madima72

Text-based Image Editing for Food Images with CLIP

Kohei Yamamoto, Keiji Yanai

The University of Electro-Communications, Tokyo, Japan

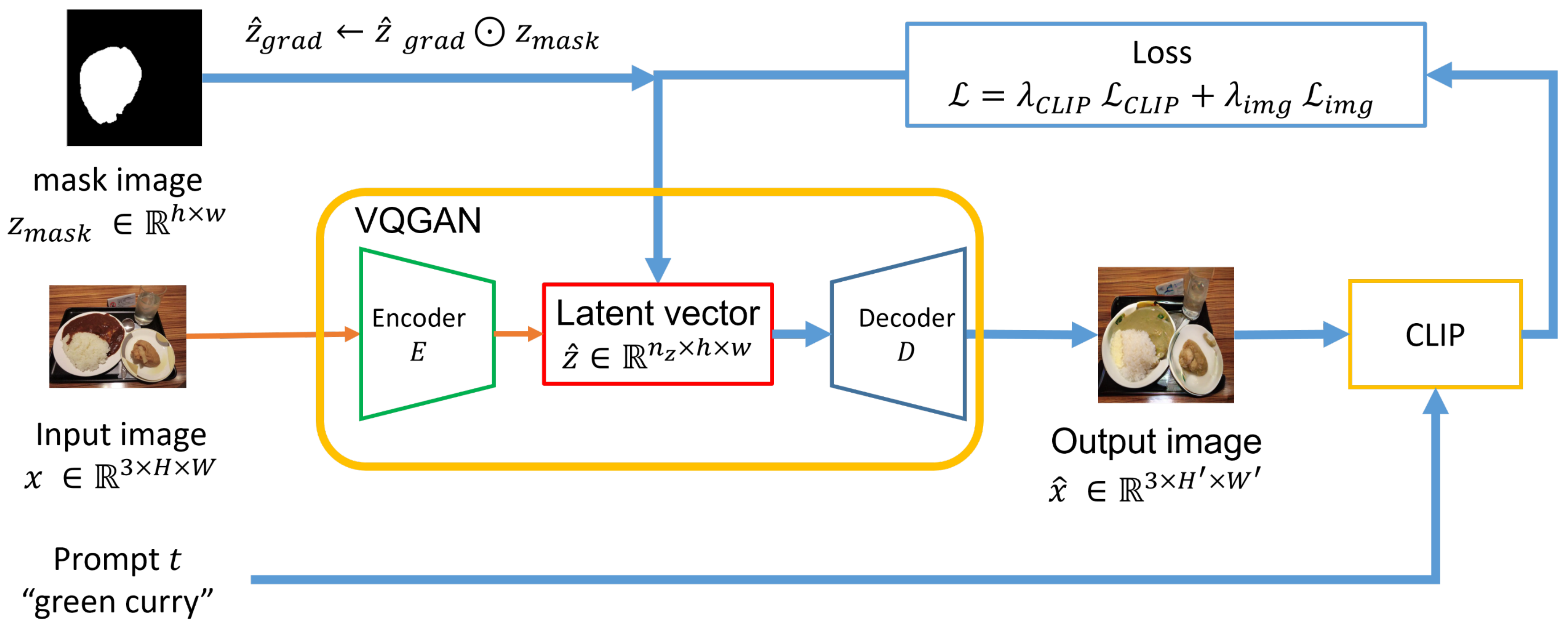
1.INTRODUCTION

There are some image synthesis models used CLIP[1] and GAN. However, **their effectiveness in the food domain has not been examined** comprehensively yet.

We reported the results of the experiments on text-based food image manipulation using VQGAN-CLIP[2].

