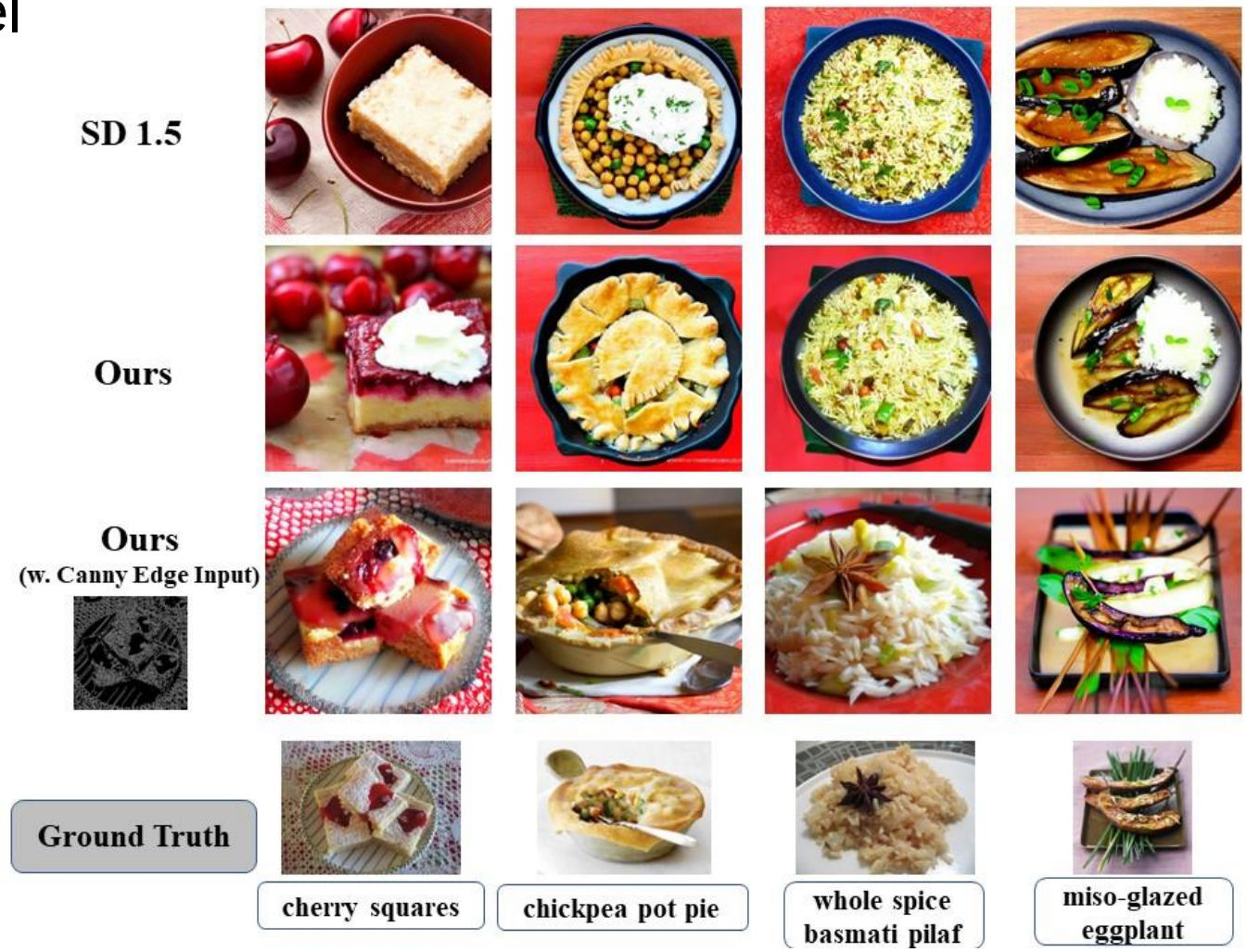
RecipeSD: Injecting Recipe into Food Image Synthesis with Stable Diffusion

Jing Yang Junwen Chen Keiji Yanai

The University of Electro-Communications

☐ RecipeSD

- Transfer cross-modal retrieval model knowledge to SD
- Introduce recipe text embeddings
- Incorporate with ControlNets



