AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

UNIVERSITI MALAYA

Yuan: Yielding Unblemished Aesthetics through A Unified Network for Visual Imperfections Removal in Generated Images

Zhenyu Yu, Chee Seng Chan
Universiti Malaya

Motivation & Background

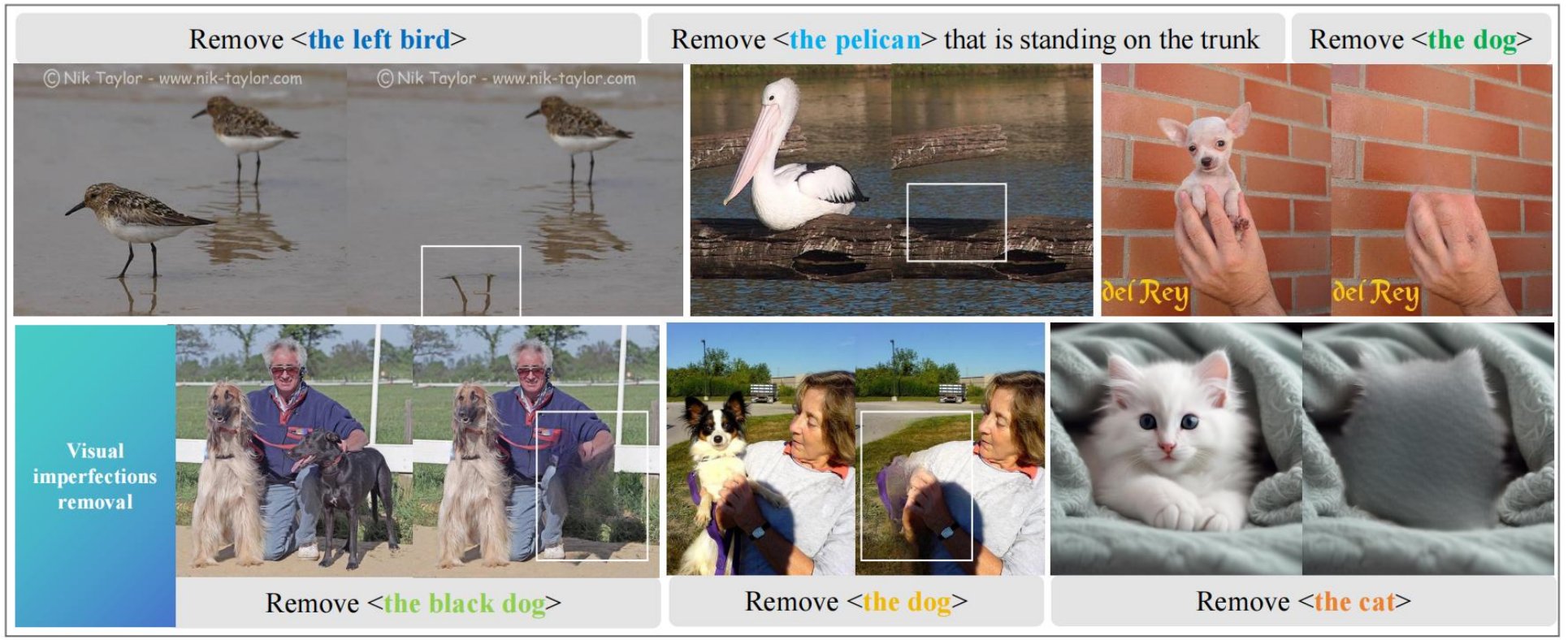


Figure 1: **Motivation for the study:** Existing algorithms for target removal often fall short in addressing related elements, such as reflections and shadows, resulting in incomplete or unnatural outcomes. Additionally, the removal of specified content can leave behind visual inconsistencies, such as unnatural postures or actions, necessitating further corrections. These challenges underscore the need for more advanced methods to achieve coherent and realistic image modifications.

