RecipeSD: Injecting Recipe into Food Image Synthesis with Stable Diffusion

Jing Yang
yang-j@mm.inf.uec.ac.jp
The University of Electro-Communications
Tokyo, Japan

Junwen Chen
chen-j@mm.inf.uec.ac.jp
The University of Electro-Communications
Tokyo, Japan

Keiji Yanai
yanai@mm.inf.uec.ac.jp
The University of Electro-Communications
Tokyo, Japan

Abstract

In this paper, we introduce RecipeSD, a novel approach for food image synthesis using Stable Diffusion, enhanced by integrating recipe text information. RecipeSD leverages a pretrained recipe encoder from a cross-modal retrieval task to extract embeddings from recipe texts, transforming these embeddings into image-like representations using Image-like Recipe Transformation (IRT). Specifically, we propose CookNet, a model that utilizes these transformed embeddings to inject detailed recipe information into the diffusion model, significantly improving the quality and realism of generated food images. Our method also supports the integration of other ControlNets, providing additional control and further enhancing image fidelity. Experimental results demonstrate that RecipeSD can generate high-quality food images that closely align with the corresponding recipe texts, marking a significant advancement in cross-modal image synthesis.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

image synthesis, diffusion model, ControlNet, cross-modal recipe embedding

ACM Reference Format:
Jing Yang, Junwen Chen, and Keiji Yanai. 2024. RecipeSD: Injecting Recipe into Food Image Synthesis with Stable Diffusion. In *Proceedings of the 2nd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3688867.3690173

1 INTRODUCTION

In recent years, researches on image generation using diffusion models have become widespread. Diffusion models are a technique that gradually diffuses noise into an image, enabling stable image generation. In comparison to GANs, diffusion models have shown significant improvements in the quality and diversity of generated images. As an implementation of diffusion models, Stable Diffusion [9] which is trained with five billion text-image pairs is the most popular and easy to use since it is fully open-sourced.

McGE '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1194-7/24/10
https://doi.org/10.1145/3688867.3690173

However, Stable Diffusion faces challenges in reproducing the appearance of objects or interactions between objects in images because it relies on prompts for image generation. ControlNet [15] addressed this issue by incorporating conditional images, which enable the specification of desired layouts and interactions between objects, leading to high-quality image generation. However, the drawback of ControlNet lies in the high cost associated with creating conditional images. Particularly, preparing conditional images for food images, where the arrangement of dishes and mixing of ingredients present difficulties. Moreover, summarizing complex and lengthy textual information, such as recipe texts, into effective prompts for image generation is also presumed to be challenging.

To address the aforementioned issues, this study proposes RecipeSD. To overcome the challenge of utilizing recipe text to specify the layout of food images in Stable Diffusion, RecipeSD leverages a text encoder pretrained on cross-modal search tasks to extract recipe embeddings from recipe text, providing control information for image generation. Furthermore, to enable recipe text to control Stable Diffusion, this study introduces Image-like Recipe Transformation (IRT) to transform recipe embeddings. Consequently, the need to prepare conditional images is eliminated, enabling the generation of high-quality food images conditioned on recipe text. the contributions are summarized as follows:

(1) We propose a novel approach that utilizes pretrained recipe text embeddings from a cross-modal retrieval task for image generation using Stable Diffusion.
(2) We propose CookNet which can generate food images of high quality by controlling the Stable Diffusion model based on structural document information, which is recipe text.
(3) Furthermore, the proposed method CookNet can incorporate with other pretrained ControlNets to generate realistic food images with extra control.
(4) We are the first to adopt a diffusion model for recipe image generation. The experimental results demonstrate the ability of the proposed method to generate high-quality food images.

2 RELATED WORK

2.1 Food Image Synthesis with Generative Adversarial Networks

The method of synthesizing food images from recipe text on the Recipe1M dataset using Generative Adversarial Networks (GAN) [4] was proposed. CookGAN by Zhu et al.[16] utilizes a cooking simulator subnetwork to incrementally modify food images based on the interaction of ingredients and cooking methods, mimicking visual