Recipe Retrieval and Dataset

☐ Modality

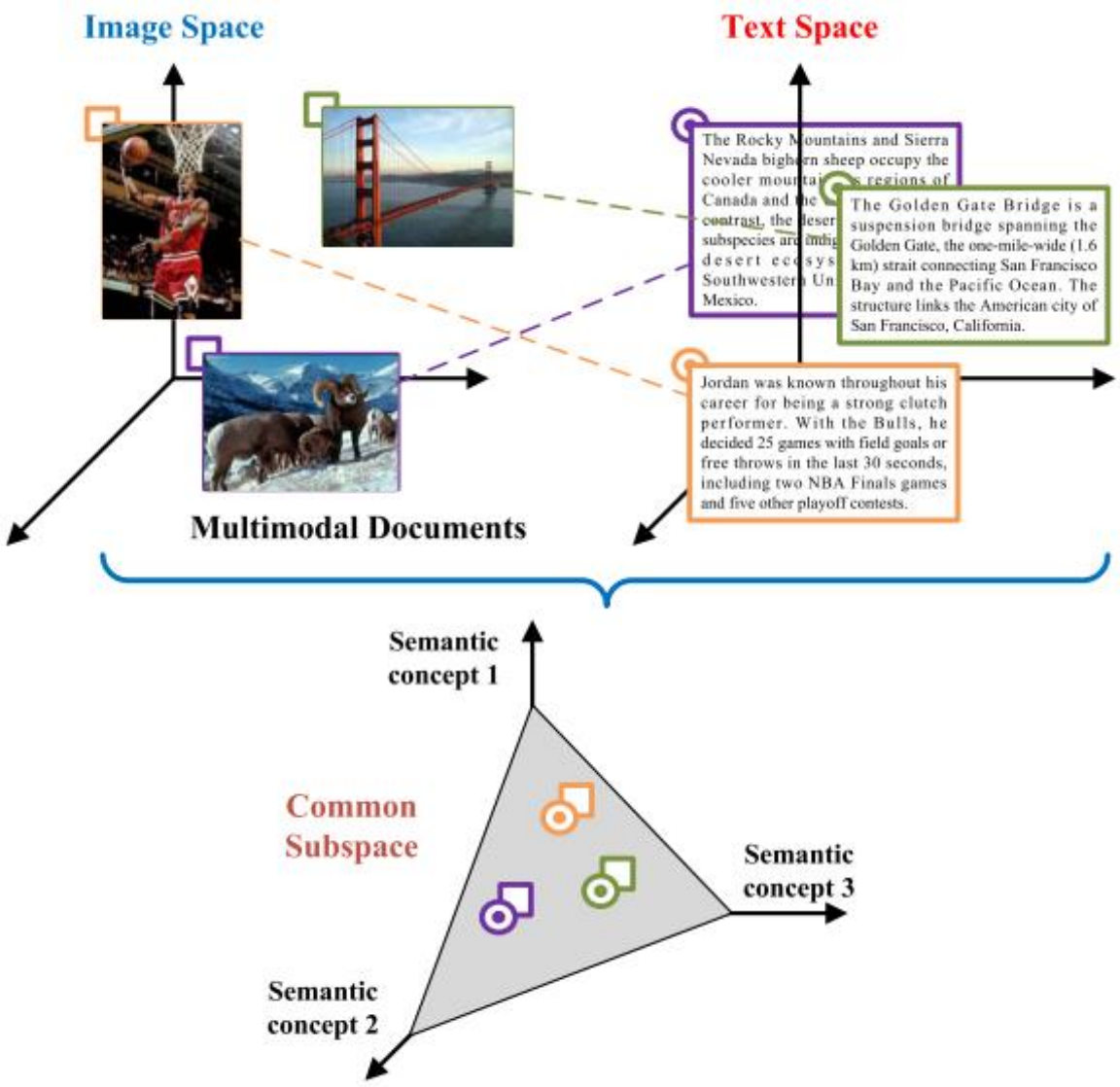
- Text, image, audio, video...

☐ Cross-modal image-text retrieval

- Build the connection is difficult
  - ➡ The gap between modalities

☐ Solution

- Embeddings & Distance Learning
- A large number of data pairs

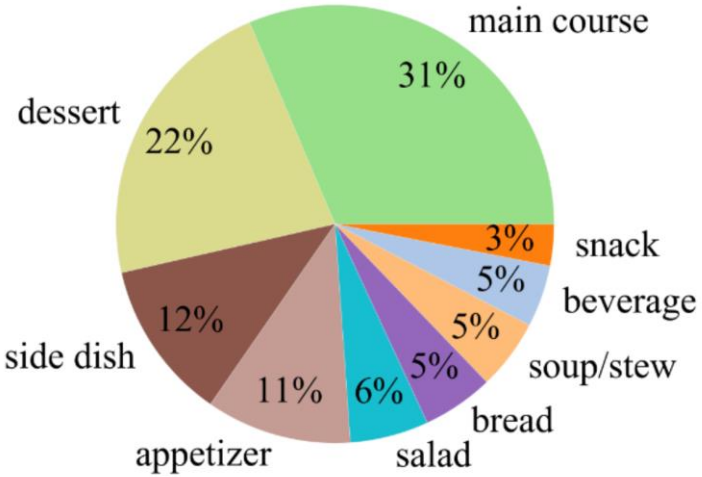


Guo, Wenzhong, Jianwen Wang, and Shiping Wang. "Deep multimodal representation learning: A survey."

Recipe Retrieval and Dataset

☐ Recipe1M

- One of the applications of cross-modal retrieval
- 1 million pairs of recipe images and recipe texts



Query Image

Retrieved Recipe

Ingredients	Instructions
butter olive oil sweet onions portabella mushrooms celery carrot garlic cloves ...	1. Melt 1 tablespoon butter with 1/2 tablespoon olive oil in saucepan over medium heat. 2. Add onions and saute, stirring every few minutes, until they are caramelized, about 15-20 minutes. ... 3. (If soup is too thick, thin with a little more hot broth). 4. Season to suit your taste with salt and freshly-cracked black pepper. 5. Serve in deep bowls, garnished with a sprinkle of minced, fresh parsley.

Query Recipe

Ingredients	Instructions
hamburger rigatoni pasta Ragu pizza sauce mushrooms onion pepperoni mozzarella cheese	1. Cook hamburger until done and drain off the fat. 2. Add mushrooms and onion and fry until translucent. 3. Add pepperoni. 4. ... 5. Lay noodles on top of hamburger mix in crockpot. 6. Turn crock on low and leave 4-5 hours. 7. Pour over the remainder of pizza sauce over the noodles. 8. Top with the cheese.