Yuan Framework Overview

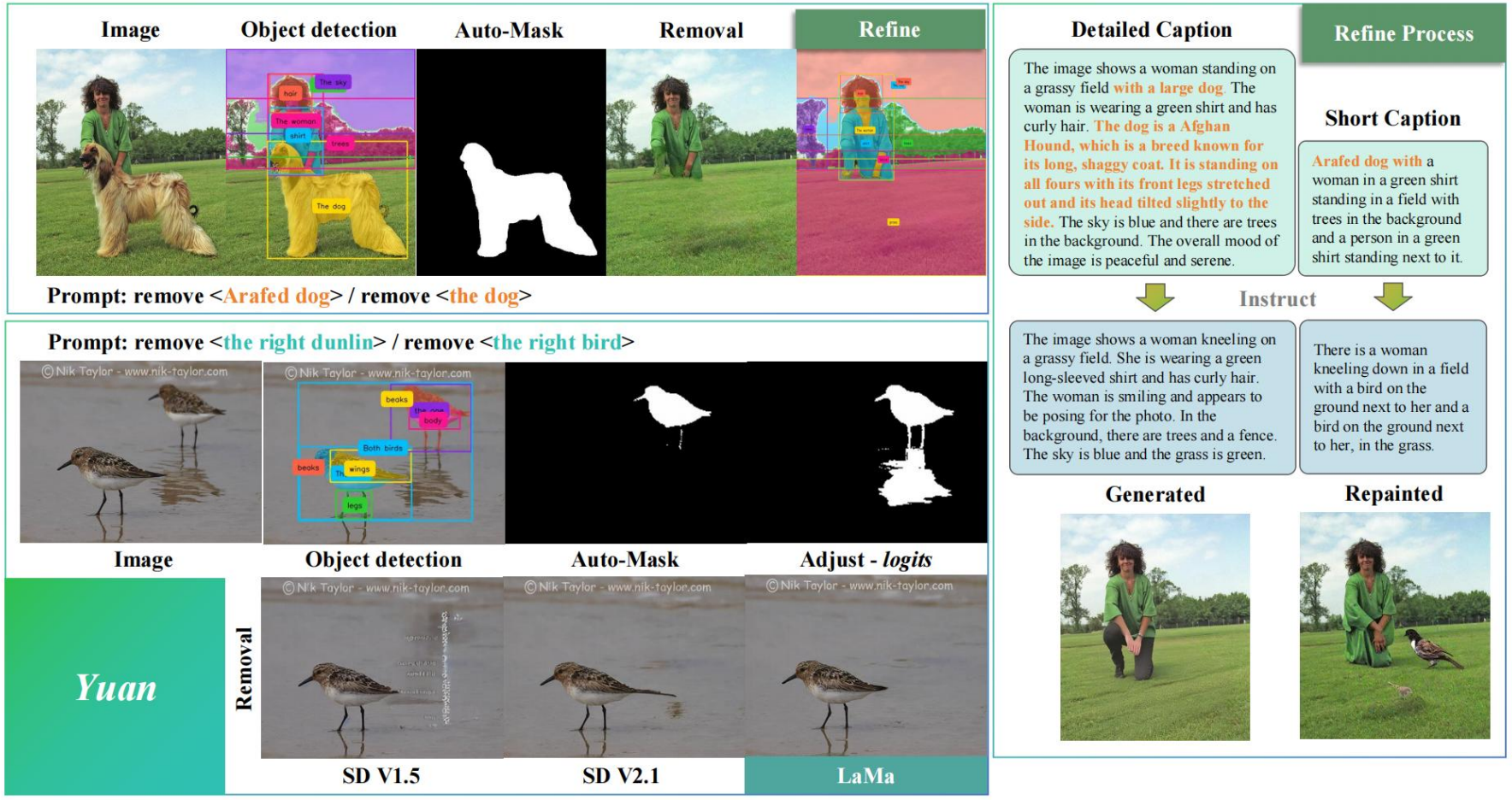


Figure 2: Our *Yuan* framework: (a) Object detection by user prompt, (b) Automatic mask generation, (c) Object removal, (d) Inpainting and preserving original context, and (e) Refined image.

Yuan Framework Overview

- **Input**: Synthetic image from any text-to-image (T2I) model.
- **Step 1:** Grounded SAM generates automatic masks based on user prompts.
- **Step 2:** LaMa inpainting model repairs identified imperfections.
- **Step 3:** Optional refinement using Prompt-to-Prompt techniques if initial results are insufficient.

Algorithm 1: *Yuan* - Object Removal

Require: Synthetic image I from any T2I model
Prompt P from user input
Ensure: Refined image *output*
$D_{GDINO} \leftarrow \text{GDINO}(I, P)$ {Detect objects}
$M_{SAM} \leftarrow \text{SAM}(D_{GDINO})$ {Generate mask}
$I_{masked} \leftarrow$ Apply ΔM_{SAM} to I
$I_{inpaint} \leftarrow \text{LaMa}(I_{masked}, \Delta M_{SAM})$ {Inpaint}
$output \leftarrow I_{inpaint}$
if $I_{inpaint}$ is insufficient **then**
  $\Delta M_{SAM} \leftarrow logit(t)$ {Adjust mask}
  $\Delta I_{masked} \leftarrow$ Apply ΔM_{SAM} to I
  $I_{inpaint2} \leftarrow \text{LaMa}(\Delta I_{masked}, \Delta M_{SAM})$ {Inpaint}
  $output \leftarrow I_{inpaint2}$
  if $I_{inpaint2}$ is insufficient **then**
    $C_I \leftarrow \text{Caption}(I)$ {Generate caption}
    $C_r \leftarrow \text{GPT}_{\text{fine-tuned}}(P, C_I)$ {Generate new caption}
    $I_{refined} \leftarrow \text{Generate}(\Delta C_r, I)$
    $output \leftarrow I_{refined}$
  end if
end if
return *output*

Experimental Setup

- **Datasets:**
  - **ImageNet-100:** 60,000 training images and 10,000 validation images across 100 categories.
  - **Stanford Dogs:** 20,580 images of 120 dog breeds.
  - **Generated Cats:** Custom dataset created using Stable Diffusion.
- **Environment:**
  - NVIDIA GeForce RTX 4090 GPU with 24 GB memory.

Results Comparison

- **Metrics:**
  - **NIQE**, **BRISQUE**, and **PI:**
  - ***Yuan*** consistently outperforms other models (see Table 1).

Table 1: Comparison of object removal performance across different models. It compares the performance of Grounded SAM+SD, +LaMa, and *Yuan* on object removal tasks across three datasets: ImageNet100, Stanford-dogs, and Generated-cats.

Metrics	ImageNet100				Stanford-dogs				Generated-cats			
	Image	+SD	+LaMa	*Yuan*	Image	+SD	+LaMa	*Yuan*	Image	+SD	+LaMa	*Yuan*
NIQE↓	*3.7425*	5.2829	3.0905	**3.0890**	*3.3380*	4.6187	4.0785	**3.4691**	*5.2829*	6.2217	**5.0716**	5.2465
BRISQUE↓	*26.6525*	32.1852	**24.7853**	25.6086	*9.6086*	25.1237	22.0275	**16.2062**	*32.1852*	**37.4372**	45.6096	39.4333
PI↓	*2.5558*	5.9084	2.0921	**2.0124**	*2.2204*	3.2685	2.6190	**2.2841**	*5.9084*	6.7089	5.5097	**5.4679**

