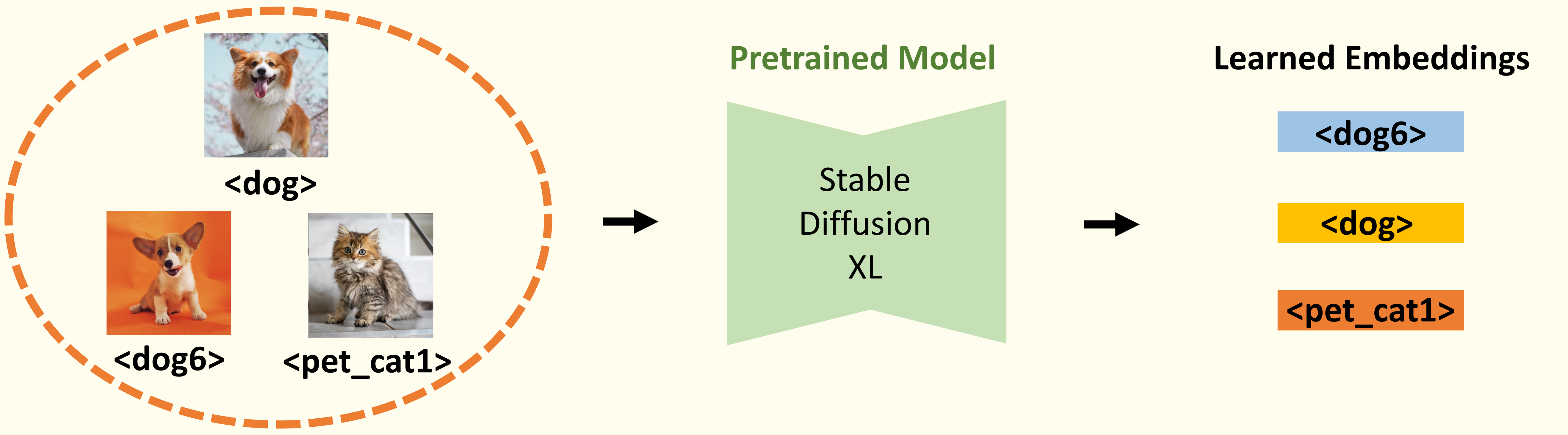
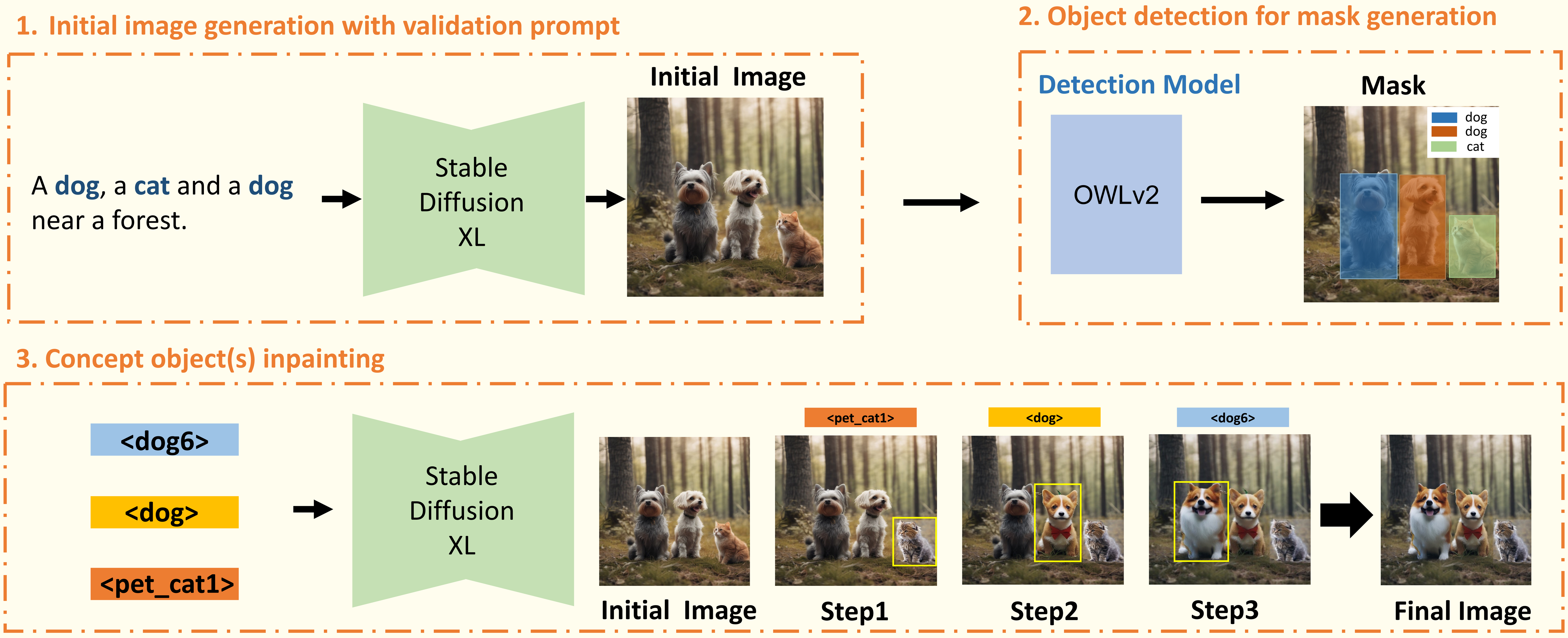
OneCAT – One Concept A Time