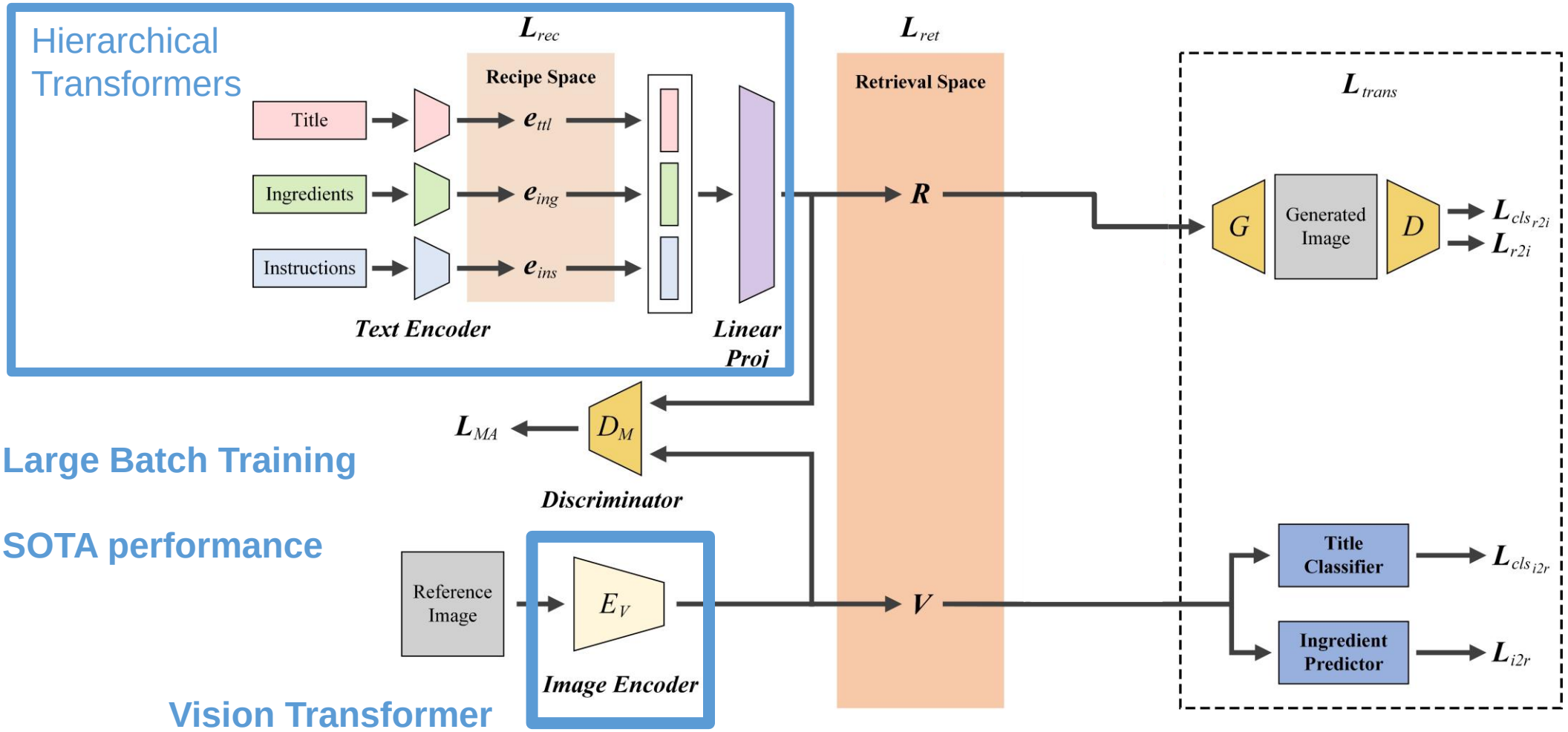
Retrieved Image

Salvador, Amaia, et al. "Learning cross-modal embeddings for cooking recipes and food images."