Results Comparison

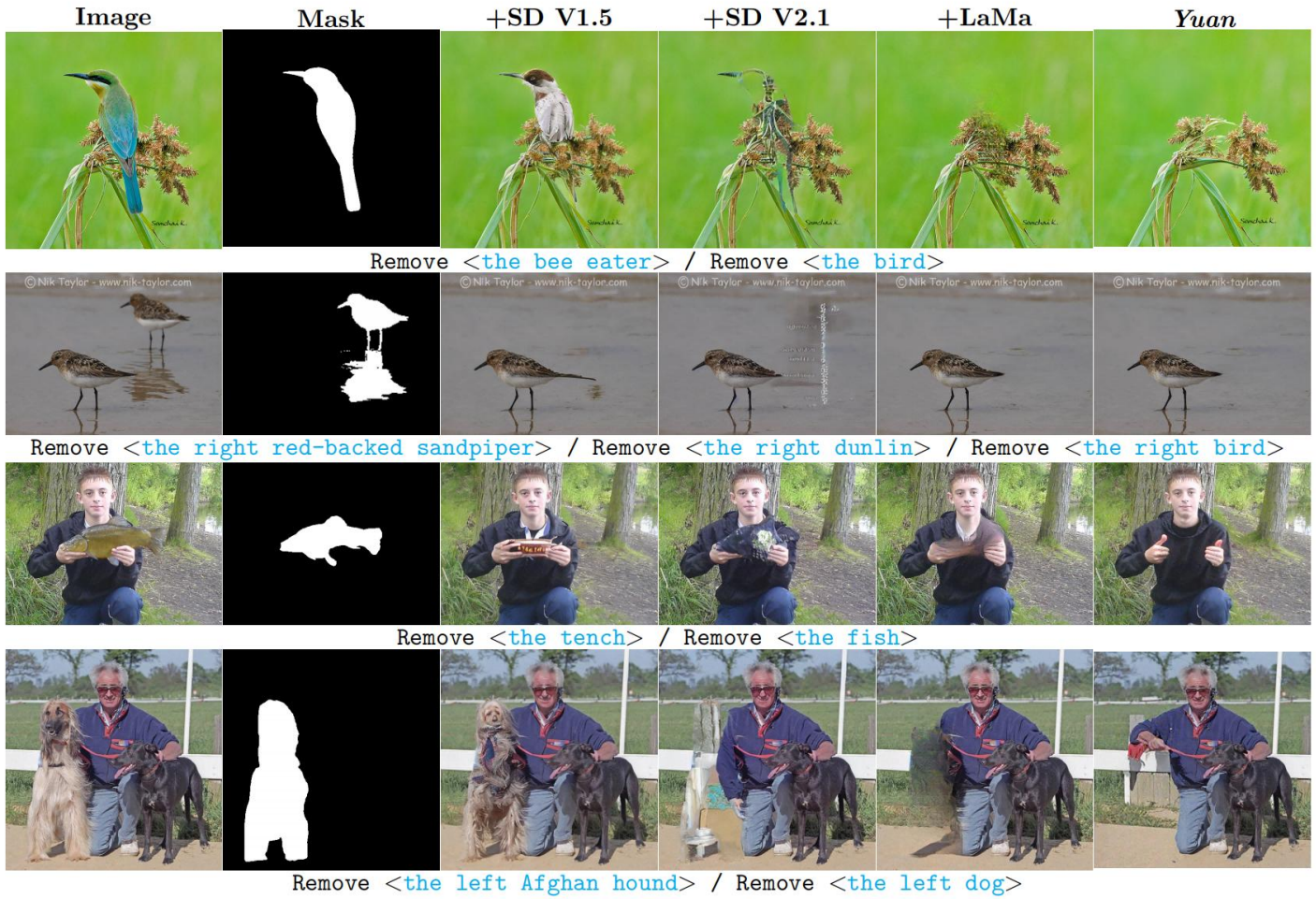
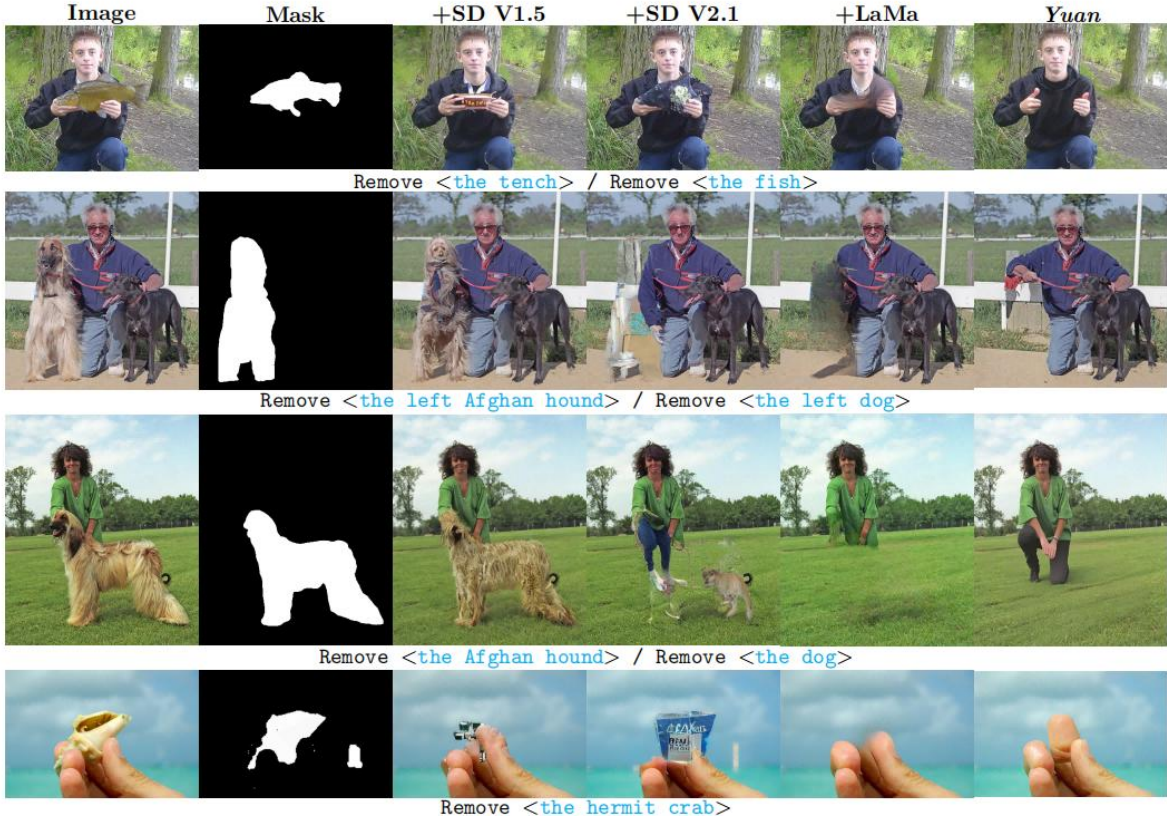
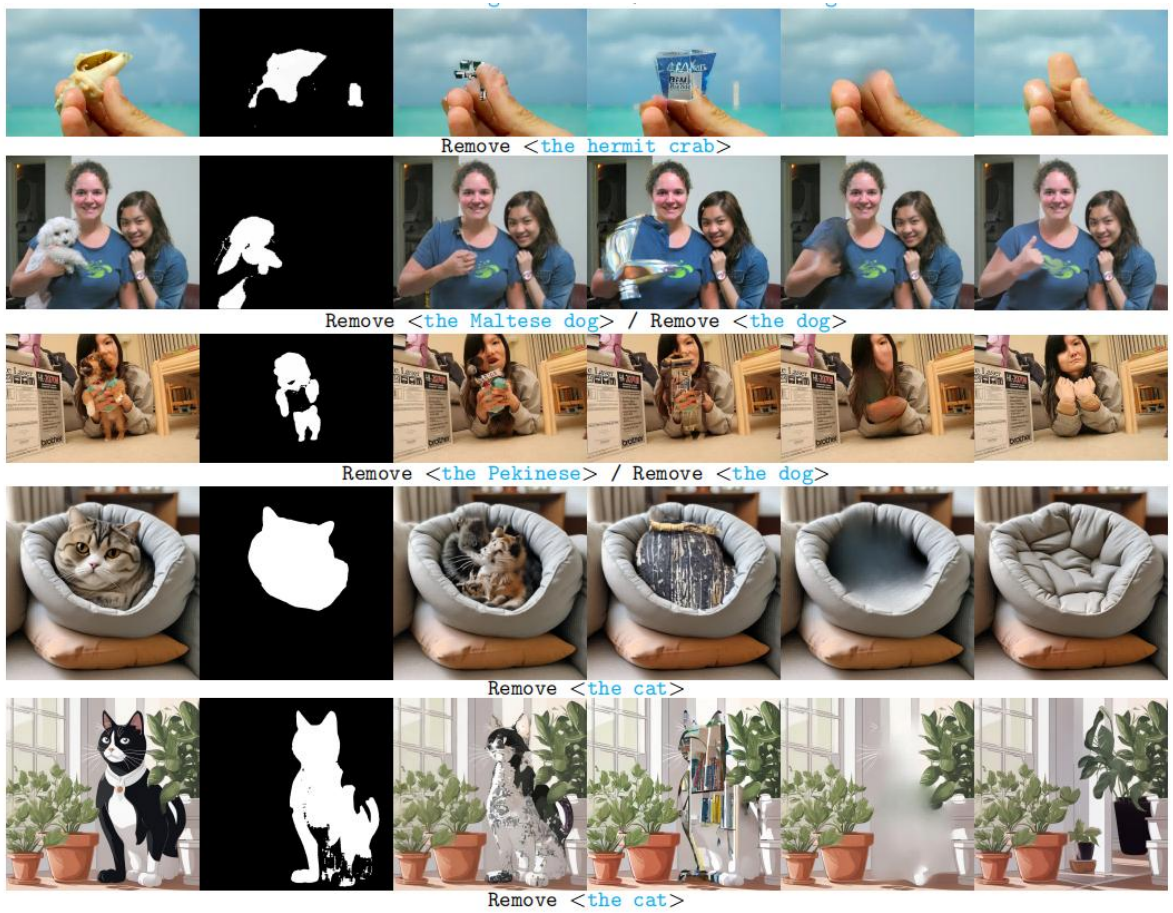