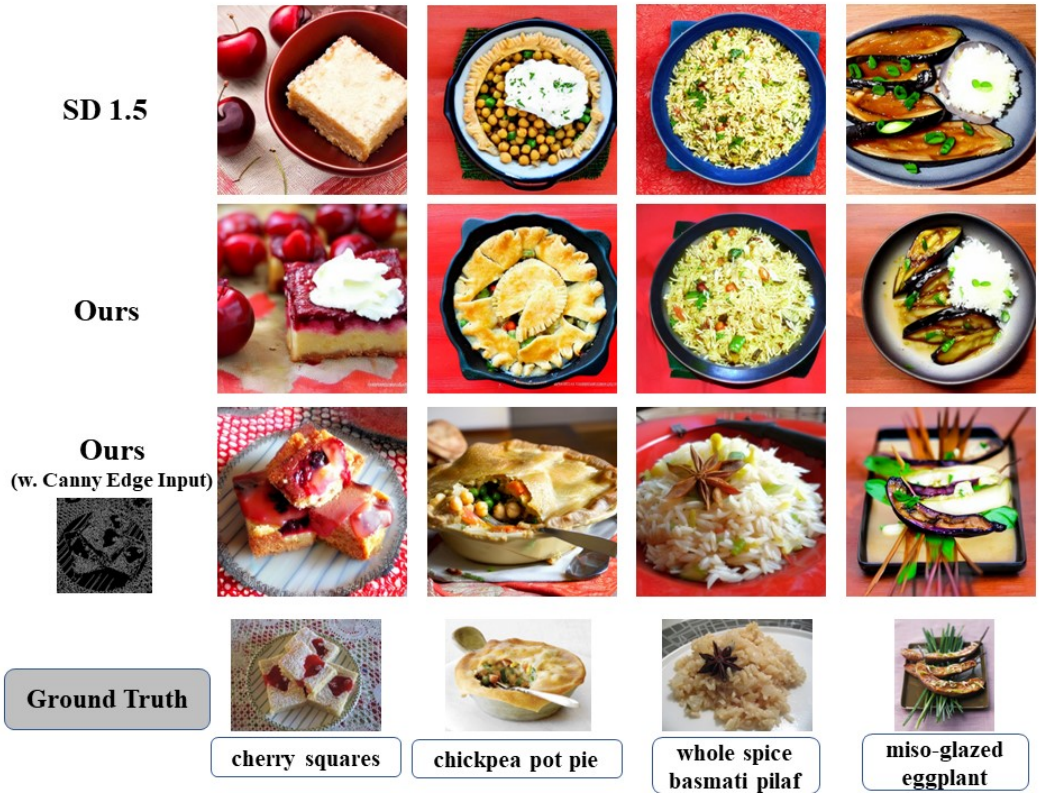

Figure 1: Comparison between RecipeSD and Stable Diffusion v1.5. The first row shows generated images using Stable Diffusion v1.5 as baseline. The second and third rows depict images generated using the proposed method CookNet.

effects. Another CookGAN by Han et al.[5] employs an attention-based ingredient-image relational model and applies the constraint of cycle consistency to ensure the quality of generated images. ChefGAN by Pan et al. [7] applies an additional regularization by utilizing a joint image-recipe embedding model before the generation task. Compared to the previous related works using GAN, we first introduce diffusion models for a food image synthesis task.

2.2 Food Image Synthesis with Diffusion Model

Latent Diffusion Models (LDM)[10] generate images in the latent space by iteratively adding noise during diffusion steps, enabling high-quality and stable image generation. Stable Diffusion (SD) [9] is a significant work on LDM. One of its major contributions is the realization of text-to-image generation using the large-scale dataset LION-5B in addition to LDM, establishing itself as a foundational model for image generation. Text embeddings are gradually mapped to high-resolution images using a pyramid structure in Imagen [13]. However, controlling the generation of images depicting object appearance or interactions between objects remains challenging for diffusion models relying on prompts. MakeAScene [3] and SpaText [2] controlled image generation through segmentation masks. GLIGEN [6] controlled image generation by learning new parameters in the attention layer of the diffusion model. DreamBooth [12] employed a reconstruction loss and a loss function for the object class in both Image-to-Image and Text-to-Image, achieving controlled image generation. Moreover, an efficient method was proposed in ControlNet [15]. Instead of directly training Stable Diffusion, ControlNet inserts a trainable additional network alongside the original Stable Diffusion for training. However, ControlNet has the drawback of requiring costly preparation of conditional images, such as poses. Users need to provide the appropriate conditional images for their desired images.

We propose CookNet, which addresses the high cost of conditional images by controlling image generation using embeddings learned from recipe text. Specifically, We use recipe text embeddings learned from cross-modal search tasks to control Stable Diffusion through Recipe Injector. While ControlNet controls image generation with costly conditional images, CookNet controls image generation using a 1024-dimensional recipe embedding, mitigating the issue of high conditional image costs.

