SUPPLEMENTARY MATERIAL

1. MORE ABLATION STUDIES

1.1. Comparison of performance before and after finetuning the text encoder

This ablation study is performed in the case when the text encoder is only finetuned without any other additional loss function, with only reconstruction loss included.

Table 1 shows the comparison of YOLO object detection performance on each image generation method with/without text encoder finetuned on VFN dataset. We observe that finetuning both the text encoder and UNet, even without specialized loss terms, significantly boosts detection performance, whereas fine-tuning only the UNet yields comparatively poor results. This infers that the pre-trained text encoder lacks sufficient food domain knowledge, and enhancing its semantics is key to generating more accurate images.

1.2. Quantitative results on multi-noun categories

Table 2 shows the quantitative comparison of image generation performance with related works on VFN and UEC-256 dataset for multi-noun categories. It can be shown that our method for generating food images on multi-noun categories can outperform previous related works.

Table 3 shows the ablation studies of generating food images on multi-noun categories for VFN dataset. Each part of our method can contribute to the improvement of food image generation performance on multi-noun categories.

1.3. More image generation results

Figure 1 shows the image generation results for more food categories and comparison with related work. It can still be shown that our method can eliminate the generation of redundant food objects.

For example, when generating food item “cheese corn snack”, stable diffusion, structured diffusion can only generate corns, and Syngen can only generate corns with shape of “French fries.” Our method, after finetuning the text encoder, it can at least generate food items close to the “snack.” The difference is between whether the text encoder is finetuned or not. If the text encoder is finetuned (FDALA included), then the generated food object is almost like “cheese corn snack.” However, if text encoder is not finetuned(CFIG only), the generated food items is more like popcorns but still looks like snack with “corn” details.

1.4. Inter-class similarity issue

We would like to show that it is necessary to have image-concept alignment in our method. Or the generated image will have inter-class similarity problem. Figure 2 shows the confusion matrix between two easily confused food categories. It can be seen that with image-text alignment to learn local image features in model, the generated images can have lower chance to confuse YOLO object detection model. For example, the generated image on chicken salad can be distinguishable from tuna salad.

1.5. Attention map on different methods

Figure 3, Figure 4 and Figure 5 show the attention map of each word on the generated image. It can be seen that stable diffusion method can easily treat corn and dog as two separate food categories when generating “corn dog” because two words’ attention have very different area on the image. A similar phenomenon is also happened on generating “egg sandwich”, where “egg” and “sandwich” have distinct attention area on the image. When generating the “crab cake” food image, in stable diffusion, “crab” is also focused on an area that is very different from the word “cake.”

Therefore, it is important to finetune CLIP text encoder to make it learn the relationship between words.

2. REFERENCES

[1] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” *International Conference on Learning Representations*, 2022.

[2] W. Feng, X. He, T. Fu, V. Jampani, A. Reddy Akula, P. Narayana, S. Basu, X. Wang, and W. Wang, “Training-free structured diffusion guidance for compositional text-to-image synthesis,” *The Eleventh International Conference on Learning Representations*, 2023.

[3] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, “Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment,” *Advances in Neural Information Processing Systems*, 2024.

Table 1. YOLOv8 detection and FID score results on generated images for VFN dataset

	No finetuning on CLIP text encoder				CLIP text encoder finetuned			
Generation Method	Precision	Recall	F-1 score	FID score	Precision	Recall	F-1 score	FID score
Real images	0.728	0.726	0.727	–	0.728	0.726	0.727	–
Stable diffusion [1]	0.42	0.414	0.417	**38.6**	0.53	0.562	0.545	**32.8**
Structured diffusion [2]	0.434	0.412	0.423	37.4	0.522	0.547	0.534	37.0
Syngen [3]	0.417	0.348	0.379	42.1	0.415	0.449	0.431	32.1
CFIG(Ours)	**0.457**	**0.422**	**0.439**	38.8	**0.541**	**0.575**	**0.557**	32.9