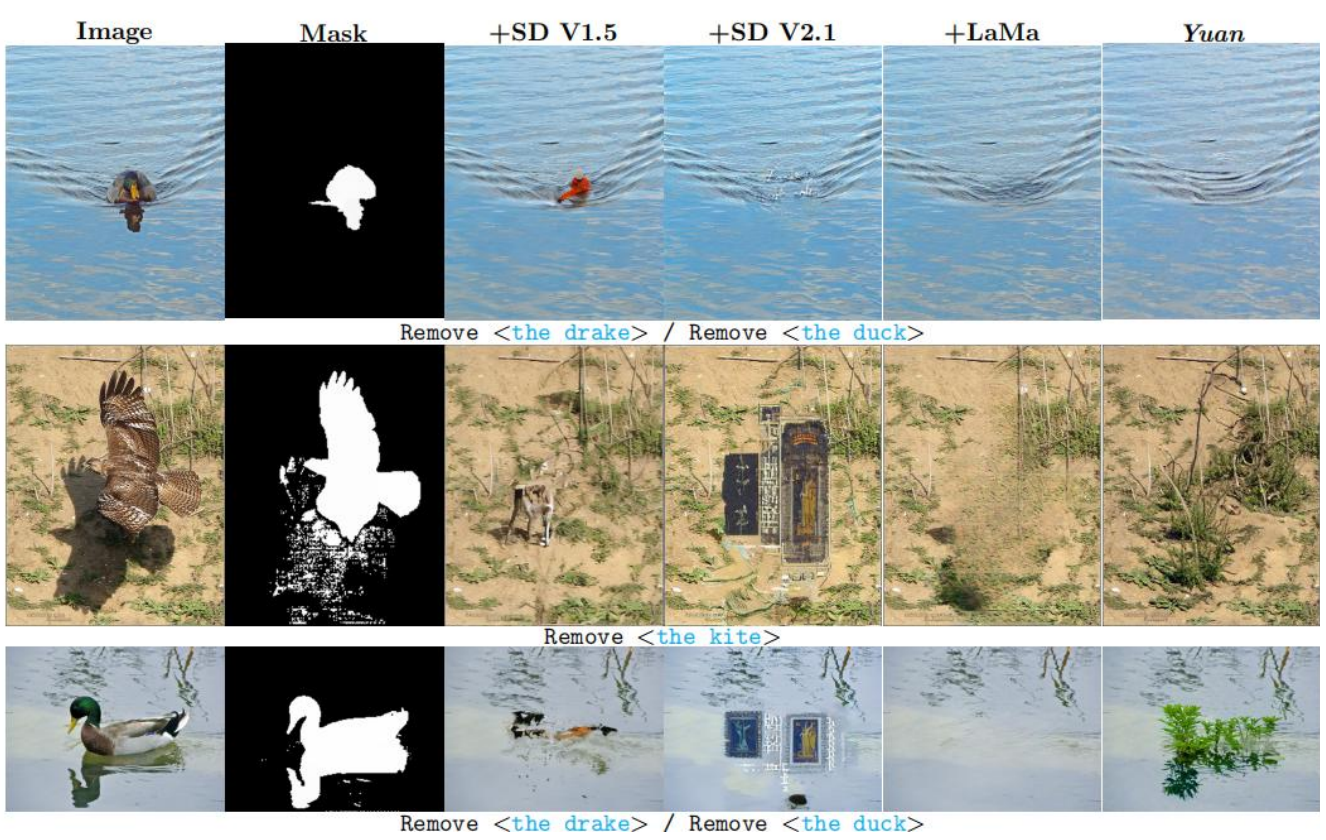
Figure 3: A comparison among Grounded SAM+SD V1.5, +SD V2.1, +LaMa, and *Yuan* for different text prompt.