3 Method

3.1 Overview

An overview of the methodology is illustrated in Figure 2. We proposed CookNet to inject the recipe information into the SD model. A pretrained recipe encoder from the cross-modal retrieval task is used to extract recipe text embeddings. Subsequently, we introduce preprocessing techniques to enable recipe text embeddings

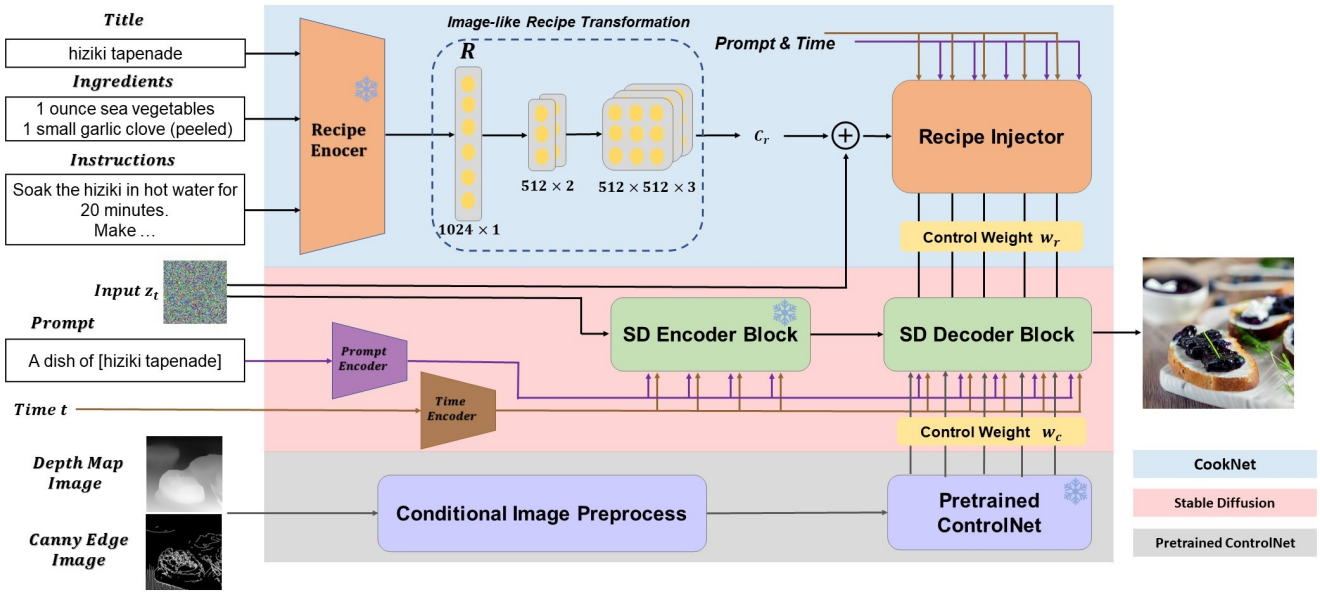

Figure 2: The overview of the proposed method, Recipe Stable Diffusion (RecipeSD), which consists of three parts: CookNet, ControlNet, and Stable Diffusion. CookNet is trained to inject recipe information into the diffusion model. Image-like Recipe Transformation (IRT) is used to transform the recipe text embeddings from the pretrained recipe encoder into an image-like representation. Additional conditional information is incorporated by using pretrained ControlNet.

to control the image generation of Stable Diffusion. Finally, we outline the learning methodology for image generation based on Stable Diffusion under the control of recipe text embeddings and additional conditional information.

3.2 Recipe Embedding Preprocessing

In this section, we introduce the preprocessing of recipe embeddings that will be input as conditions for our Recipe Injector. We adopt text encoder in TNLBT [14] in this paper. TNLBT is a framework for cross-modal recipe retrieval while incorporating CLIP model [8] and GAN [4]. The recipe embeddings R are obtained through the recipe encoder. Directly training Stable Diffusion with the recipe embeddings R as conditions is a straightforward approach. However, considering the expectation that recipe embeddings directly manipulate the generated images as conditions and the fact that the original ControlNet [15] takes images as input conditions, training with the 1024-dimensional recipe embeddings R alone is deemed insufficient. Therefore, in this study, we employ the Recipe Embedding Transformation module, Image-like Recipe Transformation (IRT), proposed in this research for preprocessing.

Figure 2 illustrates the method of preprocessing recipe text. The three components of the recipe text pass through respective encoders and are then merged through the Recipe Encoder to obtain the recipe embedding R. The encoders responsible for processing the recipe text in this stage are all frozen. These text encoders, obtained through pretraining in the recipe cross-modal retrieval, can appropriately project the recipe text into the corresponding recipe embeddings.

The recipe embedding R is transformed through the Recipe Embedding Transformation module IRT into a suitable form of embedding for ControlNet's learning. In IRT, the 1024×1 dimensional recipe embedding R is converted into a 512×2 dimensional vector. Subsequently, after expanding this vector to $512 \times 512 \times 3$ dimensions, the recipe condition embedding C_r controlling Stable Diffusion with Recipe Injector is prepared.