Table 2. Comparison with related works on generated images for VFN and UEC-256 dataset **on multi-noun categories**

		VFN dataset				UEC-256 dataset			
Method	Text encoder finetuning	Precision ↑	Recall↑	F-1 score↑	FID score↓	Precision↑	Recall↑	F-1 score↑	FID score↓
Real images (Upper bound)	–	0.744	0.75	0.747	–	0.853	0.798	0.825	–
Stable diffusion [1]	No	0.37	0.342	0.355	86.2	0.236	0.231	0.233	36.3
Structured diffusion [2]	No	0.411	0.366	0.387	84.8	0.203	0.206	0.204	39.0
Syngen [3]	No	0.411	0.237	0.301	92.0	0.199	0.072	0.106	56.7
TextCraftor[4]	Yes	0.476	0.498	0.486	**72.7**	0.518	**0.478**	0.497	**35.5**
FoCULR(Ours)	Yes	**0.551**	**0.542**	**0.546**	73.7	**0.55**	0.471	**0.507**	**35.5**

Table 3. Ablation studies of our method for VFN dataset on **multi-noun categories**

FDALA	CFIG	Text encoder finetuning	Precision	Recall	F-1 score	FID score
No	Yes	No	0.386	0.372	0.379	85.9
Yes	No	Yes	0.483	0.541	0.51	**72.5**
Yes	Yes	Yes	**0.551**	**0.542**	**0.546**	73.7

[4] Y. Li, X. Liu, A. Kag, J. Hu, Y. Idelbayev, D. Sagar, Y. Wang, S. Tulyakov, and J. Ren, "Textcraftor: Your text encoder can be image quality controller," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7985–7995, 2024.

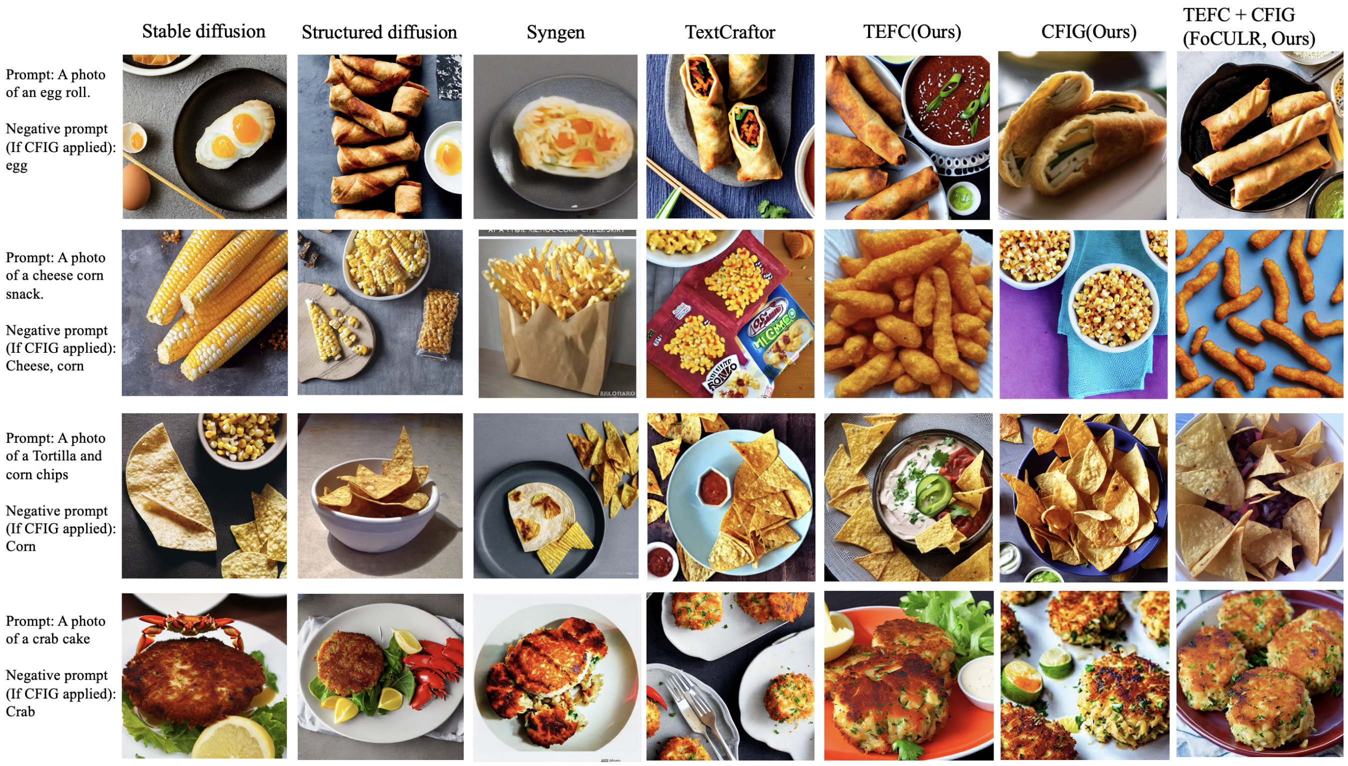
Stable diffusion
Structured diffusion
Syngen
TextCraftor
TEFC(Ours)
CFIG(Ours)
TEFC + CFIG (FoCULR, Ours)
Prompt: A photo of an egg roll.
Negative prompt (If CFIG applied): egg
Prompt: A photo of a cheese corn snack.
Negative prompt (If CFIG applied): Cheese, corn
Prompt: A photo of a Tortilla and corn chips
Negative prompt (If CFIG applied): Corn
Prompt: A photo of a crab cake
Negative prompt (If CFIG applied): Crab