2. ARCHITECTURE

Based on VQGAN-CLIP


$\hat{z}_{grad} \leftarrow \hat{z}_{grad} \odot z_{mask}$
Loss
$\mathcal{L} = \lambda_{CLIP}\,\mathcal{L}_{CLIP} + \lambda_{img}\,\mathcal{L}_{img}$
mask image
$z_{mask} \in \mathbb{R}^{h \times w}$
VQGAN
Encoder E
Latent vector $\hat{z} \in \mathbb{R}^{n_z \times h \times w}$
Decoder D
CLIP
Input image $x \in \mathbb{R}^{3 \times H \times W}$
Output image $\hat{x} \in \mathbb{R}^{3 \times H' \times W'}$
Prompt t
"green curry"

1. VQGAN encoder generates the latent vector from input images.
2. VQGAN decoder generates output images from latent vector.
3. CLIP calculate similarities between images and prompts.
4. The model calculates the loss by similarities.
5. (Optional) The latent vector gradient $\hat{\boldsymbol{z}}_{grad}$ updates by based on z_{mask}.
6. The loss function updates the latent vector by gradient descent method.

$$\lambda_{CLIP} = \lambda_{img} = 1, \mathcal{L}_{CLIP} = 2\arcsin^2\left(\frac{I - T}{2}\right), \mathcal{L}_{img} = 2\arcsin^2\left(\frac{I - I_{img}}{2}\right)$$

(I: output image token, I_{img}:input image token, T:text token)

3. EXPERIMENT

1. **Compare the prompts** for food image editing
2. **Fine-tune VQGAN and CLIP** on food datasets

The list of food datasets

datasets name	number of categories	number of images
Magical Rice Bowl[3]	10	80,408
Foodx251[4]	251	158,846
Food500[5]	500	399,726
Recipe1M(Train, Valid)[6]	-	753,251