Recipe Retrieval Method

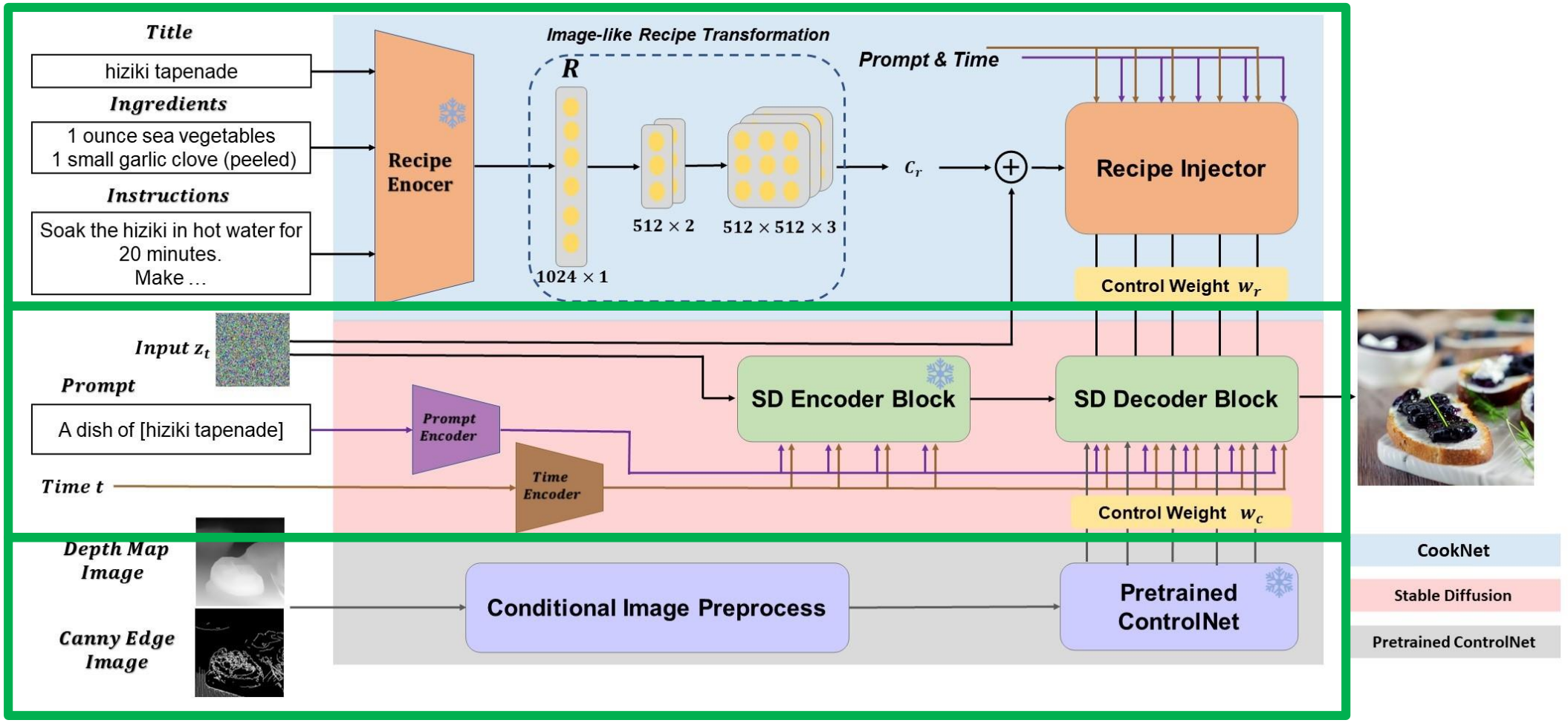
☐ TNLBT

- Improving the representation capability of the recipe embeddings

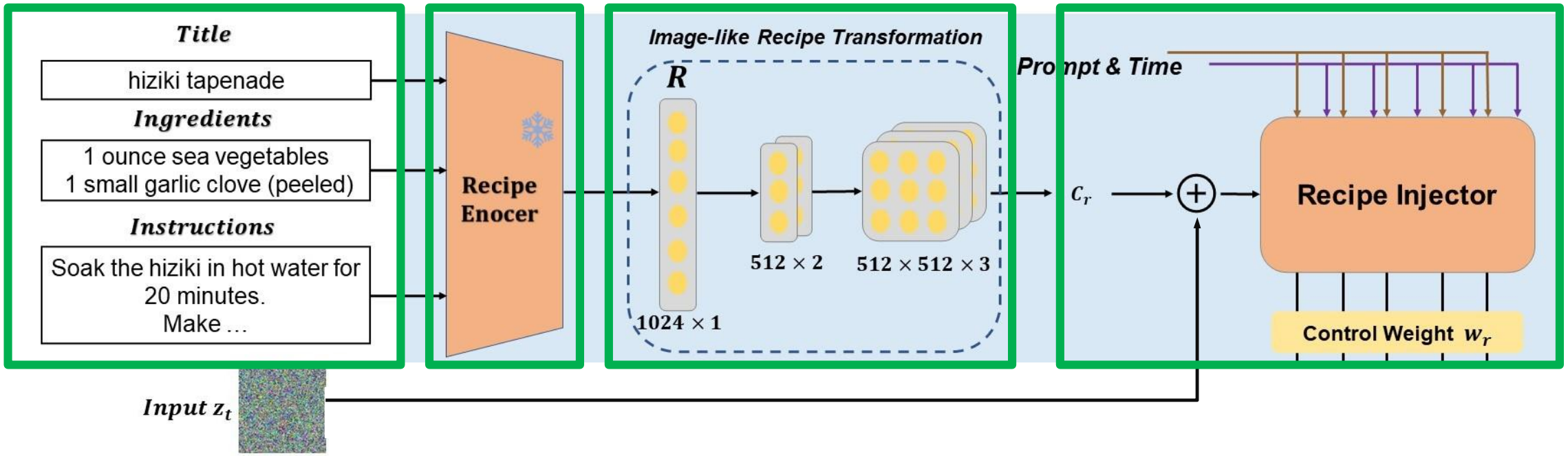


TNLBT : Yang, Jing, Junwen Chen, and Keiji Yanai. "Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training." MMM 2023