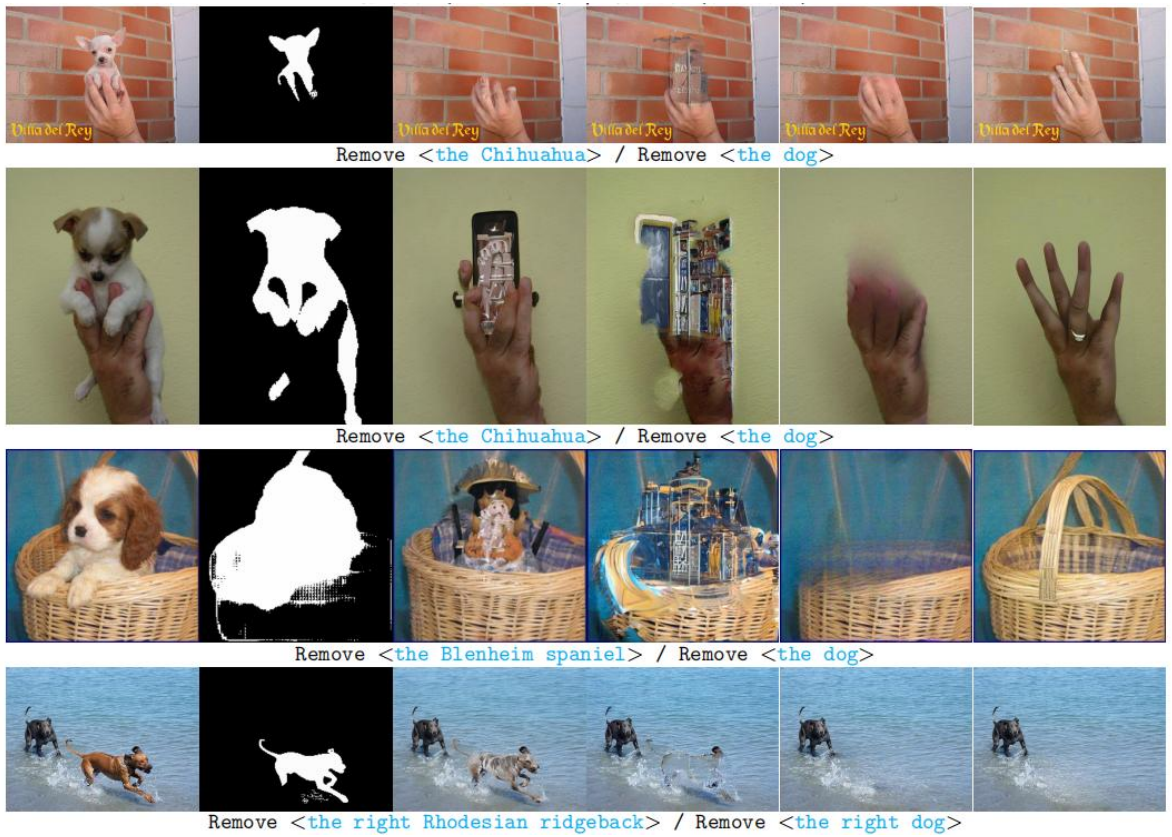
Results Comparison



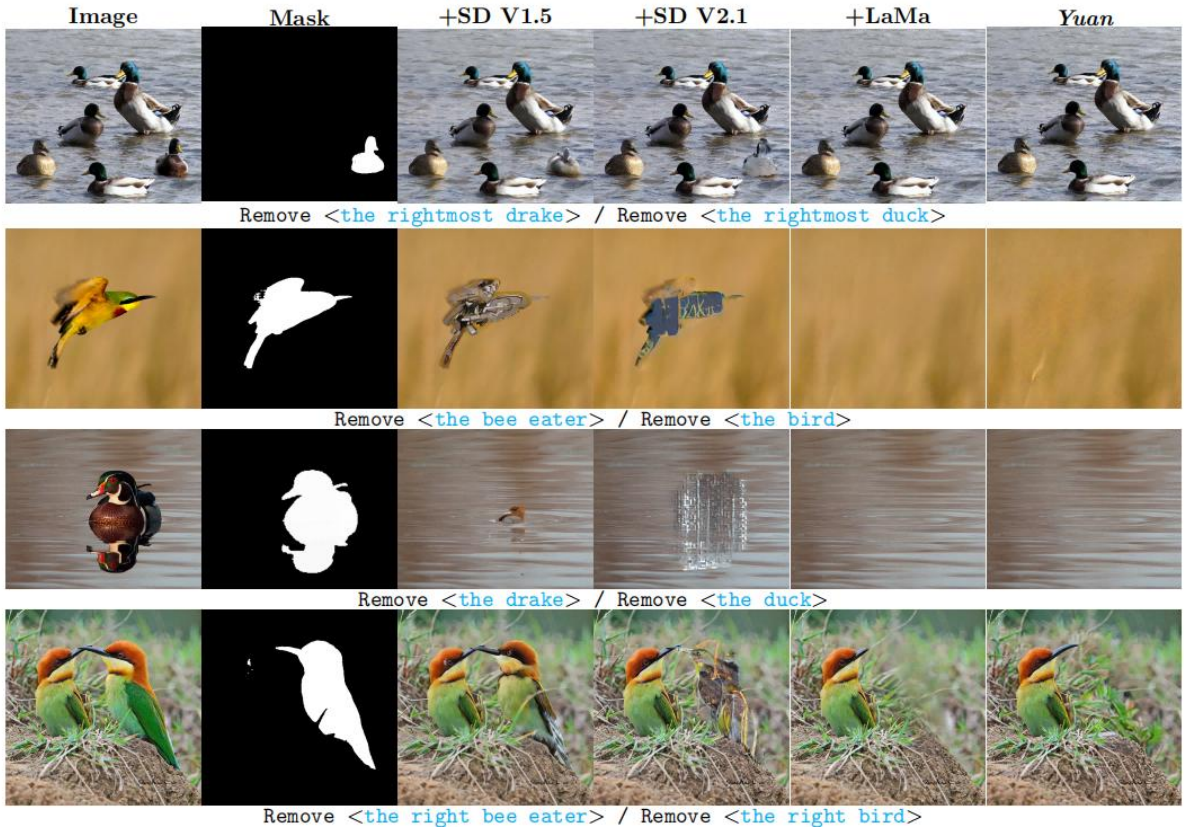


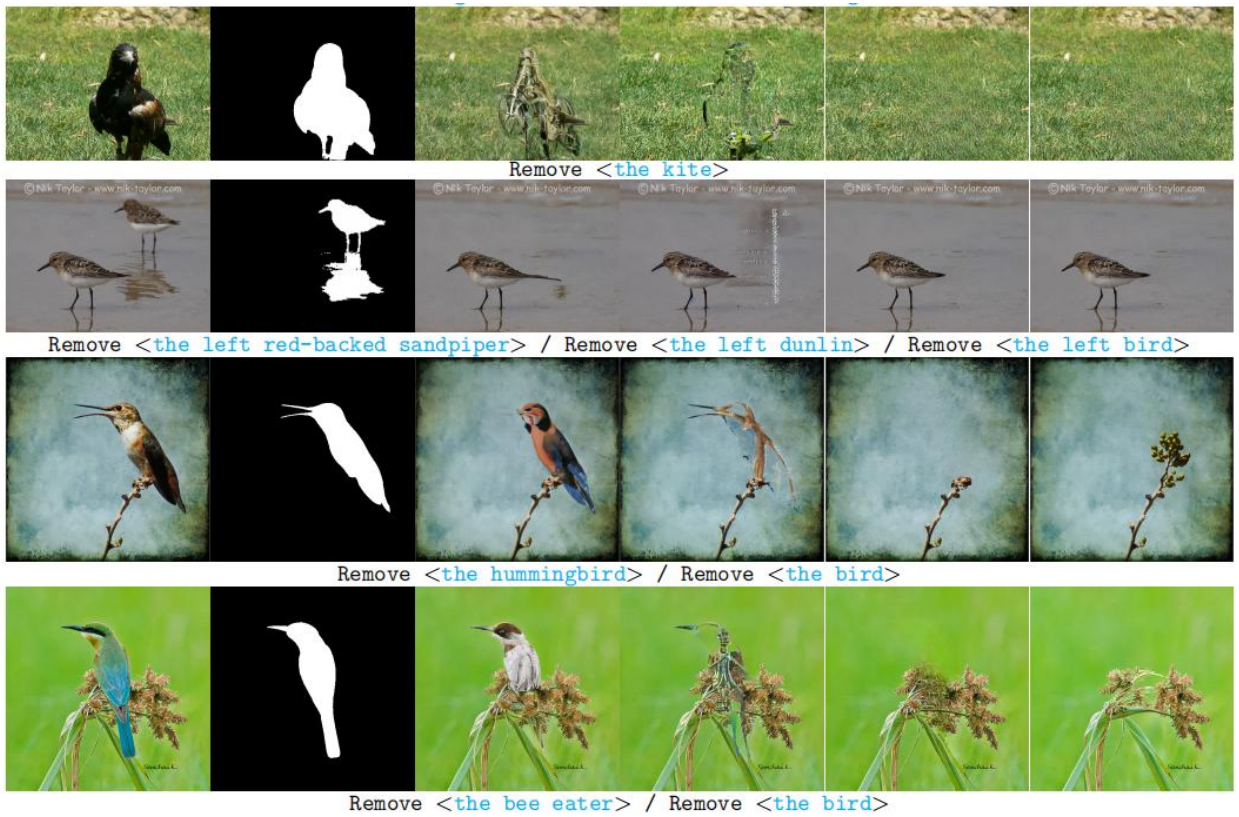
Results Comparison



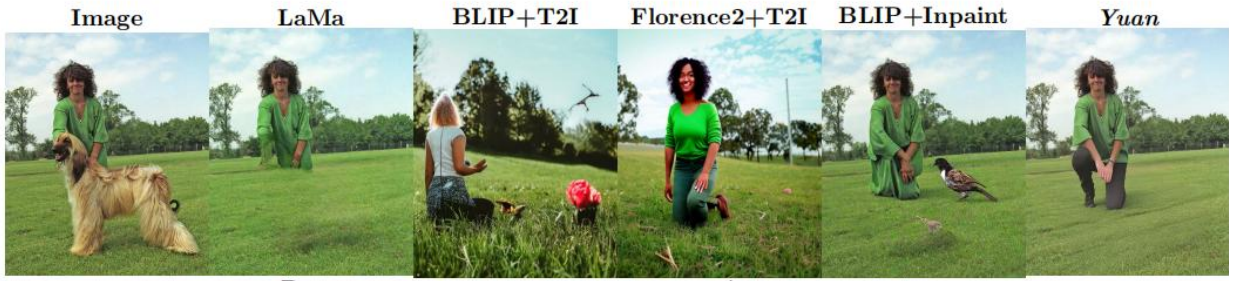


Results Comparison





Results Comparison



Prompt: Remove <the Afghan hound> / Remove <the dog>

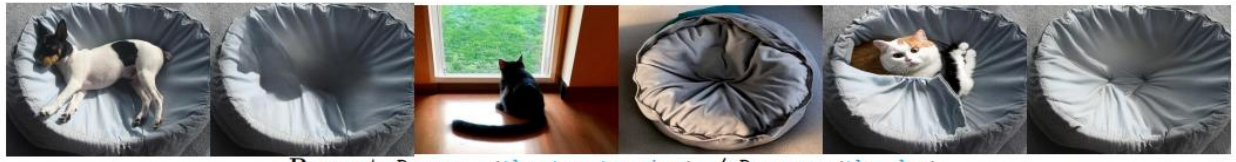
Caption: The image shows a woman kneeling on a grassy field. She is wearing a green long-sleeved shirt and has curly hair. The woman is smiling and appears to be posing for the photo. In the background, there are trees and a fence. The sky is blue and the grass is green.

Prompt: Remove <the Maltese dog> / Remove <the dog>

Caption: The image shows two young women posing for a selfie in a living room. The woman on the left is wearing a blue t-shirt with a green logo on it and has curly hair. She is smiling and looking directly at the camera. The other woman is standing next to her, also smiling. Both women are wearing denim jackets and have their hair pulled back in a ponytail. They are standing in front of a white wall and there is a couch in the background.

Prompt: Remove <the Pekinese> / Remove <the dog>

Caption: The image shows a young woman sitting on the floor in a living room. She is wearing a grey sweatshirt and has long dark hair. Her arms are crossed over her chest and she is looking directly at the camera with a serious expression on her face. On the left side of the image, there is a large cardboard box with the brand name "Brother" printed on it. The box appears to be a laser printer, as indicated by the text on the box. In the background, there are bookshelves and a coffee table.



Prompt: Remove <the toy terrier> / Remove <the dog>

Caption: The image shows a round, light grey bean bag chair on a gray carpeted floor. The bean bag is made of a soft, plush material and has a scalloped edge. The edges of the bean bag are slightly frayed, indicating that it has been used before. The chair appears to be empty and is in good condition.

