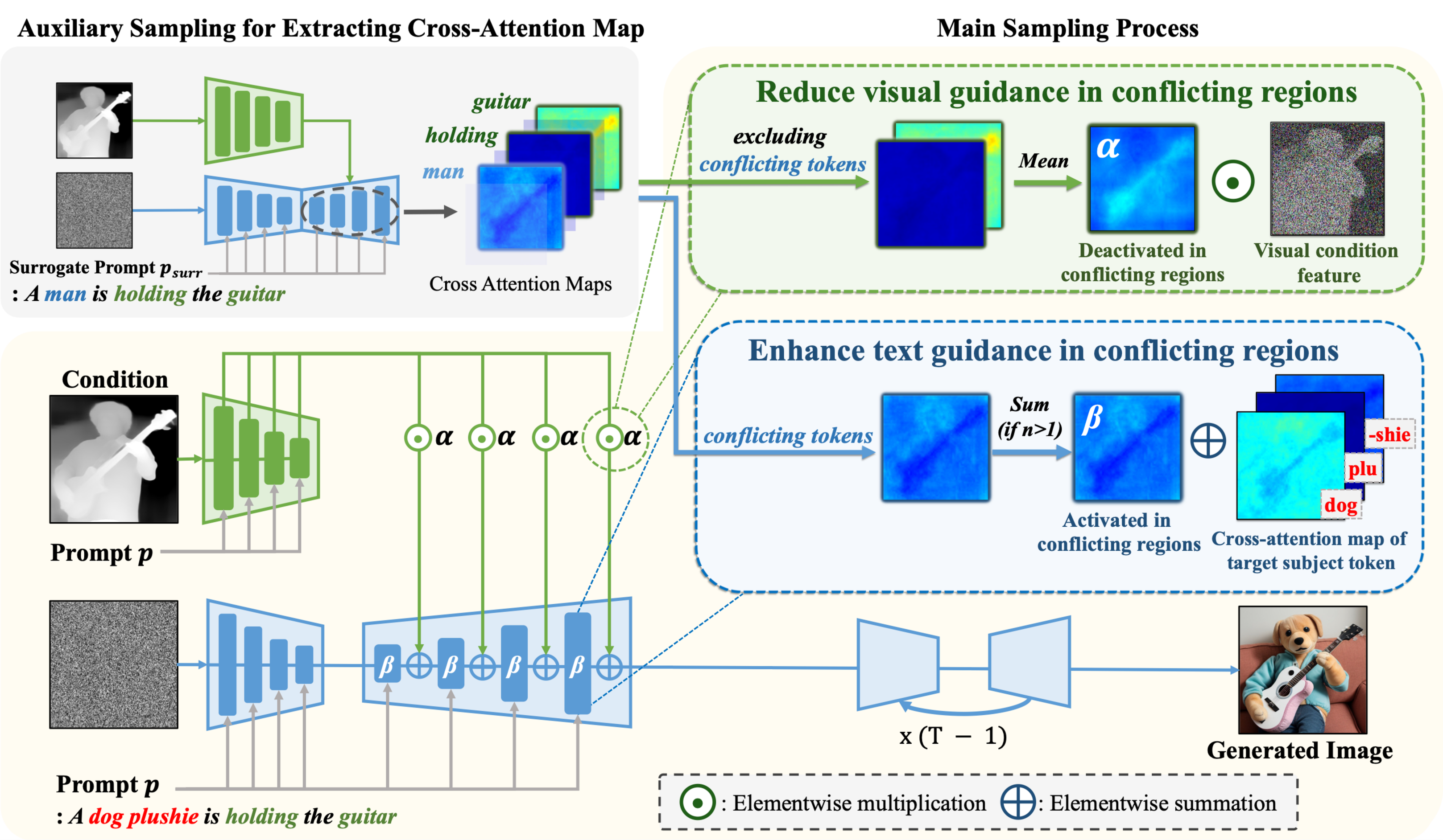
Cross-attention maps reflect how much each text token contributes to specific spatial regions in the generated image.