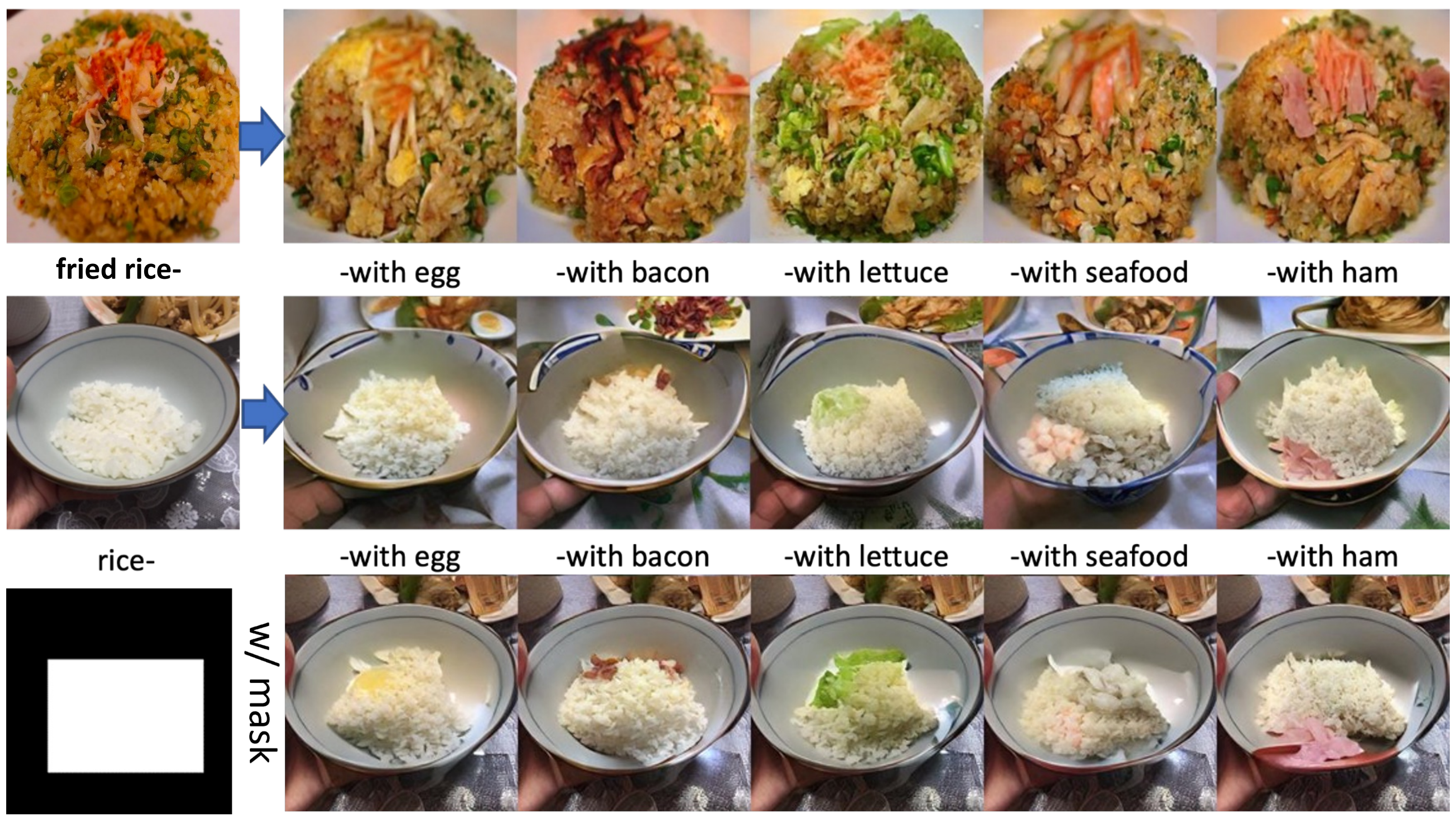
Prompt for CLIP training

abbreviation	prompts for training	pre-train
title_NoPretrain	**some_title**	×
title	**some_title**	∘
ingredients	**ingredients**	∘
ingredients_title	**ingredients** + ' are ingredients in ' + **some_title** + ' .'	∘
APhotoOf	'A photo of a ' + **some_title** + ' .'	∘
APhotoOf_ATypeOfFood	'A photo of a ' + **some_title** + ' , a type of food .'	∘

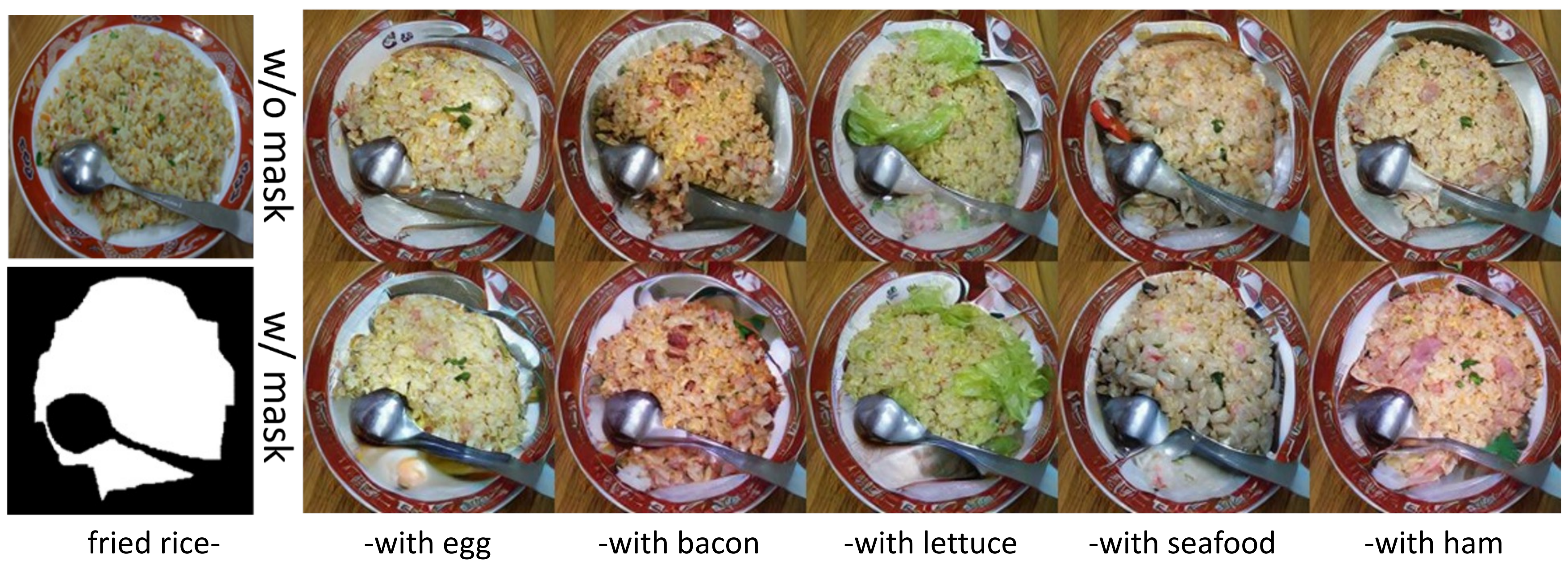
- $x, \hat{x} \in \mathbb{R}^{3 \times 256 \times 256}$
- $\hat{\boldsymbol{z}} \in \mathbb{R}^{256 \times 16 \times 16}$
- 1000 times iteration
- 4-6 minutes per images for image editing