Fig. 1. More image generation results and comparison of different method

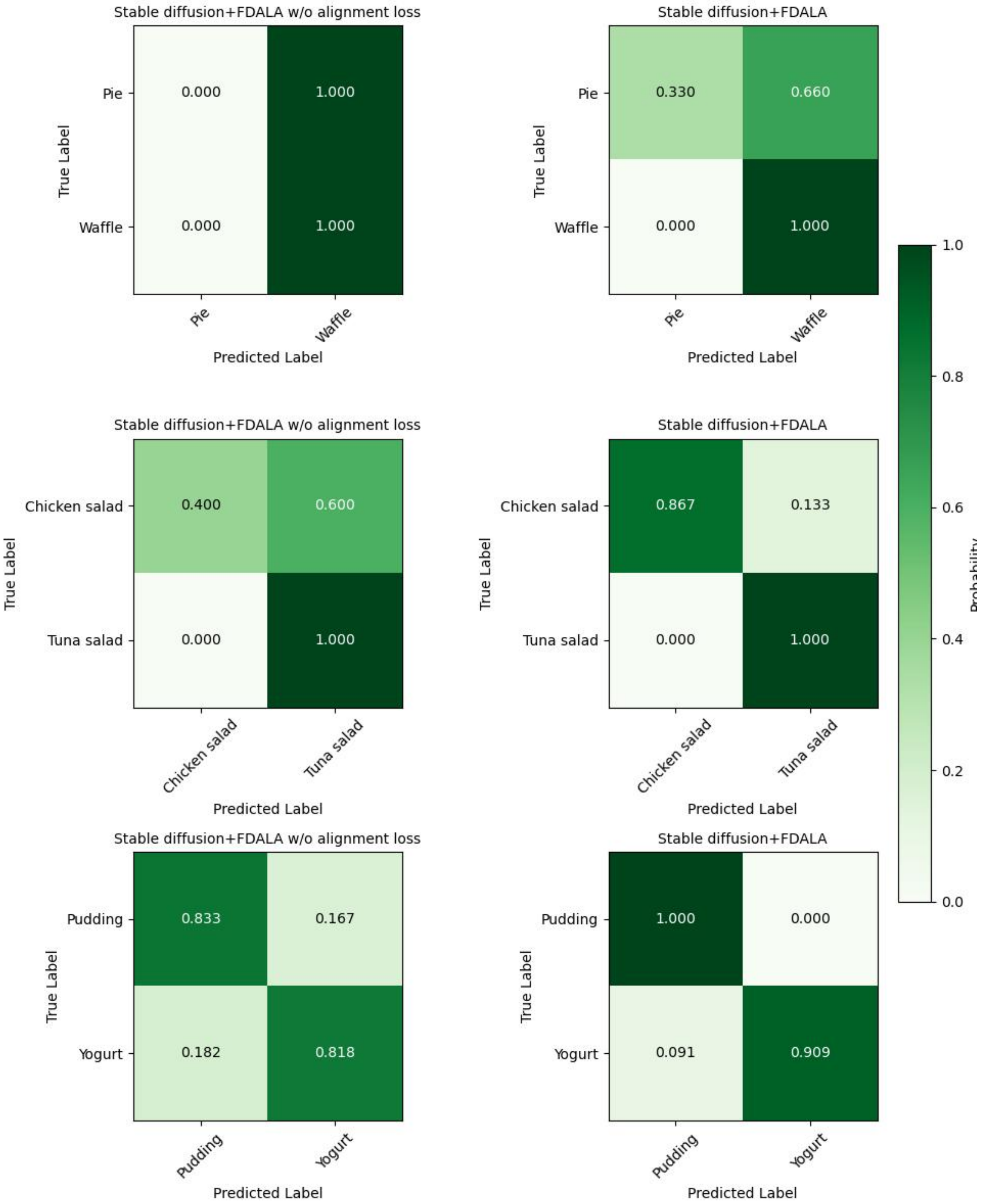

Fig. 2. Confusion matrix between easily confused categories with/without image-concept alignment