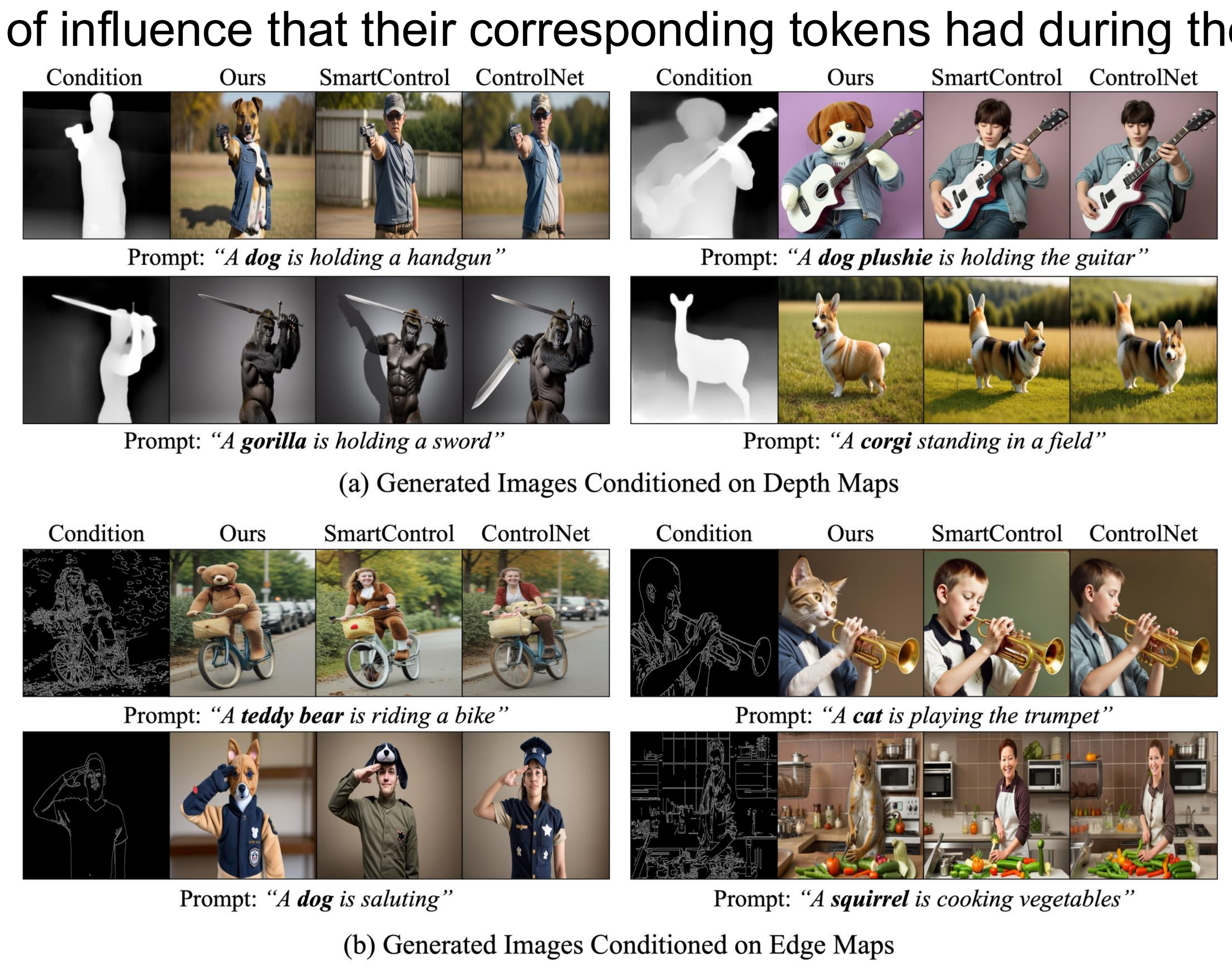


- ❑ Misaligned outputs from ControlNet implies that the resulting attention maps may not meaningfully reflect how much attention each token should receive.
- ❑ Important tokens from the text prompt that contradict the condition, such as *<dog>* on the second row, are receiving insufficient cross-attention.
- ❑ Our idea is to leverage cross-attention maps from text prompts aligned with the visual condition, which reliably reflect the intended contribution of each token.