Bo-Rui Chen[1], Chun-Huan Chou[1], Shao-Xiang Yuan[1], Ye-Shin Yang[2]
(B10901028, B10901025, B11901055, B10508026)
[1]National Taiwan University Department of Electrical Engineering
[2]National Taiwan University Department of Biomedical Engineering

(1) Textual Inversion Training



(2) Custom Image Generation Pipeline



Introduction

Current multi-concept personalization methods suffer from **blending similar concepts**, such as cats and dogs, which results in mixed features or the omission of certain concepts. To address this issue, we propose One Concept-A-Time (One CAT), **a method that separates the generation of each concept to avoid blending while maintaining the natural and harmonious appearance of the entire image**.

Methods

Our pipeline consists of three steps:

- **Draft Generation**: Similar to how humans sketch, we first create a "draft" using dummy tokens (e.g., "cat") to represent each concept.
- **Mask Generation**: Using a fast zero-shot object detection method, we generate binary masks to identify the regions of interest (bounding boxes of the dummy objects).
- **Inpainting**: With the masks, we perform inpainting to sequentially replace the dummy objects with the refined concepts generated via textual inversion.

"A <dog>, a <dog6>, a <cat2>" and a <cat> near a forest.

Figure1. We can also generate four distinct concepts in the image.

Results

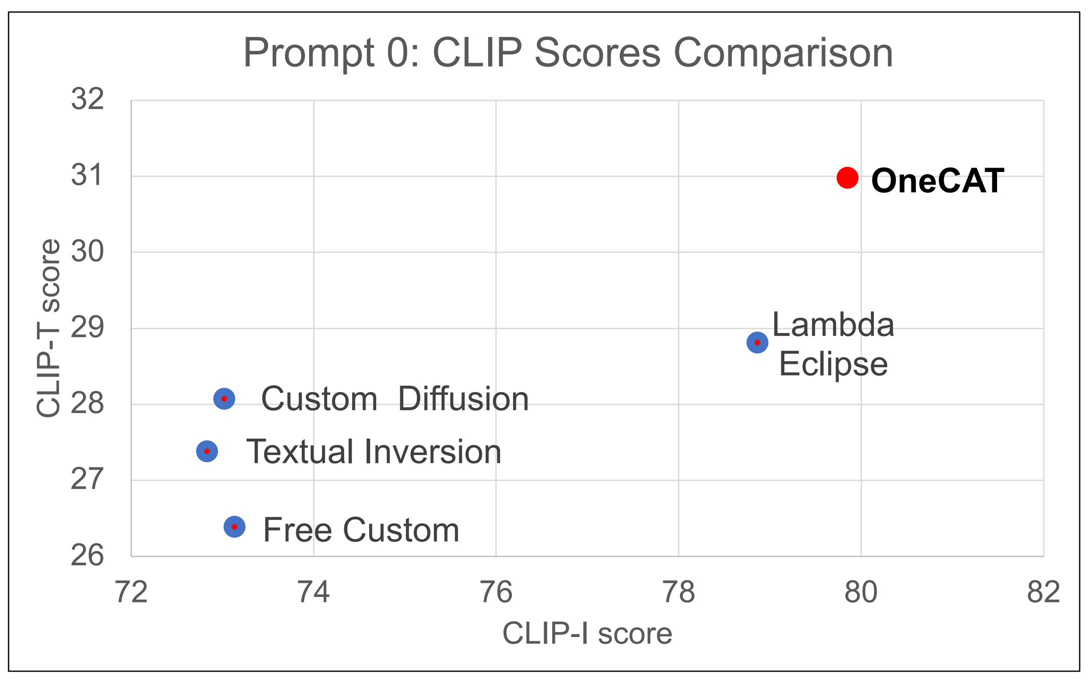


Figure2. CLIP score comparison across different methods

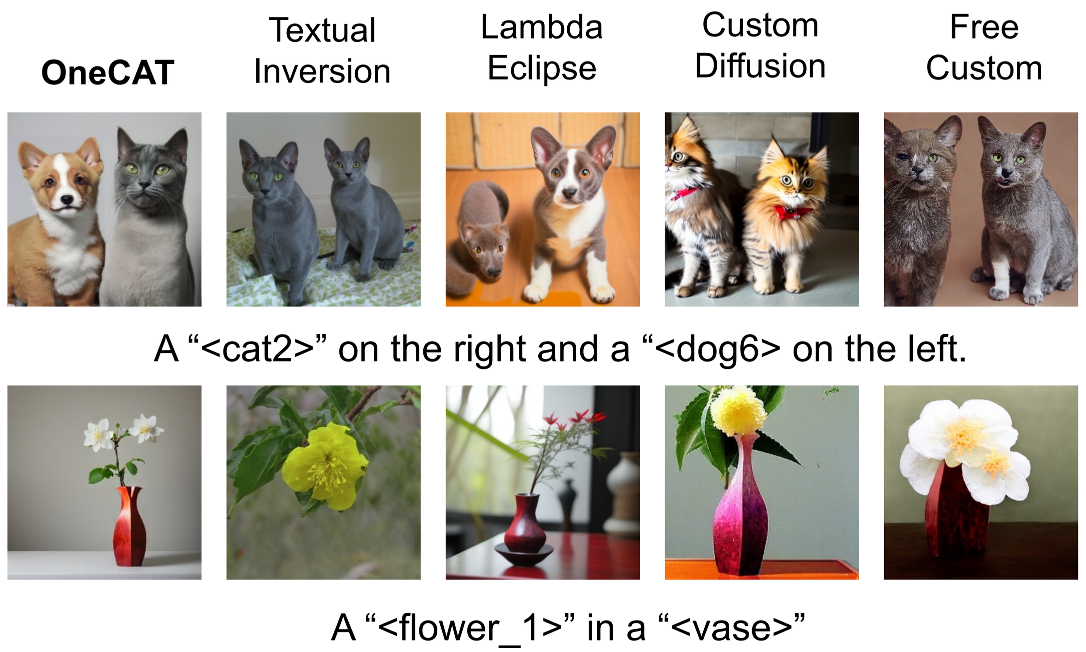


Figure3. Comparison between each methods

Ablation Studies