3.3 CookNet

Here, we introduce the learning of recipe embeddings to control image generation using Stable Diffusion. Recipe text embeddings serve as conditions to control the image generation of Stable Diffusion. Expecting recipe embeddings to have the ability to control image generation, the following properties are anticipated:

- Ability to reconstruct image information from recipe text.
- Capability for image generation based on recipe image embeddings.
- Distinct control between different recipe texts, with similar recipe texts exhibiting similar control.
- Leadership of recipe text in the process of generating images with Stable Diffusion, accurately reproducing the specified ingredients.

With these considerations, we decided to learn recipe text embeddings using a cross-modal retrieval method. Distance learning in cross-modal retrieval is suitable for similarity learning between similar entities while pushing away dissimilar ones. Therefore, it is suitable for similarity learning among recipe texts. Additionally, by using GAN for image generation from recipe text during the cross-modal retrieval learning process, it is possible to reconstruct

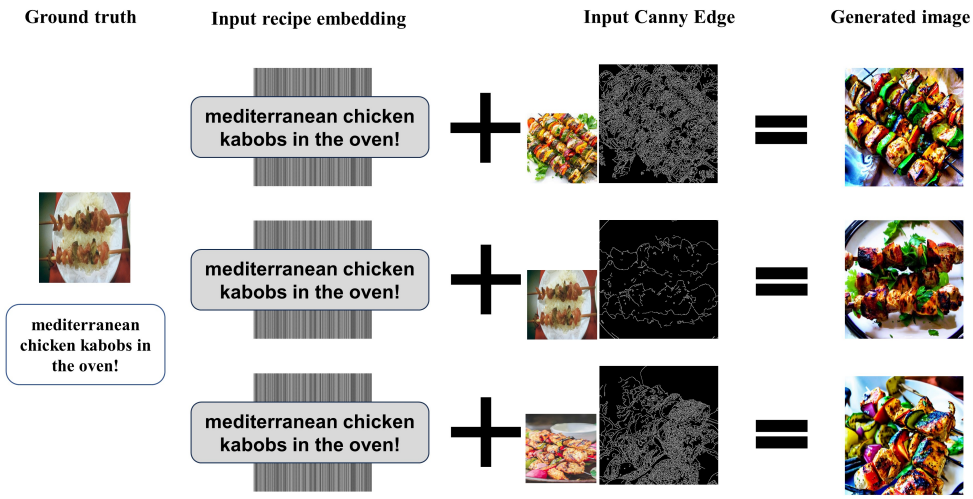

Figure 3: Image generation with a fixed recipe embedding combined with different Canny Edge images.

image information from recipe text. Finally, to ensure that recipe text leads the generation of correct ingredients in the image generation process, it is necessary to fix the embedding of recipe text. Therefore, utilizing a cross-modal retrieval with the large-scale language-image model CLIP seems appropriate, as it ensures sufficient learning in recipe text, suitable for controlling the final Stable Diffusion for image generation.

Figure 2 depicts the architecture of the proposed method CookNet. The SD Encoder Block represents the encoder part of Stable Diffusion, consisting of convolutional blocks with resolutions 64×64, 32×32, 16×16, 8×8 from top to bottom. The SD Decoder Block constitutes the bottleneck and decoder parts of Stable Diffusion, comprising a 8×8 Middle Block and convolutional blocks from 8×8 to 64×64. Note that while the SD encoder block is frozen, the SD decoder is not frozen during learning to adapt the model to recipe image generation. Finally, processed recipe embedding controls stable diffusion to generate food images with Recipe Injector, which is a copy of SD encoder. During image generation, the prompt is set to "A dish of [dish name]." For instance, when generating images for the dish "Ramen", the input is set as "A dish of Ramen," along with the time step and input noise z_t, and fed into Stable Diffusion.

Recipe Injector is basically based on ControlNet [15], which utilizes the encoder of U-Net [11], consisting of convolutional blocks from 64×64 to 8×8 for the encoder part and a bottleneck with a convolutional block of 8×8. The preprocessed conditional recipe embedding C_r from the IRT module, after passing through the zero convolution proposed in ControlNet [15], is added to the noise of Stable Diffusion and input into Recipe Injector. The outputs of each block in Recipe Injector are injected into the decoder of Stable Diffusion through zero convolutions, which consist of respective zero convolution blocks, to control image generation. A control weight w_r is used to adjust the importance of the recipe information.

3.4 Food Image Synthesis with Extra ControlNets

Due to the flexibility of ControlNet, the proposed method can generate even higher-quality food images by combining recipe embeddings with other conditions. As shown in Figure 2, by introducing a pretrained ControlNet and a corresponding control weight w_c for different conditions such as edge-based control, each controlling the same Stable Diffusion, faithful images under multiple conditions can be generated.