RecipeSD

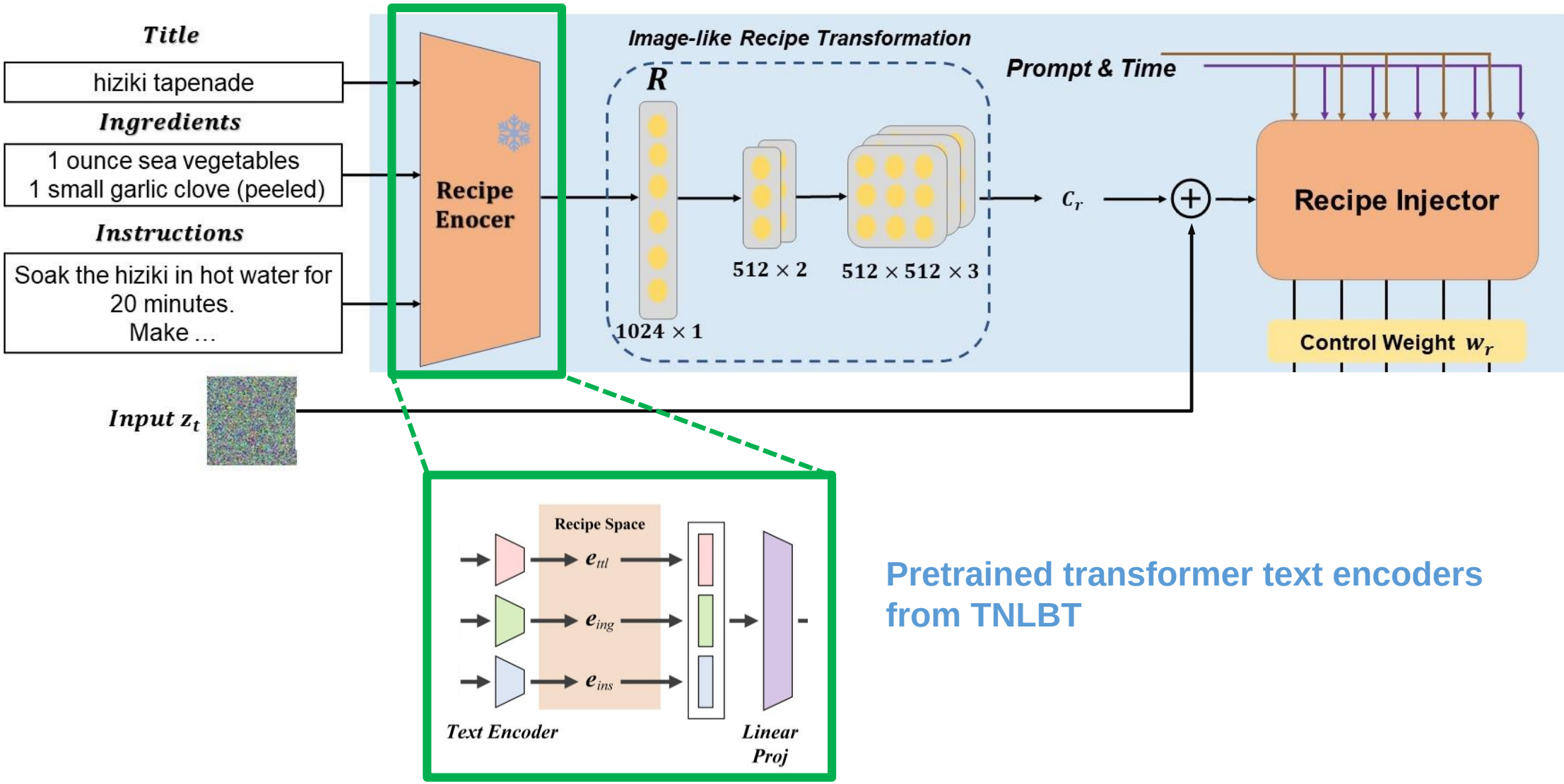


- Our RecipeSD consists of three parts: Stable Diffusion, CookNet, and ControlNet
- Image-like Recipe Transformation (IRT) transforms the recipe text embeddings to image-like embeddings



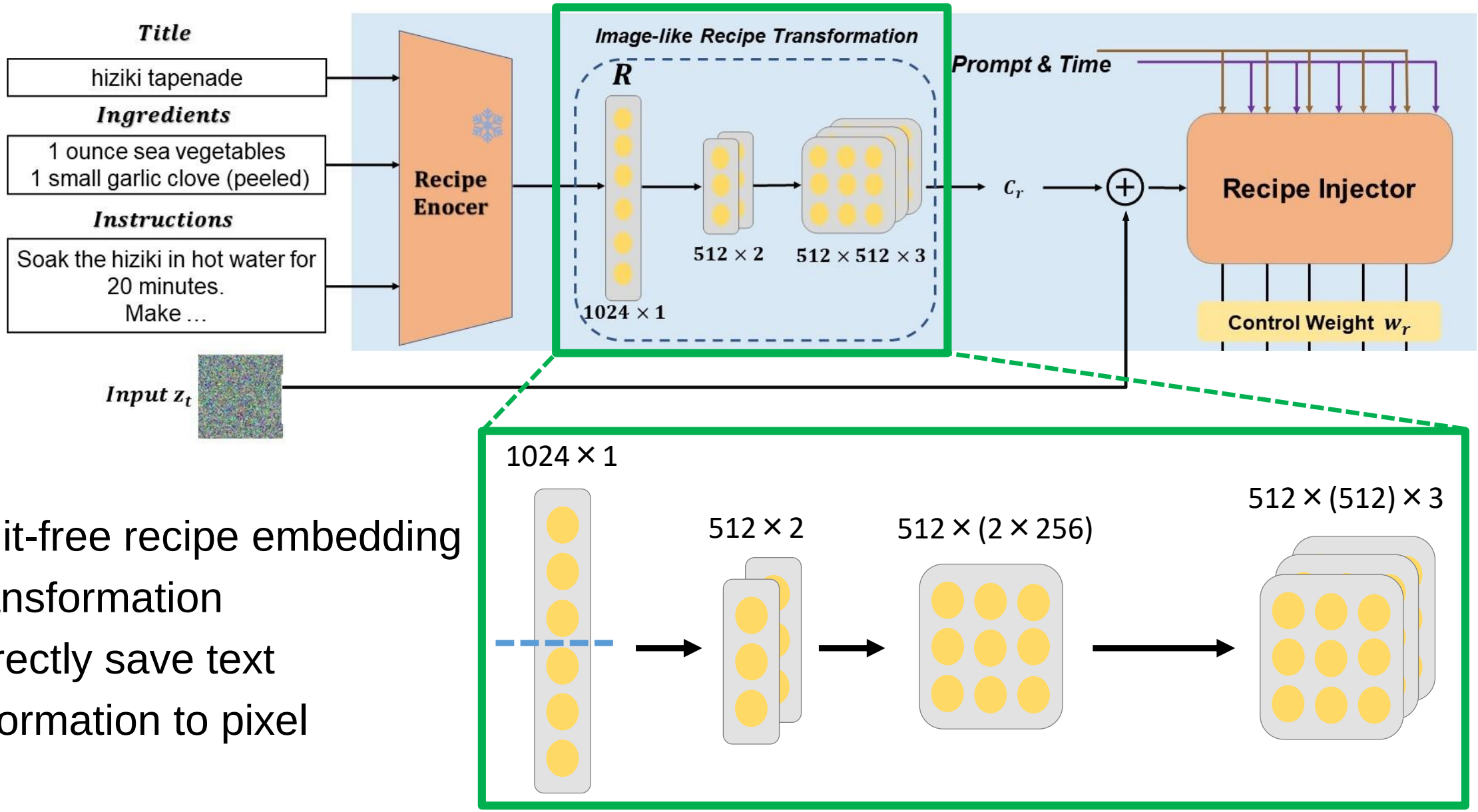
- Distinct control between different detailed recipe texts
- Image generation based on recipe image embeddings
- Reconstruct image information from recipe text
- Accurately reproducing the specified ingredients