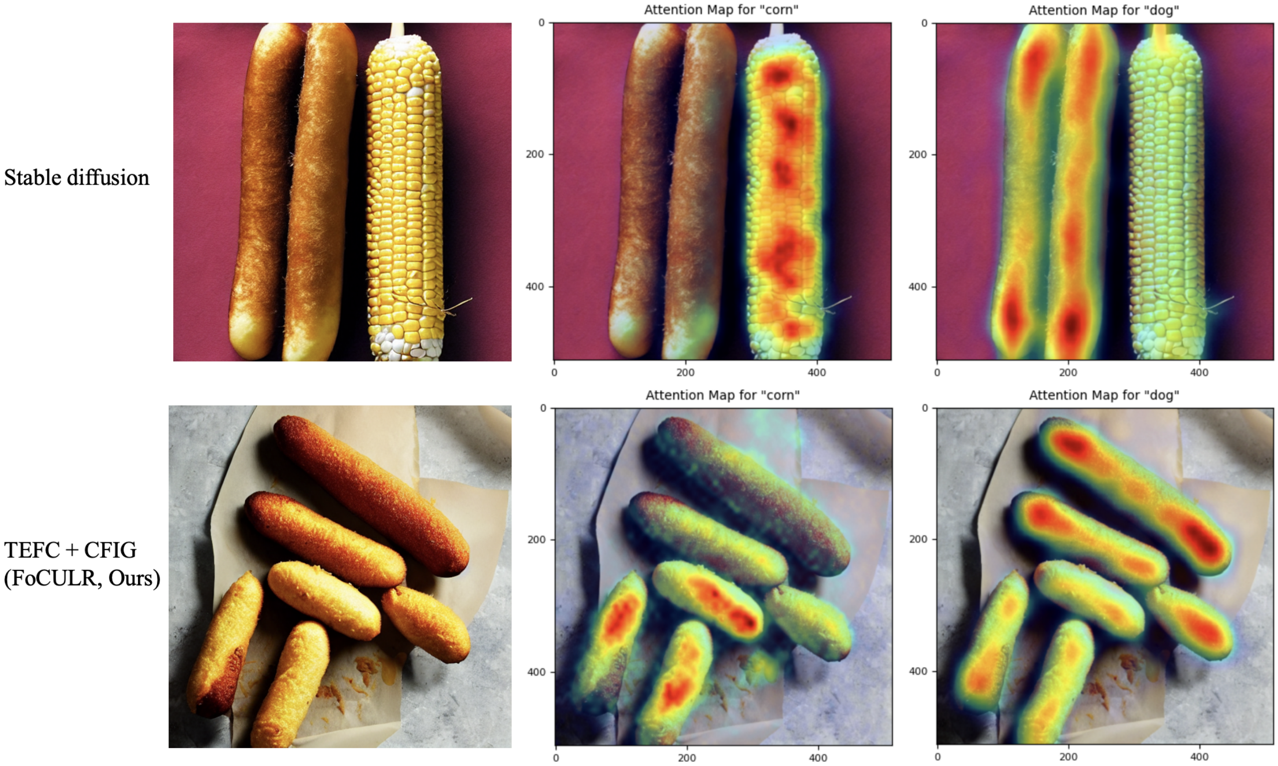

Fig. 3. Attention map for generated image with prompt of “A photo of a corn dog.” and negative prompt of “corn” if CFIG applied.