Prompt: Remove <the Chihuahua> / Remove <the dog>

Caption: The image shows a hand of a person cleaning a brick wall. The wall is made up of red bricks arranged in a horizontal pattern. The bricks appear to be old and weathered, with some areas of the bricks having a rough texture. The hand is in the process of cleaning the wall, with the fingers slightly curled and the thumb and index finger pointing towards the right side of the image. The text "Villa del Rey" is written in yellow on the bottom right corner.

Ablation Study

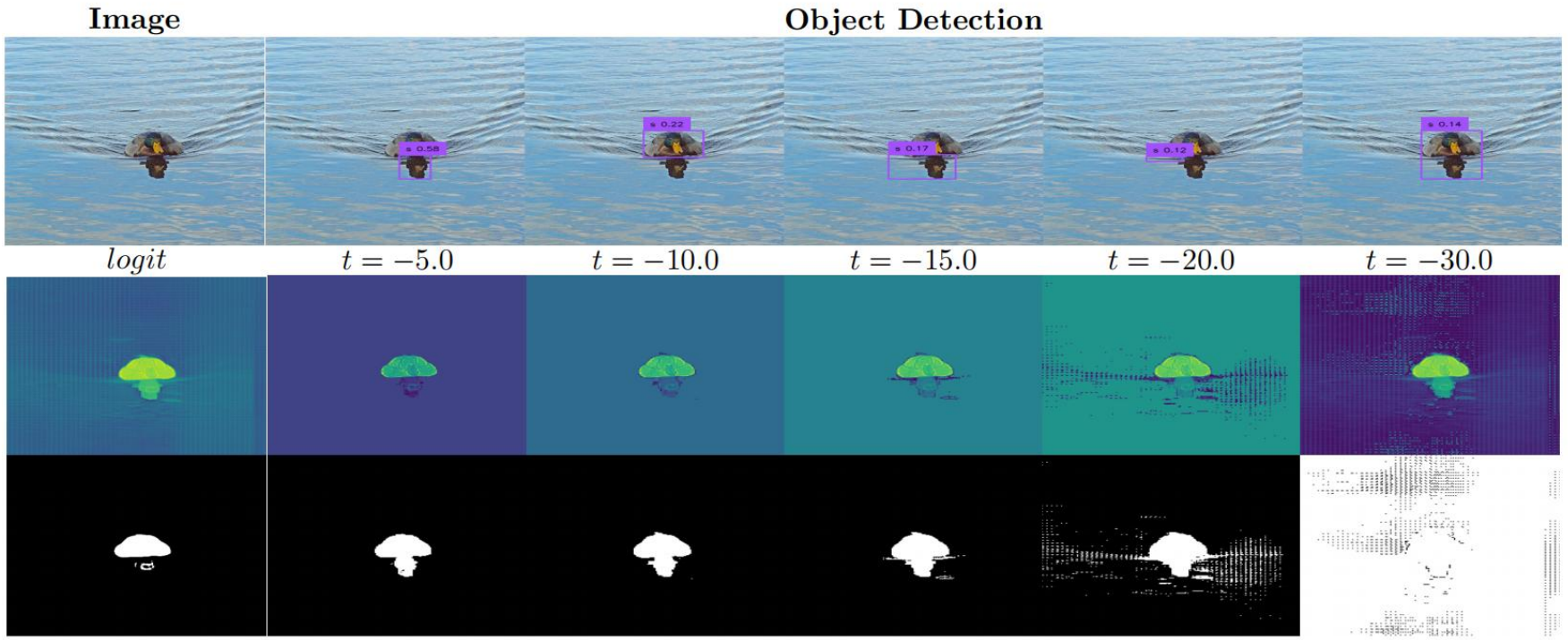


Figure 4: Ablation study results on $logits$ threshold (t) adjustment for automatic mask generation.