CookNet

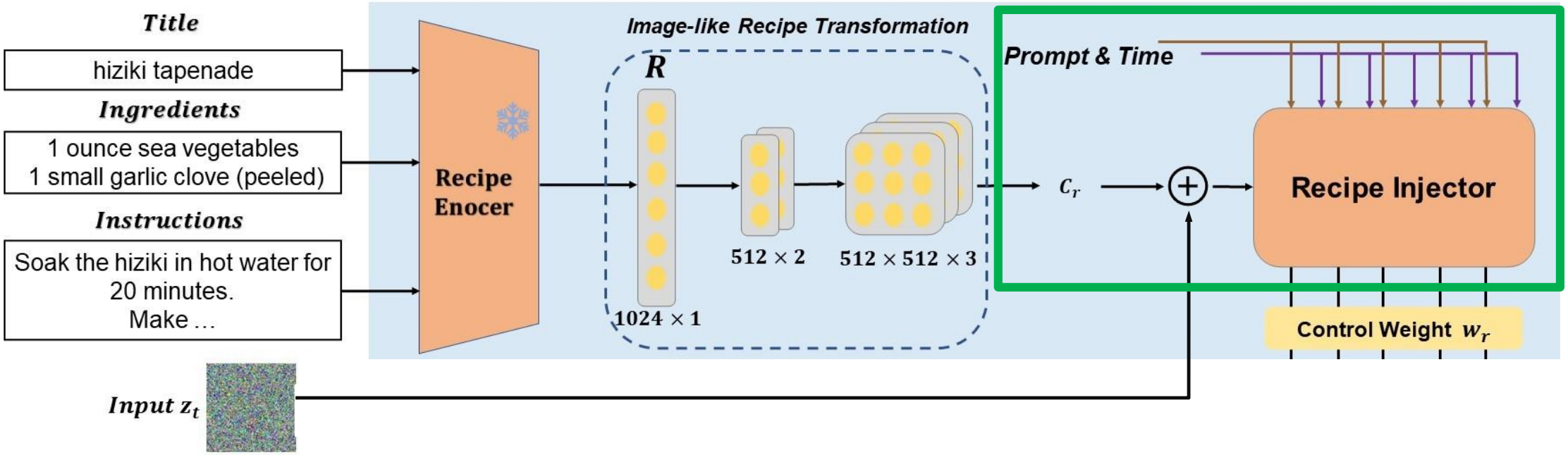


Pretrained transformer text encoders from TNLBT

- Prior knowledge from cross-modal retrieval task

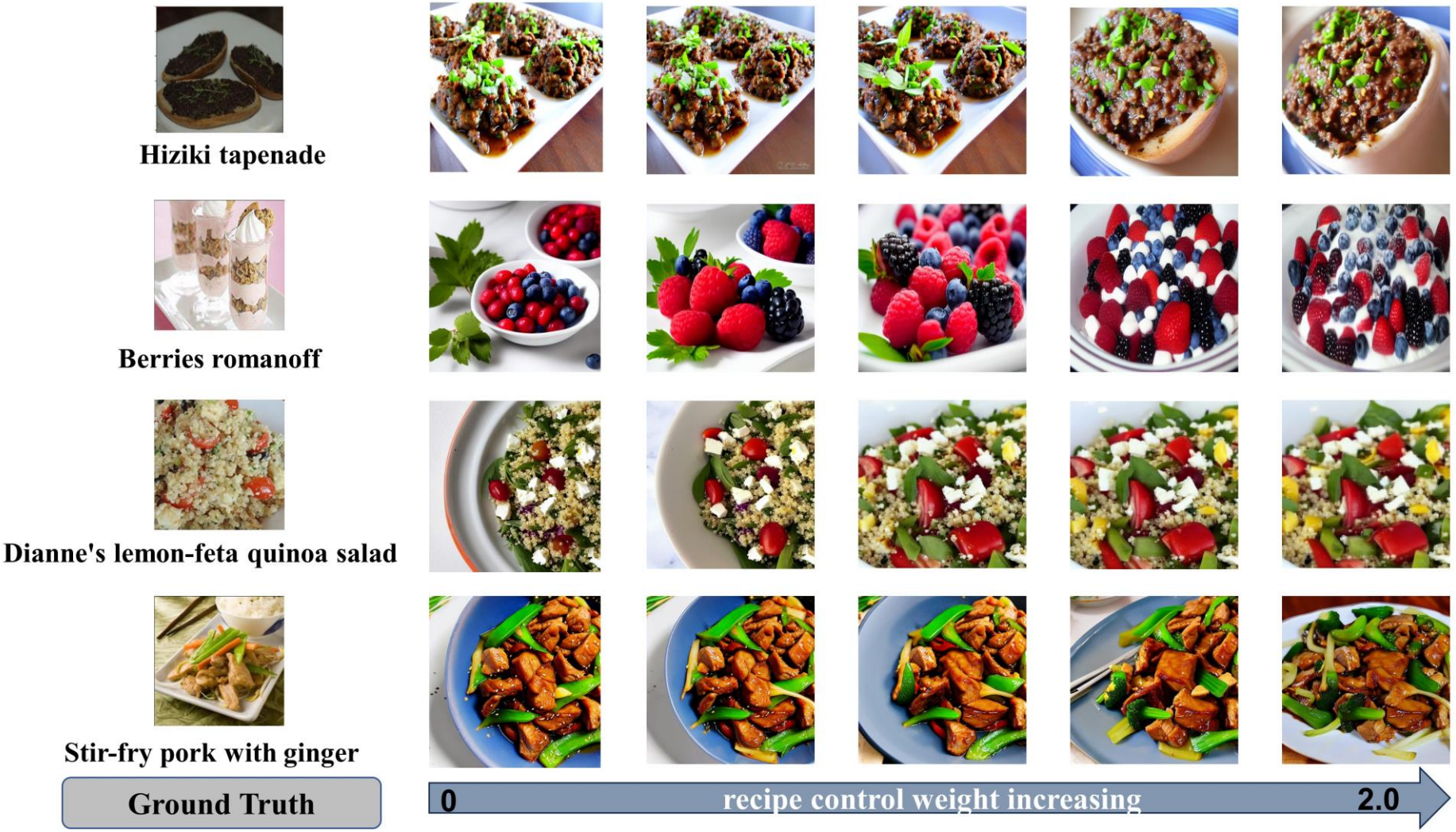


- Edit-free recipe embedding transformation
- Directly save text information to pixel