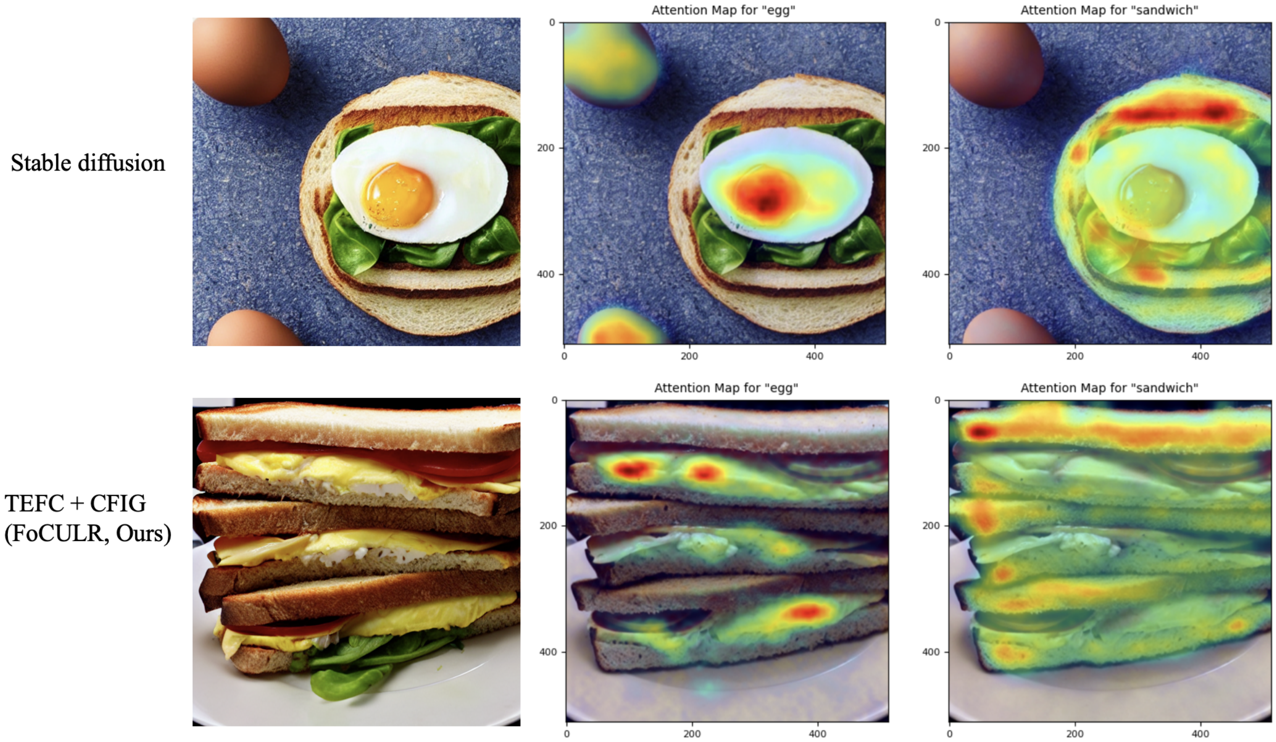

Fig. 4. Attention map for generated image with prompt of “A photo of a egg sandwich.” and negative prompt of “egg” if CFIG applied.