4 EXPERIMENTS

In this section, we present the experiments to validate the effectiveness of our proposed RecipeSD on food image generation from recipe texts.

4.1 Implementation Details

We used the train set in Recipe1M containing 238,999 pairs of data as training data, and used remained 102,422 pairs of data as test data. The recipe images in the training data were preprocessed by extending them to 512×512 through linear interpolation. Additionally, for each dish image, training was conducted with the corresponding prompt, "A dish of [recipe title]." For example, when training on data for berries romanoff, the input included the meal image along with "A dish of berries romanoff". The learning rate was set to 2×10^{-6}, and the batch size was set to 2. Training of RecipeSD took about 30 days on a single RTX 3090.

During the inference phase of the experiments, if no additional instructions were provided, the prompt "A dish of [food title]" was used for image generation. In the inference experiments combining recipe embeddings with other conditions, the environment utilized the Web UI developed by AUTOMATIC1111 [1].

4.2 Image Generation

In this section, we evaluate image generation controlled by recipe texts with RecipeSD. Particularly, as the original ControlNet [15]

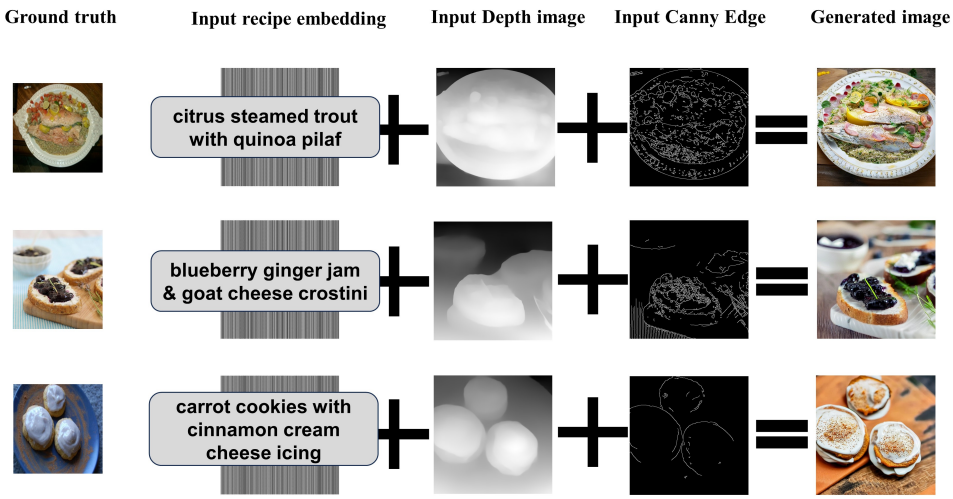

Figure 4: Image generation controlled by three conditions, recipe embeddings, Canny Edge, and Depth Map.

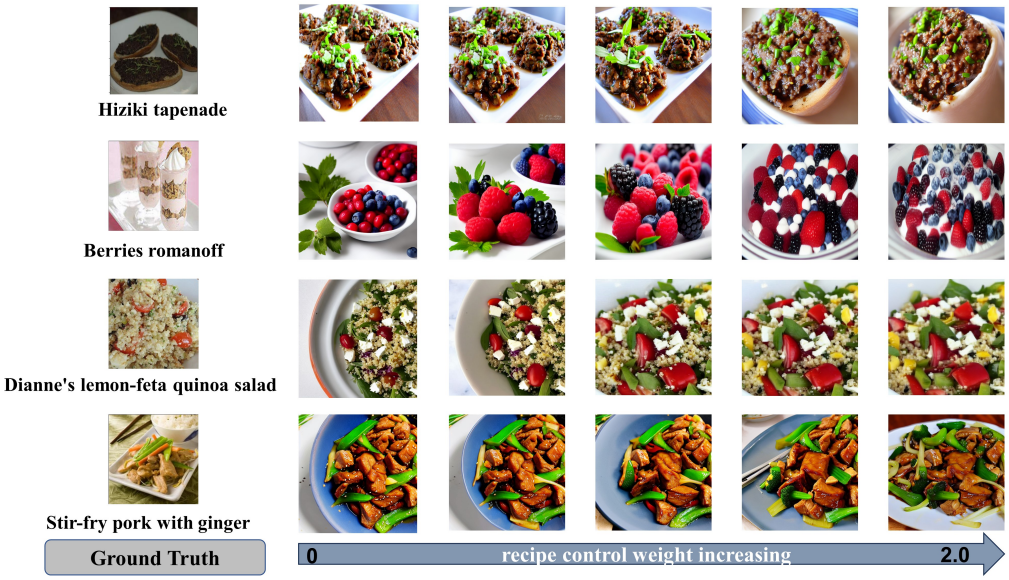

Figure 5: Image generation with different recipe embedding weights. The weights for the recipe embeddings increase gradually from 0 to 0.5, left to right, in increments, and finally settle at 2.0.