Overall pipeline of SemanticControl

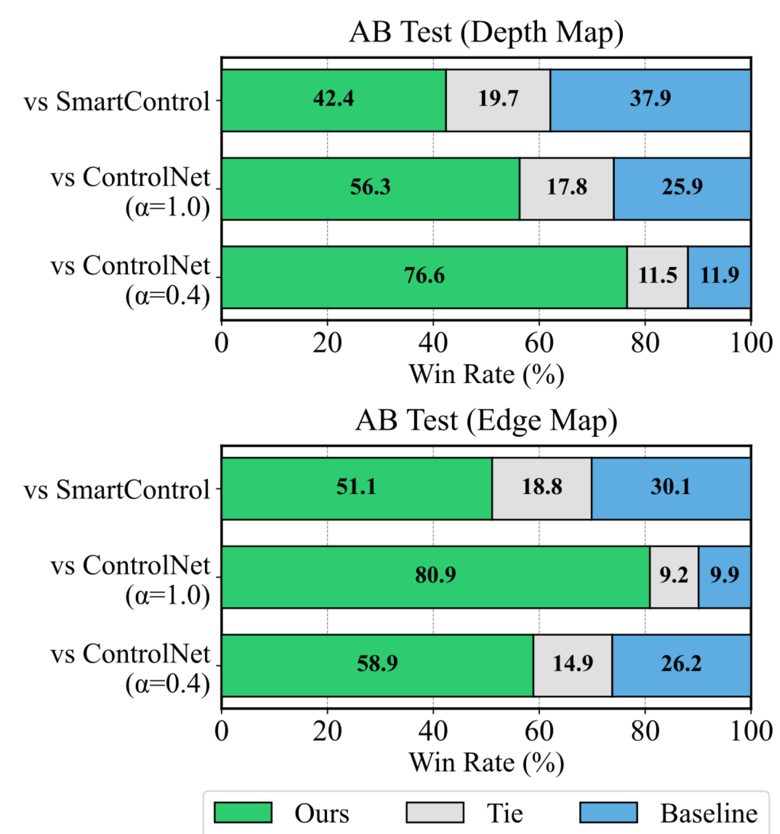
- ❑ **Control scale mask** from surrogate prompt p_{surr}, where conflicting tokens are replaced, suppresses conflicting regions in the ControlNet feature.
- ❑ **Cross-attention bias** ensures that target tokens retain at least the level of influence that their corresponding tokens had during the preliminary sampling.