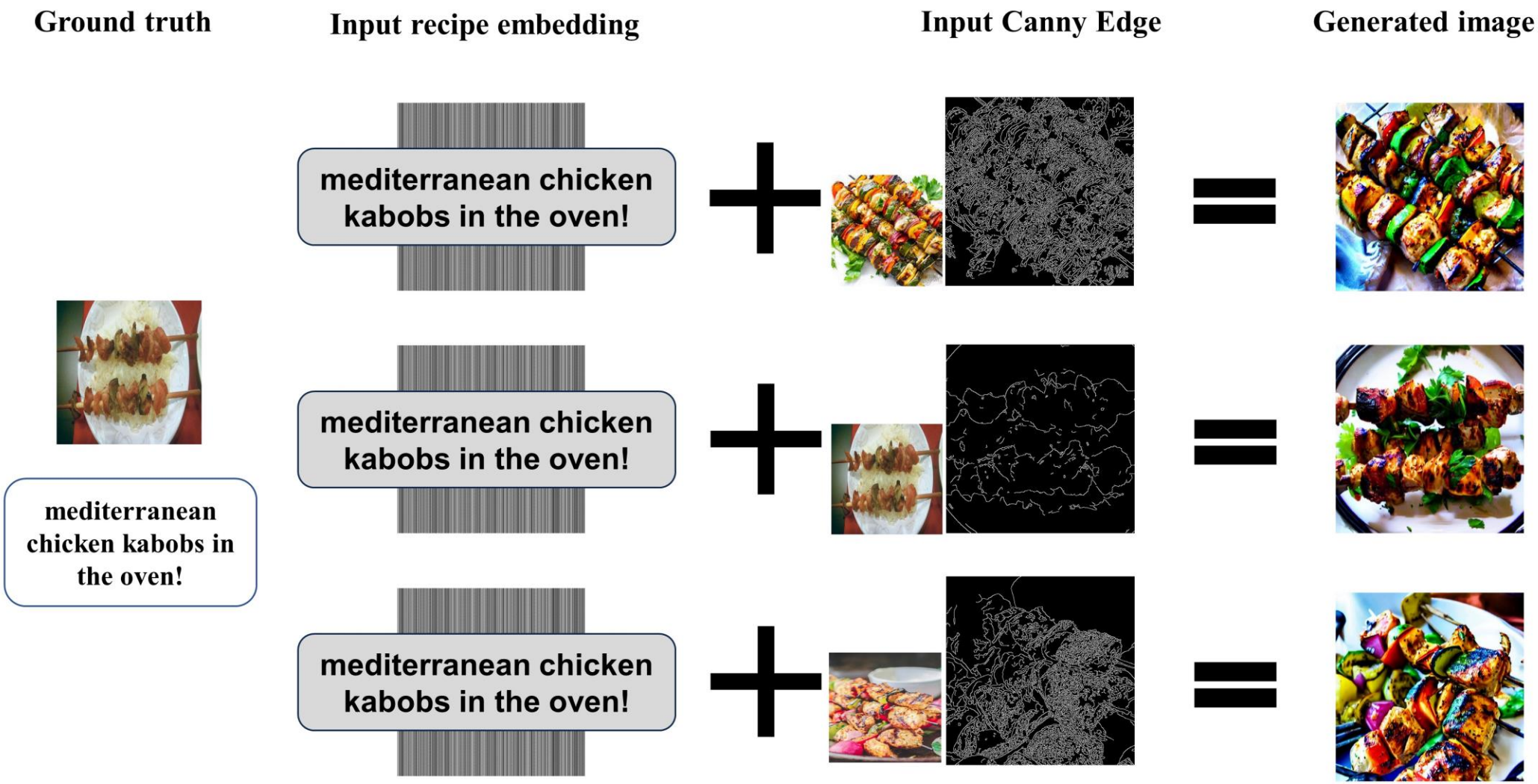
- Our Recipe Injector incorporate reconstruct text information from the beginning
- Accurately reproducing the specified ingredients

Effectiveness of Recipe Embeddings



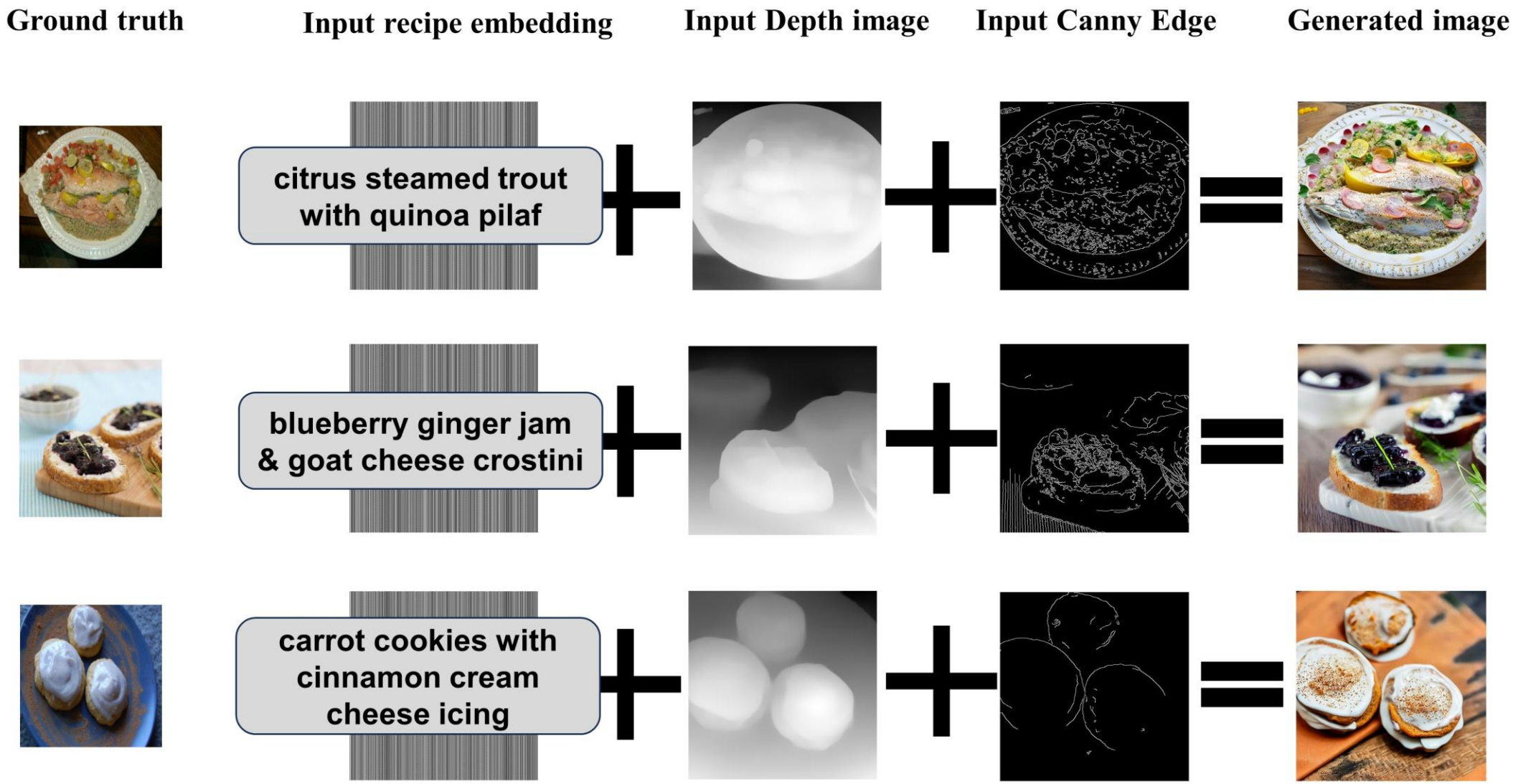
With the detailed recipe information from recipe text encoders and IRT in our CookNet, we increasing the control weight to improve the quality of generated images