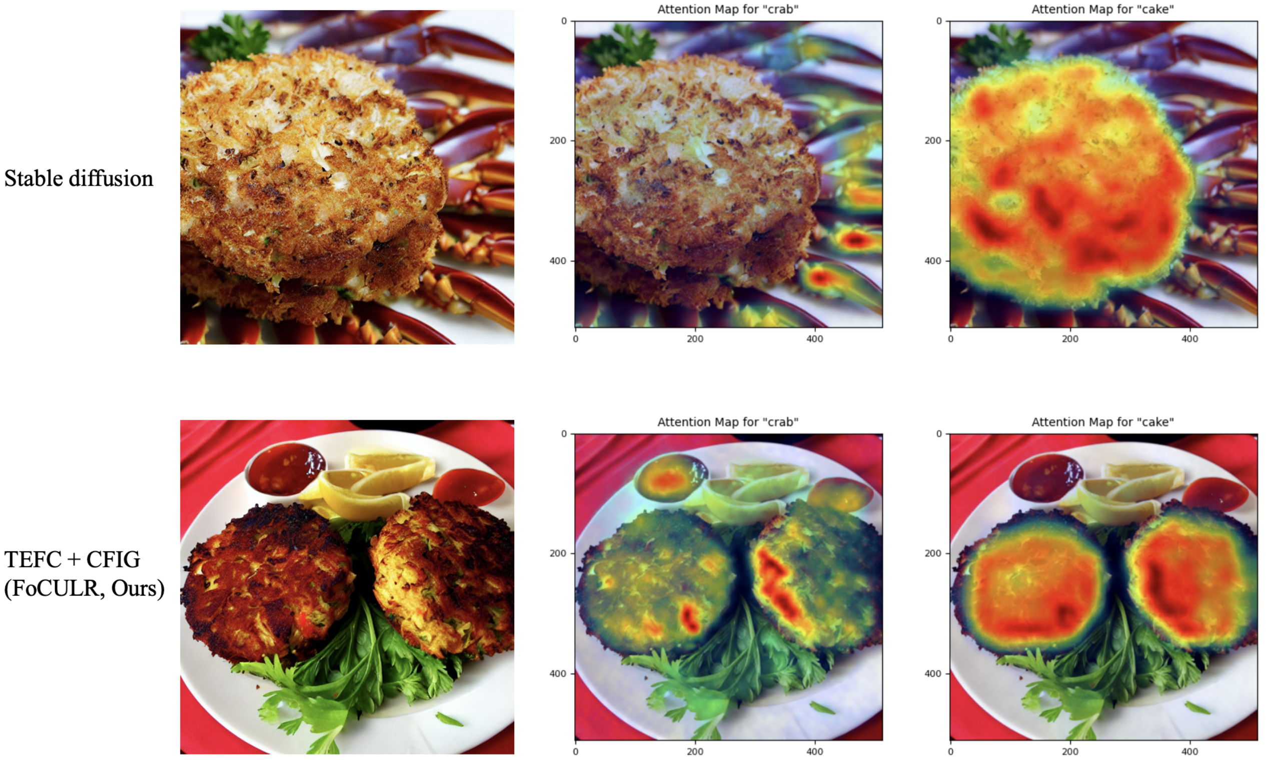

Fig. 5. Attention map for generated image with prompt of "A photo of a crab cake." and negative prompt of "crab" if CFIG applied.