(a) Generated Images Conditioned on Depth Maps

(b) Generated Images Conditioned on Edge Maps

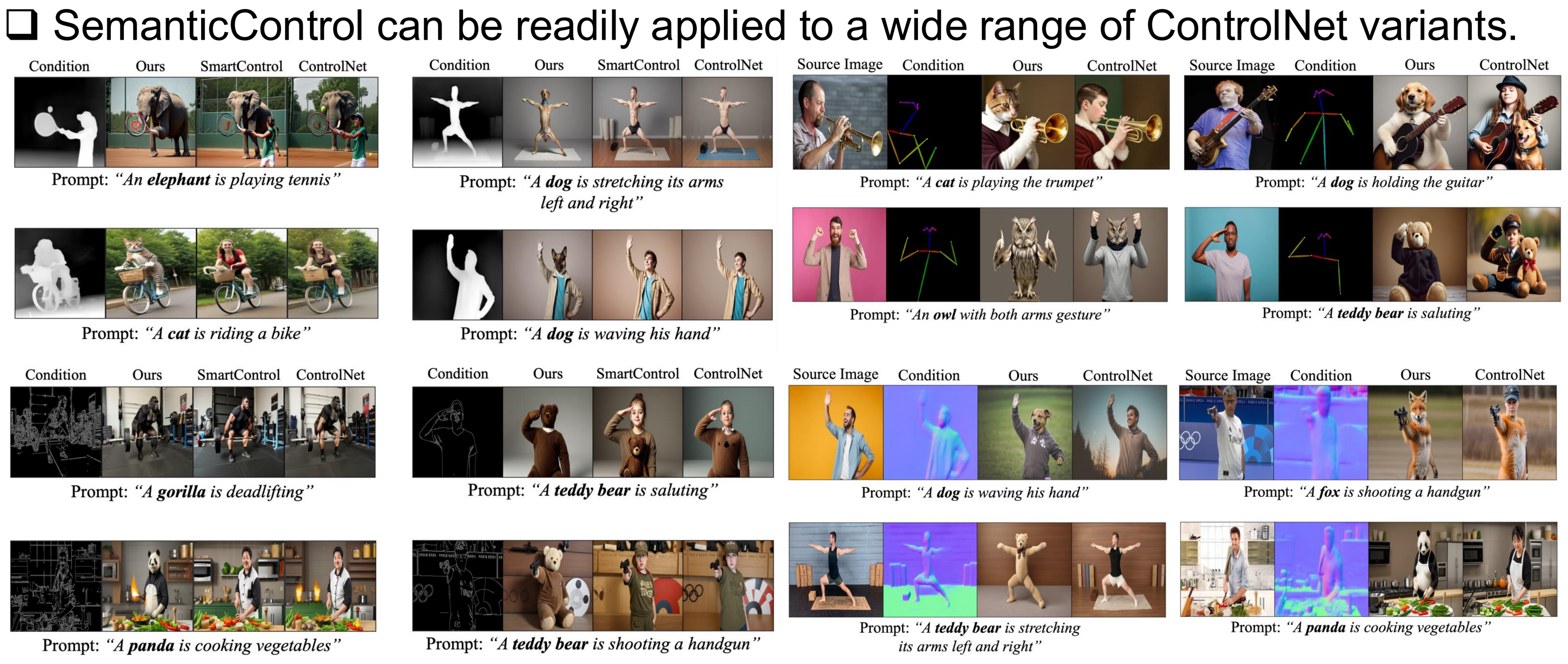


Experiments

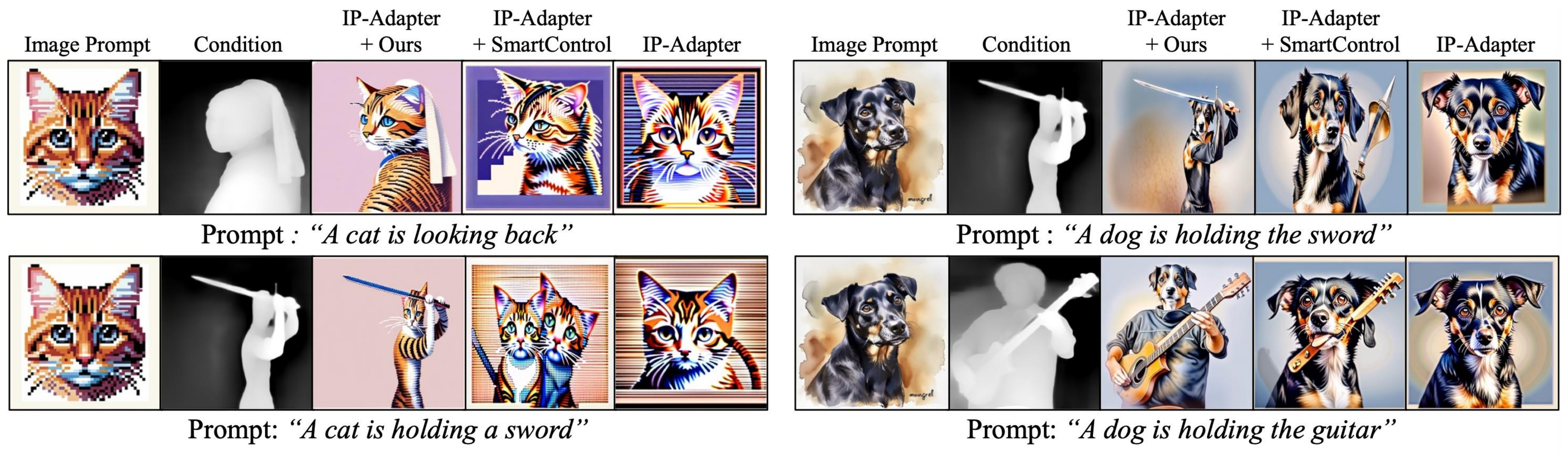
❑ Quantitative comparison between our method and baselines.