allows for multiple combinations, the proposed RecipeSD can also be simultaneously used with the pretrained Canny Edge ControlNet and Depth Map ControlNet. Therefore, in the experiments, we perform multi-condition image generation by combining recipe embeddings with Canny Edge and Depth Map. The pretrained ControlNet models for Canny Edge and Depth Map are obtained from the models released by the authors of ControlNet [1].

Figure 1 presents qualitative comparative experiments between the proposed method and original Stable Diffusion [9]. The first row shows generated images using Stable Diffusion v1.5, while the second and third rows depict images generated from recipe embeddings and images generated from a combination of recipe embeddings and Canny Edge, respectively. The experimental results indicate that the proposed method produces images that are more faithful to the recipe texts. It enhances the visual quality of the generated images on top of those from Stable Diffusion, providing a richer representation of the ingredients in the dish. Especially noteworthy is that for the same recipe embedding, even when combined with different Canny Edge inputs, the proposed method consistently generates high-quality images, demonstrating the versatility of recipe embeddings, as shown in Figure 3.

Moreover, not only with Canny Edge but also by including Depth Map as an input along with recipe embeddings, we achieve results

[1] https://huggingface.co/lllyasviel/ControlNet

that bring the generated images even closer to real images, as shown in Figure 4.

4.3 Adjusting Recipe Embedding Weights

Here, we analyze the role of recipe embeddings in image generation. As depicted in Figure 5, as the weight of the recipe embedding, w_r, increases, the quality of generated images is improved, showing that recipe embedding plays a crucial role in manipulating image generation. For instance, in the first row depicting the dish "Hiziki Tapenade" when the weight of the recipe embedding is 0, only the meat takes shape, and there is no bread. Increasing the weight of the recipe embedding transforms the plate into bread, ultimately resulting in a realistic image of a meat bun dish. Similarly, in the second row featuring "Berries Romanoff" increasing the weight of the recipe embedding changes an image originally consisting only of berries into a Berries Romanoff closer to a real image. This observation highlights the significant role of recipe embeddings in image generation.

4.4 Role of Each Condition in Multi-Condition Image Generation

In this subsection, we explore the impact of recipe embeddings on multi-condition image generation. Figure 6 records the variations in weight when generating images of a soup dish called "creampy kielbasa and potato soup" by combining recipe embeddings with Canny Edge. From the figure, it is evident that recipe embeddings control the ingredients in the soup within the range provided by Canny Edge. Particularly, the first row illustrates the control over the number of potatoes in the soup. Additionally, when the weight of Canny Edge is 1.5, its control becomes more prominent, and changes in the weight of recipe embeddings are less pronounced. Combining this result with Figure 5, we can infer that recipe embeddings do not arbitrarily change the shape of the dish but rather control changes in accordance with the given conditions.

Furthermore, Figure 7 records the variations in weight when generating images of a sandwich dish called "curried turkey wraps" by combining recipe embeddings with Depth Map. Due to the higher degree of freedom in Depth Map and the sandwich dish compared to the soup dish, the changes in the generated food images based on the weight of recipe embeddings are relatively substantial. Especially when the weight of the Depth Map is 0, increasing the weight of recipe embeddings results in cleaner shapes of the food images, approaching real images. However, when the weight of Depth Map is 1.5, similar to the case with Canny Edge, its control becomes dominant, and the changes in the generated images based on the weight of recipe embeddings are small. Nevertheless, increasing the weight of recipe embeddings, in this case, results in sandwich colors closer to real images, demonstrating that recipe embeddings appropriately control meal image generation.

In conclusion, recipe embeddings were found to control meal image generation within given conditions rather than arbitrarily manipulating the food images.

4.5 Influence of Prompts in RecipeSD

Traditional diffusion models, including Stable Diffusion, control image generation based on prompts specified in the text. In this paper, we propose a method to control image generation in Stable Diffusion using recipe embeddings pretrained through cross-modal retrieval as additional conditions. Here, we explore the impact of prompts on image generation in the proposed RecipeSD.

Figure 8 illustrates image generation using the prompt "A dish" without specifying the food title. The image generation takes input from recipe embeddings and Canny Edge, where Canny Edge serves as supplementary information about the shape of the dish. The results show that high-quality images can be generated even without providing the food title in the prompt. However, as the food title is not specified, some details in the food images appear slightly distorted. This drawback could be mitigated to some extent by adjusting the prompts or extending the training time.