- **Only use textual inversion**

Figure4. (left) Generate two cats instead of a cat and a dog.
Figure5. (middle,right) Attention map for stable diffusion2 model.

- **Using segmentation instead of detection**

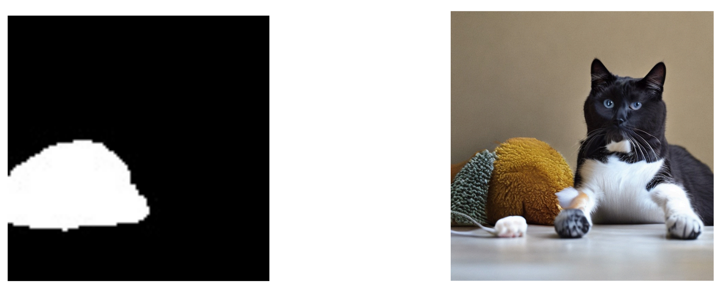


Figure6. (left) Segmentation result for the dog.
Figure7. (Right) Inpainting based on segmentation results.

- **Adjusting Inpainting Hyperparameters.**

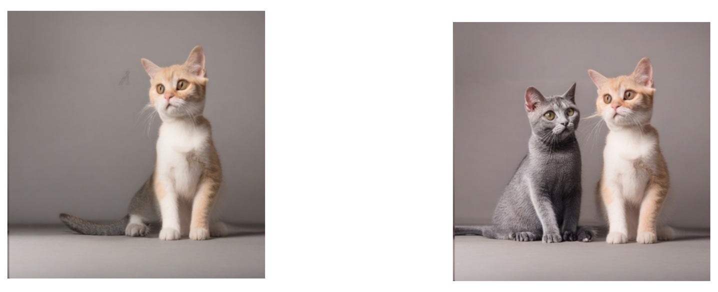


Figure8. w/o and w/ Mask padding crop and Negative prompts

Discussions

Advantages

Our Method	Effect
Object detection	Preventing merging between concepts
Inpainting	Concepts blend seamlessly with the scene
Not model-specific	Can use any diffusion model with its textual inversions

Limitations

Cause	Effect	Improvement
Complex Commands	Concepts vanish	Find better dummy tokens
Inpainting Failure	Concepts vanish	Negative prompts and mask padding
Multiple Inpainting steps	Slower inference time	Inpaint simultaneously for disjoint masks

Reference:
Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).
Ding, Ganggui, et al. "FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
Podell, Dustin, Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis." arXiv preprint arXiv:2307.01952 (2023).
Patel, Maitreya, et al. "$\lambda $-ECLIPSE: Multi-Concept Personalized Text-to-Image Diffusion Models by Leveraging CLIP Latent Space." arXiv preprint arXiv:2402.05195 (2024).
Minderer, Matthias, Alexey Gritsenko, and Neil Houlsby. "Scaling open-vocabulary object detection." Advances in Neural Information Processing Systems 36 (2024).