Methods	Visual Condition: Depth Map				Visual Condition: Canny Edge			
	CLIP(↑)	BLIP(↑)	ImageReward(↑)	PickScore(↑)	CLIP(↑)	BLIP(↑)	ImageReward(↑)	PickScore(↑)
ControlNet (α=1.0)	0.2878	0.4283	-0.3063	0.2155	0.2741	0.4042	-0.7893	0.2105
ControlNet (α=0.4)	0.3146	0.4633	0.5744	0.2248	0.3129	0.4604	0.5488	0.2242
SmartControl	0.3186	0.4735	0.8574	0.2271	0.3127	0.4647	0.5679	0.2246
SemanticControl w/o attention bias	<u>0.3274</u>	<u>0.4813</u>	<u>0.9428</u>	<u>0.2291</u>	<u>0.3265</u>	<u>0.4797</u>	<u>0.8919</u>	<u>0.2285</u>
SemanticControl (Ours)	**0.3322**	**0.4904**	**1.1538**	**0.2304**	**0.3302**	**0.4864**	**1.0289**	**0.2292**

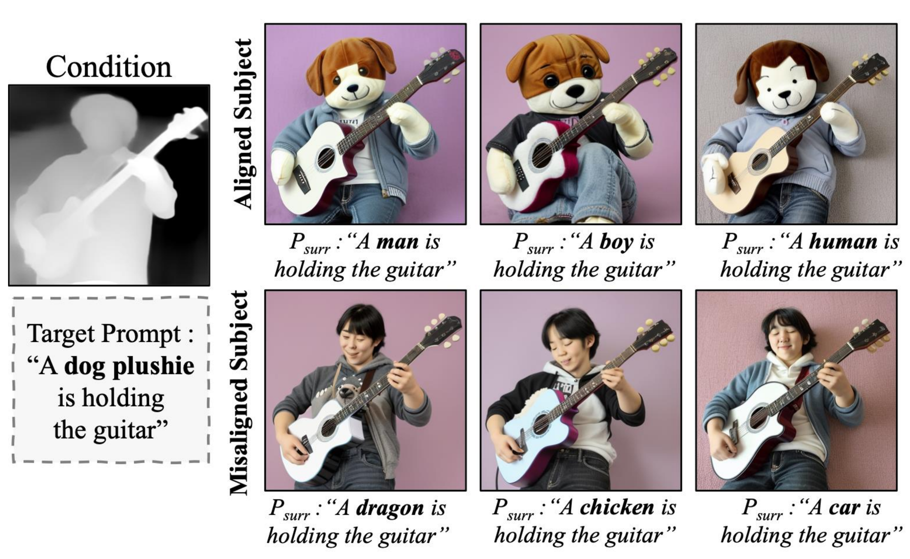
❑ SemanticControl can be readily applied to a wide range of ControlNet variants.



❑ SemanticControl can also be utilized with IP-Adapter.



❑ Ablation study on the surrogate prompt and each component in our method.



(a) Analysis on Choice of Surrogate Prompt

(b) Contribution of Each Module

Acknowledgments

This work was supported by IITP grant funded by the Korea government(MSIT) (IITP2025-RS-2024-00397085, 30%, RS-2025-02263754, Human-Centric Embodied AI Agents with Autonomous Decision-Making, 30%, IITP-2025-RS-2020-II201819, 10%, RS-2022II220043, Adaptive Personality for Intelligent Agents, 30%). Daewon Chae is supported by the Hyundai Motor Chung Mong-Koo Foundation.