4.6 Reconstructing Food Image from Image Embedding

Due to the feature of cross-modal embeddings, the proposed method RecipeSD can also generate food images under the control of image embedding of the recipe, as shown in Figure 9. Similar to the image generation from recipe text, generated images from the combination of image embedding and Canny Edge can also be of very high quality. Note that the used image embeddings are obtained from the pretrained CLIP-ViT [8] image encoder in TNLBT [14].

5 Conclusion and Future Work

In this paper, we first introduce Stable Diffusion for recipe image generation and propose a novel approach, RecipeSD for synthesizing food images by injecting recipe information into the Stable Diffusion model. Our method leverages the power of a pretrained recipe encoder from a cross-modal retrieval task to generate detailed and high-quality food images that closely align with the corresponding recipe texts. By transforming recipe embeddings into image-like representations through our proposed Image-like Recipe Transformation (IRT) and integrating them with the CookNet model, RecipeSD enhances the fidelity and realism of the generated images. Additionally, we demonstrated that our approach can be further enhanced by incorporating other ControlNets, allowing for additional layers of control in image generation. RecipeSD marks a significant advancement in food image synthesis, providing a flexible and powerful tool for generating realistic food images directly from textual descriptions. In the future, we plan to explore the integration of more fine-grained conditional controls to further enhance the quality and relevance of generated images.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers, 22H00540 and 22H00548.

Figure 6: Image generation with varying weights of recipe embeddings and Canny Edge. The weight of recipe embeddings increases gradually from 0 to 2.0 in increments of 0.5, starting from the left image. The weight of Canny Edge increases from 0 to 1.5 in increments of 0.75, starting from the top image.

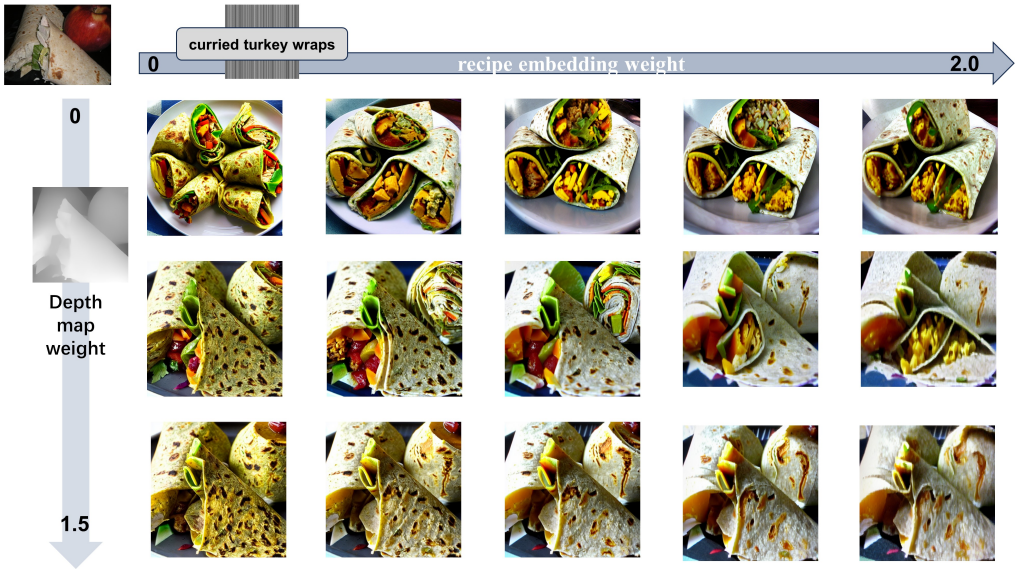

Figure 7: Image generation with varying weights of recipe embeddings and depth map. The weight of recipe embeddings increases gradually from 0 to 2.0 in increments of 0.5, starting from the left image. The weight of Depth Map increases from 0 to 1.5 in increments of 0.75, starting from the top image.