4. RESULTS

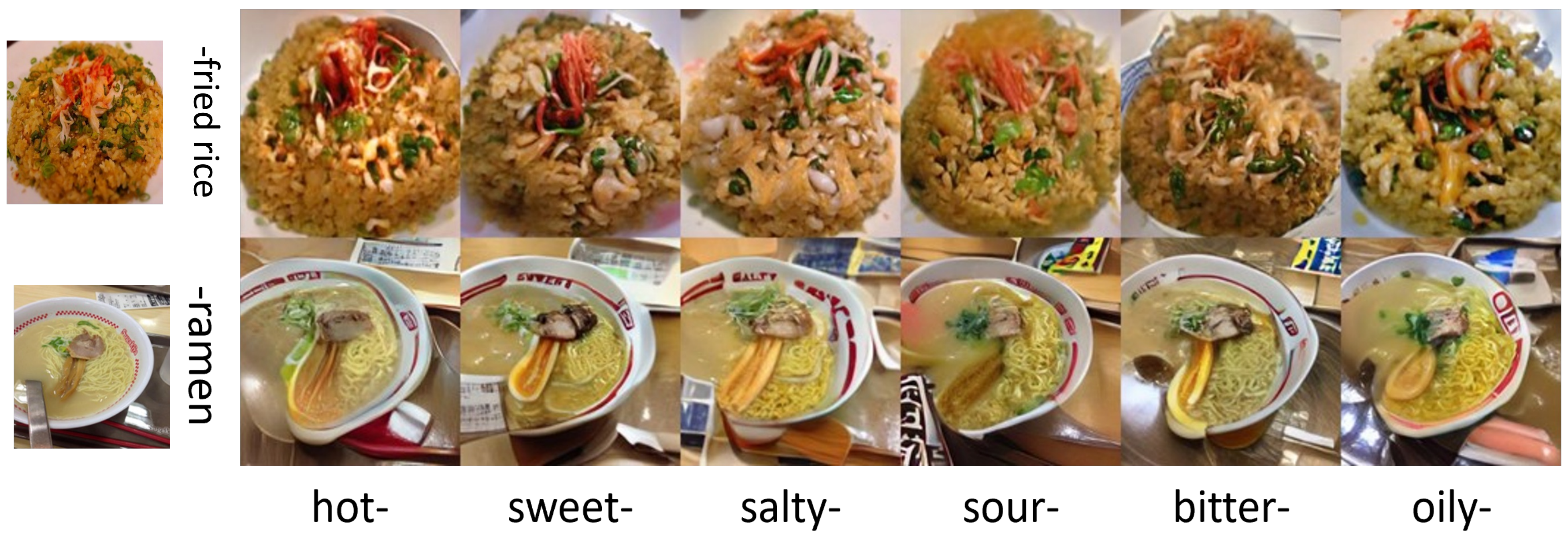
Prompts of adding topping→**Toppings were added**