Additional Conditions



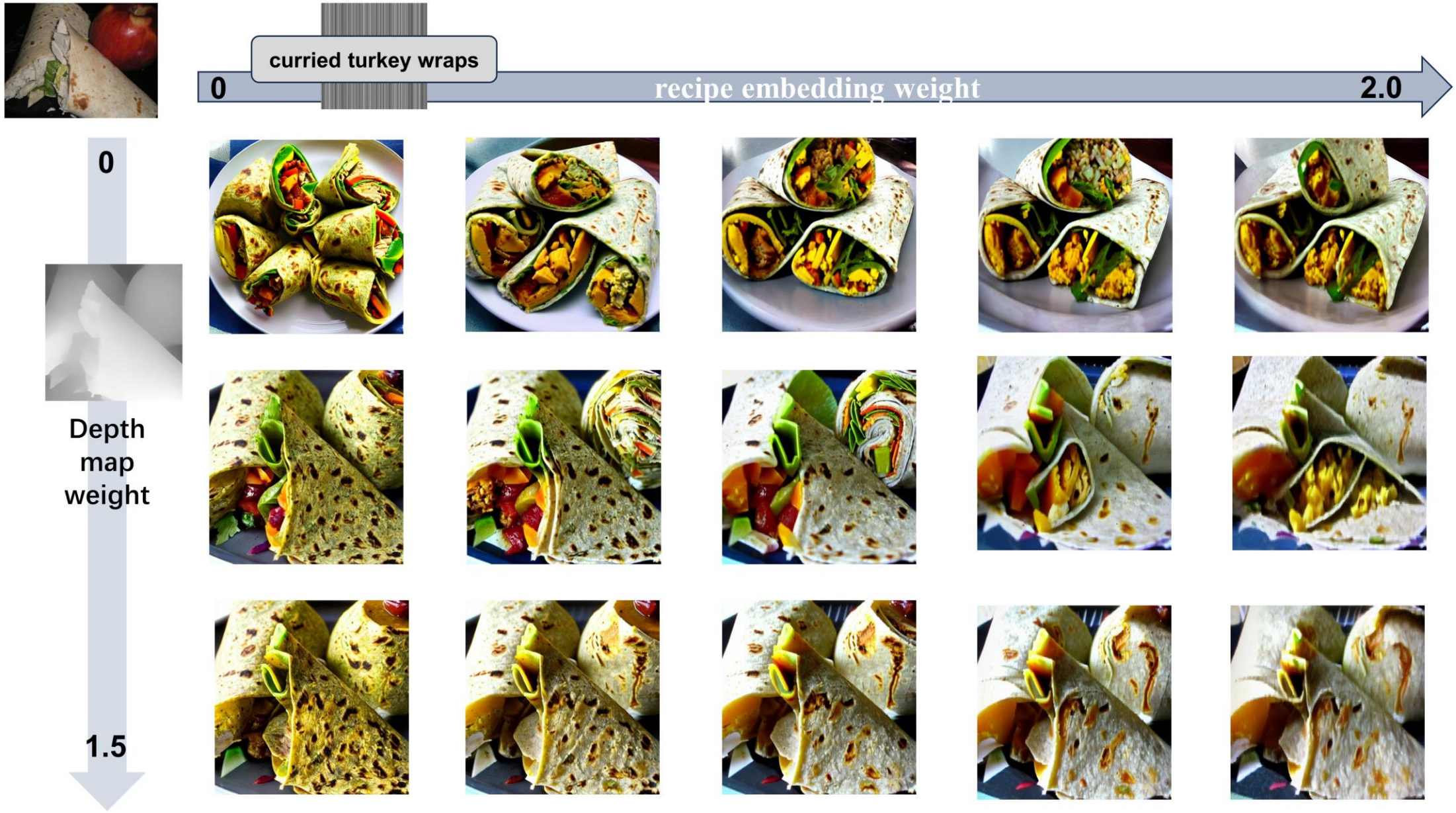
Recipe Embedding + Canny Edge

Additional Conditions



Recipe Embedding + Depth Image + Canny Edge

Weighted Condition Generation



Conclusion

- We introduce RecipeSD for synthesizing food images by injecting recipe information into the Stable Diffusion model
- We proposed CookNet with task-specific recipe encoders to align the generated images with the corresponding recipe texts
- We demonstrated that our approach can be further enhanced by incorporating other ControlNets


UEC