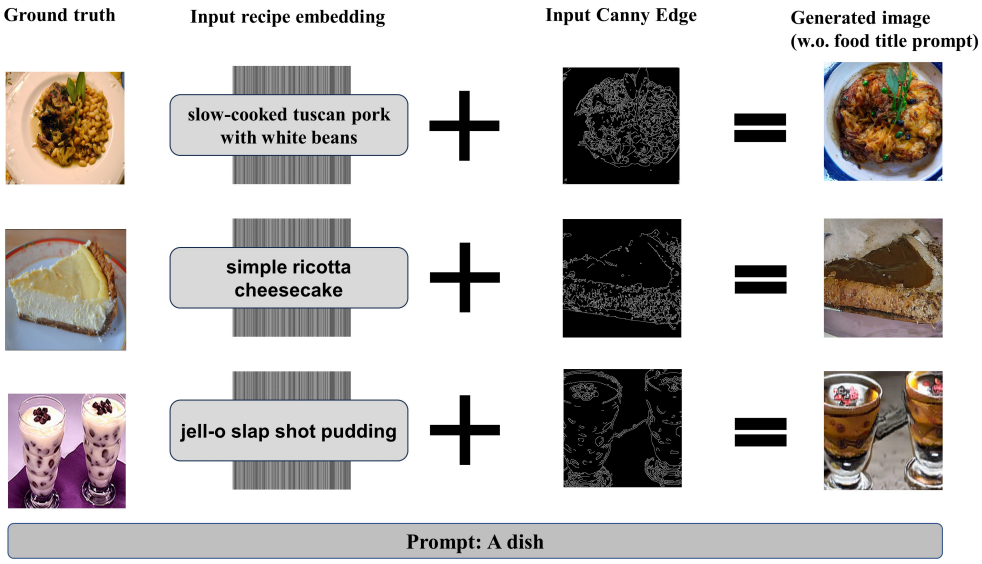

Figure 8: Image generation without providing the food title in the prompt. Food images are reproduced from the recipe text using recipe embeddings and Canny Edge.

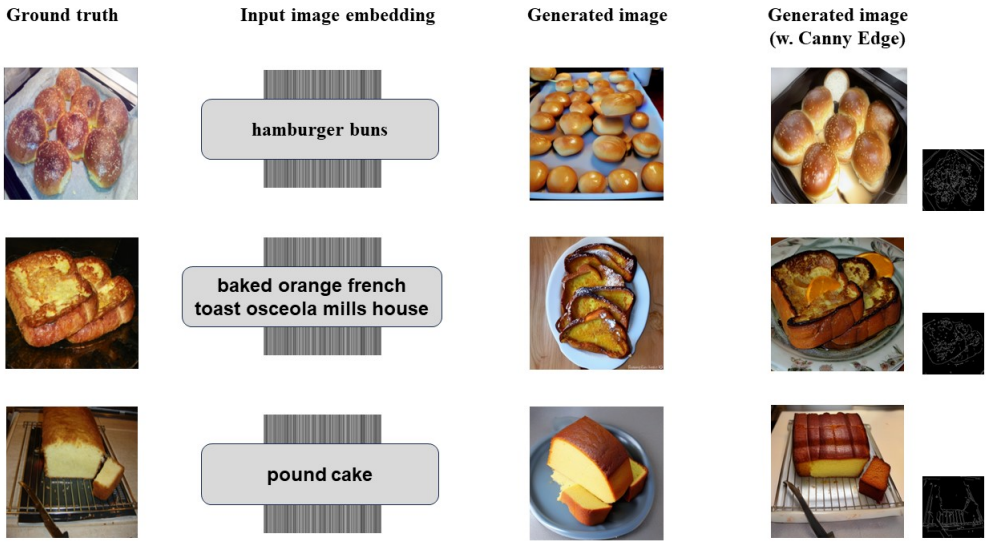

Figure 9: Image generation from image embeddings. Note that similar to recipe embeddings, the combination of Canny Edge and image embedding can also obtain high-quality results.

References

[1] AUTOMATIC1111. 2022. *Stable Diffusion Web UI.* https://github.com/AUTOMATIC1111/stable-diffusion-webui

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proc.of IEEE Computer Vision and Pattern Recognition.*

[3] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Proc.of European Conference on Computer Vision.*

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[5] Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic. 2020. CookGAN: Meal Image Synthesis from Ingredients. In *Proc. of IEEE/CFV Winter Conference on Applications of Computer Vision.*

[6] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *arXiv:2301.07093.*

[7] Siyuan Pan, Ling Dai, Xuhong Hou, Huating Li, and Bin Sheng. 2020. ChefGAN: Food Image Generation from Recipes. In *Proc.of ACM International Conference Multimedia.* 4244–4252.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of International Conference on Machine Learning*, Vol. 139. 8748–8763.

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc.of IEEE Computer Vision and Pattern Recognition.* 10684–10695.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc.of IEEE Computer Vision and Pattern Recognition.*

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI.*

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proc.of IEEE Computer Vision and Pattern Recognition.* 22500–22510.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *arXiv:2205.11487.*

[14] Jing Yang, Junwen Chen, and Keiji Yanai. 2023. Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training. In *Proc. of the International Multimedia Modeling Conference (MMM).*

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proc.of IEEE International Conference on Computer Vision.*

[16] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality Based Text-to-Image Synthesis. In *Proc.of IEEE Computer Vision and Pattern Recognition.*