Editing with a mask image → **backgrounds were saved.**



Prompts with taste adjective→**little change**



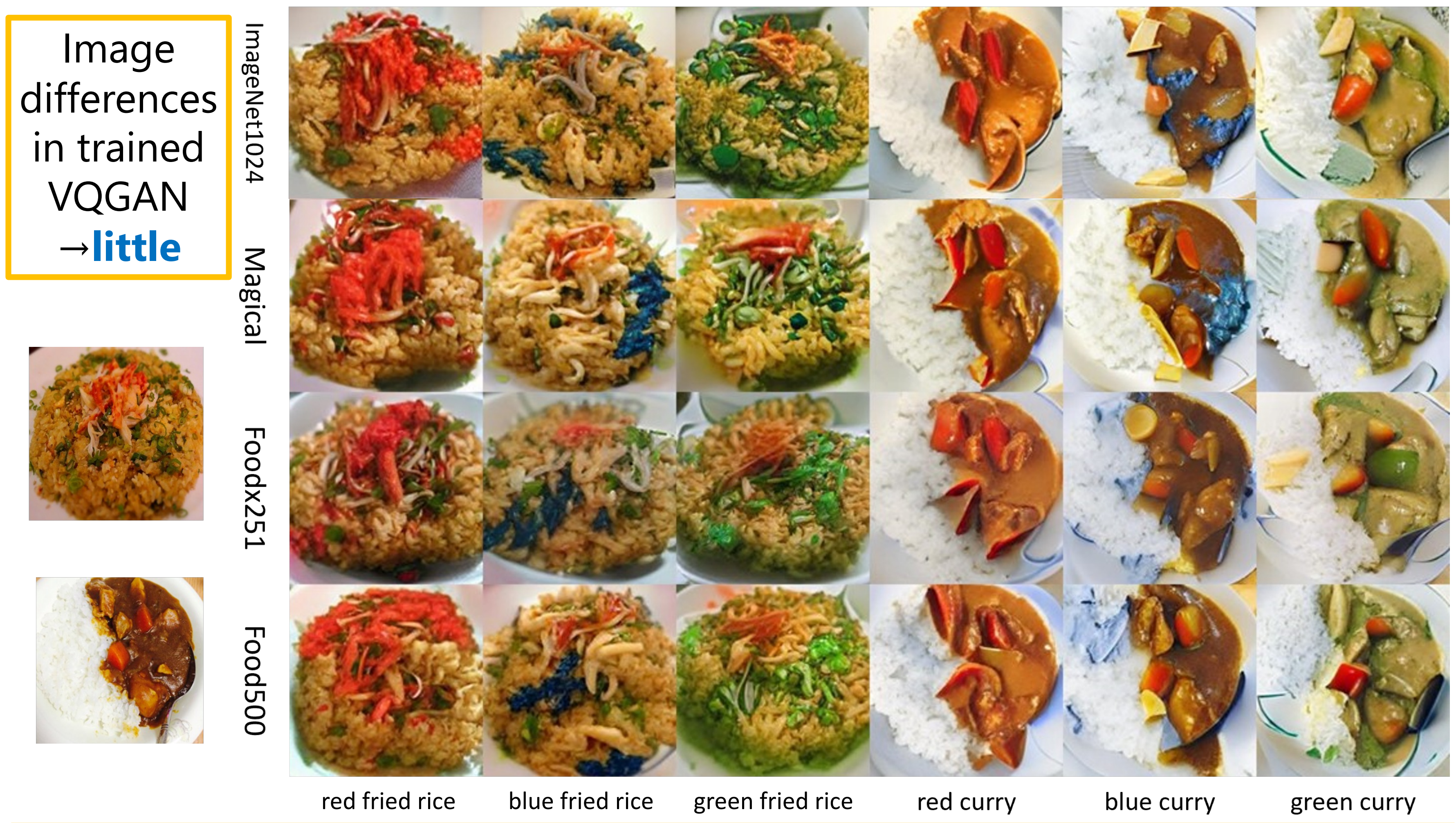


Image differences in trained VQGAN →**little**

Reconstructed quantitative evaluation of VQGAN

→**More images and categories, better results.**

50k	IS↑	FID↓	KID↓($\times 10^{-3}$)
imagenet1024	**7.09 ± 0.10**	6.73	3.92 ± 0.46
Magical Rice Bowl	5.97 ± 0.06	7.15	3.51 ± 0.48
foodx251	6.15 ± 0.06	4.59	1.85 ± 0.31
food500	6.62 ± 0.05	**4.07**	**1.66 ± 0.29**

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
[2] Crowson, Katherine, et al. "Vqgan-clip: Open domain image generation and editing with natural language guidance." arXiv preprint arXiv:2204.08583 (2022).
[3] Horita, Daichi, et al. "Food category transfer with conditional cyclegan and a large-scale food image dataset." Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. 2018.
[4] Kaur, Parneet, et al. "Foodx-251: a dataset for fine-grained food classification." arXiv preprint arXiv:1907.06167 (2019).
[5] Min, Weiqing, et al. "Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[6] Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.