Ablation Study

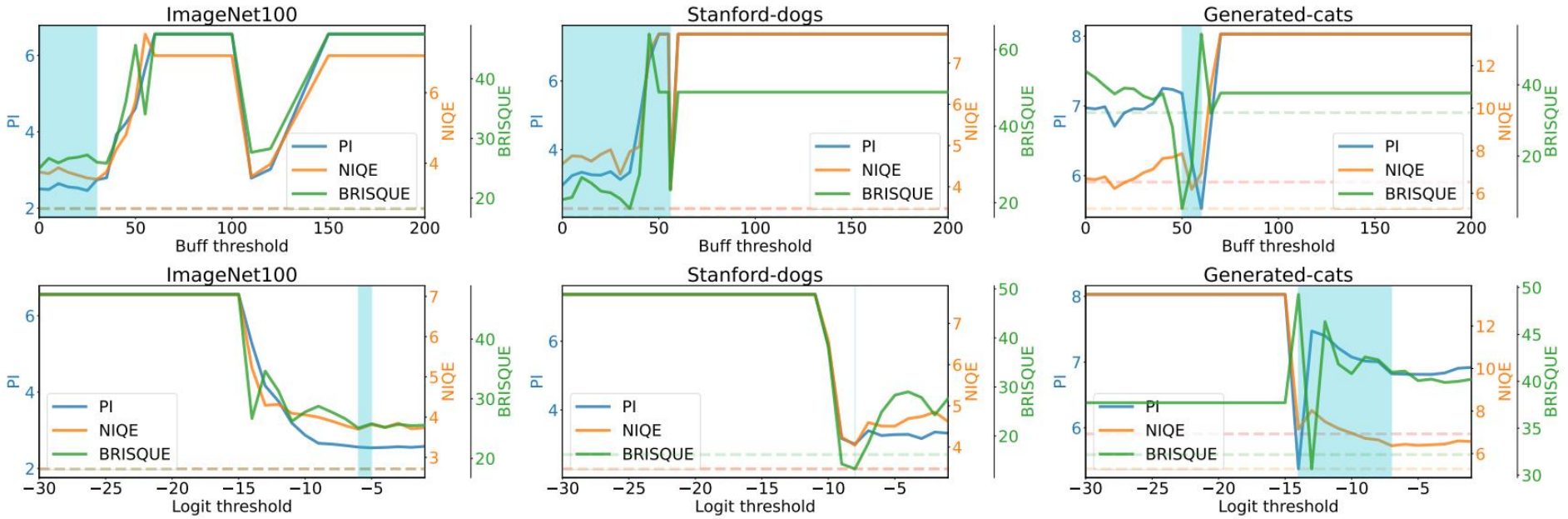
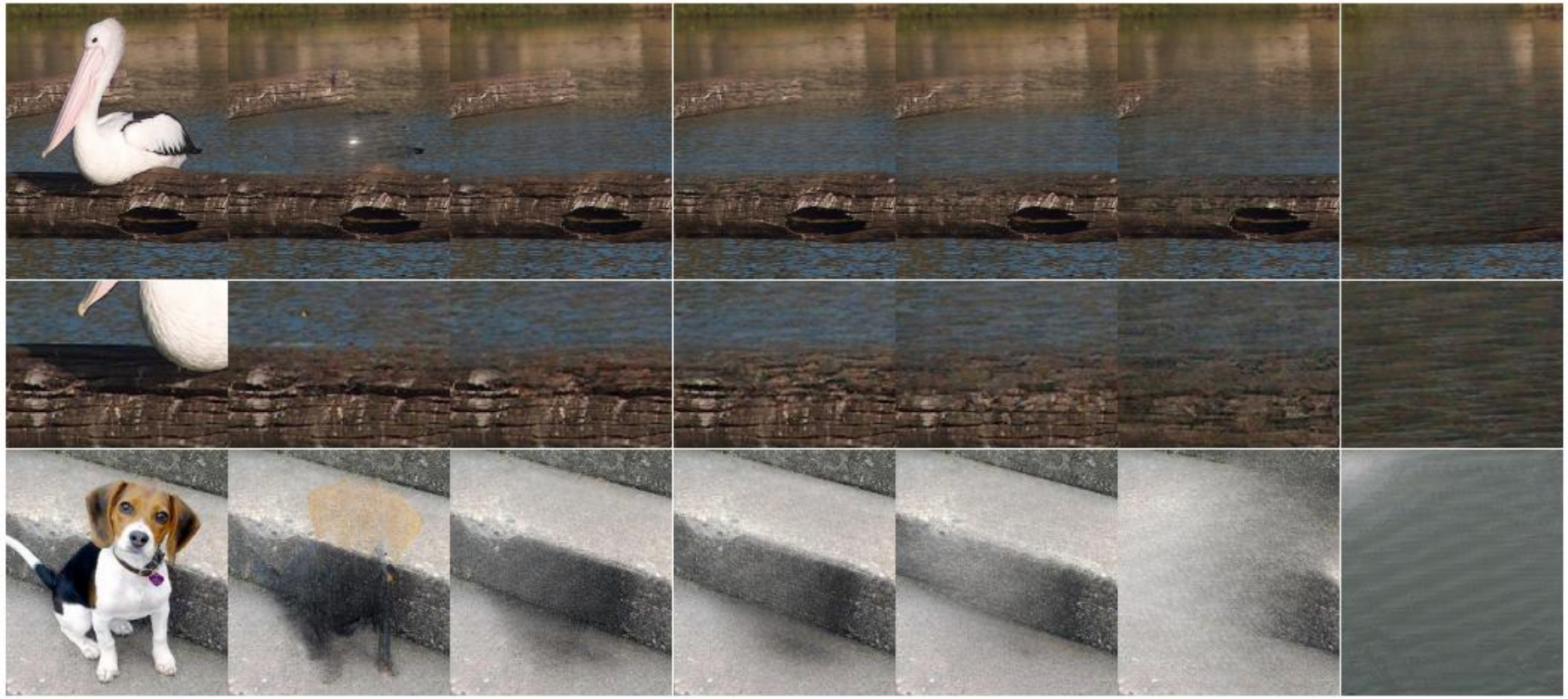


Figure 5: Ablation study results of threshold in buffer (b) and *logits* (t). For different datasets, the threshold needs to be adjusted as needed. For *Yuan*, we recommend adding two adjustable parameters, exposure b and t, based on the original settings. This will provide convenience and service for different generated images to achieve the best results.

Ablation Study



(a) Image (b) 0 (c) 15 (d) 30 (e) 50 (f) 100 (g) 200

Figure A.1: Effect of buffer zone thresholds (b) on reconstruction quality.

Ablation Study

Table A.1: Comparison of buffer zone thresholds (b) across datasets.

b	ImageNet100			Stanford-dogs			Generated-cats		
	NIQE↓	BRISQUE↓	PI↓	NIQE↓	BRISQUE↓	PI↓	NIQE↓	BRISQUE↓	PI↓
0	3.7537	**24.9772**	**2.5064**	4.5526	**20.8205**	**2.9659**	6.6916	43.7922	6.9720
15	3.7425	26.6525	2.5558	4.6187	25.1237	3.2685	**6.2217**	37.4372	**6.7089**
30	**3.5438**	25.9997	2.7441	**4.3036**	21.0217	3.1316	6.9695	36.8356	6.9524
50	5.7662	45.6513	4.6176	7.7046	48.8762	7.3630	7.8724	**35.2731**	7.1823
100	7.0501	47.5065	6.5582	7.7046	48.8762	7.3630	13.4853	37.7114	8.0294
200	7.0501	47.5065	6.5582	7.7046	48.8762	7.3630	13.4853	37.7114	8.0294

Table A.2: Impact of *logits* threshold (t) on auto-mask sensitivity and image quality.

t	ImageNet100			Stanford-dogs			Generated-cats		
	NIQE↓	BRISQUE↓	PI↓	NIQE↓	BRISQUE↓	PI↓	NIQE↓	BRISQUE↓	PI↓
0.0	**3.7350**	**25.5632**	2.5786	4.6179	**27.6525**	3.3228	6.5708	40.2131	6.9190
-5.0	3.8162	25.8107	**2.5403**	**4.5064**	28.3264	**3.2824**	**6.3903**	40.0700	**6.8134**
-10.0	4.0561	27.7858	2.8707	6.5758	38.1330	5.8810	6.9659	40.8078	7.0764
-15.0	7.0501	47.5065	6.5582	7.7046	48.8762	7.3630	13.4853	**37.7114**	8.0294

Limitations & Future Work

Figure 6: Limitations of *Yuan*. The challenge of accurately rendering human hands due to complex anatomy, and the generation of unintended content during the refinement process.

Conclusion & Code

- **Summary:**
  - ***Yuan*** offers a scalable, automated solution for visual imperfections in generative AI.
  - It achieves superior performance without manual intervention.
- **Code:**
  - Source code is available at:
  - https://github.com/YuZhenyuLindy/Yuan.git

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

UNIVERSITI MALAYA

Thank you for your attention!

Yuan: Yielding Unblemished Aesthetics through A Unified Network for Visual Imperfections Removal in Generated Images

Code

Zhenyu Yu, Chee Seng Chan
Universiti Malaya

Email Me