2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-6547-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPRW63382.2024.00377

A Generative Exploration of Cuisine Transfer

Philip Wootaek Shin* Ajay Narayanan Sridhar* Jack Sampson
Vijaykrishnan Narayanan
The Pennsylvania State University
{pws5345,afs6372,jms1257,vxn9}@psu.edu

Abstract

Recent research has made significant progress in text-to-image editing, yet numerous areas remain under explored. In this work, we propose a novel application in the culinary arts, leveraging diffusion models to adjust a range of dishes into a variety of cuisines. Our approach infuses each dish with unique twists representative of diverse culinary traditions and ingredient profiles. We introduce the ***Cuisine Transfer*** *task and a comprehensive framework for its execution, along with a curated dataset comprising over 1600 unique food samples at the ingredient level. Additionally, we propose three Cuisine Transfer task specific metrics to accurately evaluate our method and address common failure scenarios in existing image editing techniques. Our evaluations demonstrate that our method significantly outperforms baseline models on the Cuisine Transfer task.*

1. Introduction

Recent advances in generative models have opened up fundamentally new areas of research affecting nearly all aspects of society from food to finance. Traditionally, generating realistic images and videos from textual descriptions has posed a formidable challenge. Conventional approaches such as Generative Adversarial Network[8] often struggled to capture the intricate nuances of visual content, leading to outputs that lacked coherence and fidelity. However, with the introduction of diffusion models [12, 21, 22], this landscape has undergone a paradigm shift. These models, inspired by the fundamental principles of statistical physics and probabilistic inference, have emerged as powerful tools for understanding and generating complex visual data. Leveraging concepts from diffusion processes and stochastic sampling, these models excel at synthesizing high-quality images and videos that closely align with textual input, achieving unprecedented levels of realism and detail.

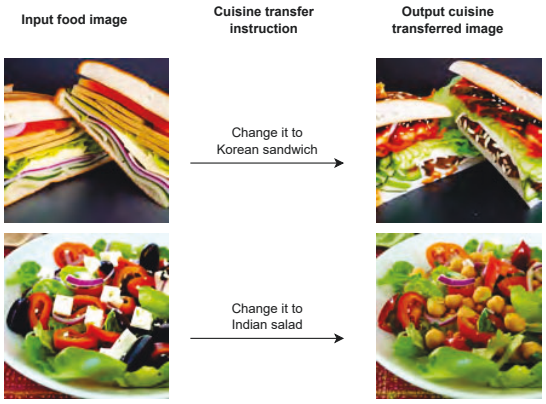

Figure 1. Illustration of our proposed Cuisine Transfer task.

Building upon the foundational principles of diffusion models, we propose an innovative application that demonstrates their efficacy in the realm of culinary arts. Inspired by the fusion of cultural diversity and culinary creativity, we leverage diffusion models to create authentic variants of a given dish across a range of cuisines. By utilizing existing cuisines as a reference, our approach, described pictorially in Fig. 1 infuses each dish with a unique twist representative of the culinary traditions and flavor profiles of various nations. Our methodology is versatile, with potential applications across multiple areas outlined in Tab. 1.

As a novel application of generative models, cuisine transfer presents some challenges to explore. Specifically, existing image generation approaches[1, 3, 5, 16] do not explicitly account for ingredient semantics, there are few datasets available for bench marking, and no task-specific quantitative metrics for the success or failure of the cuisine transfer task. In this work we examine integration of diffusion models with ingredient-level analysis, create dataset with cross-cuisine recipe synthesis and transformation, and propose advanced metric development for evaluating cultural authenticity and culinary fidelity. Our contributions are as follows,

- We present a novel task, Cuisine Transfer, and a comprehensive framework for leveraging diffusion models in the generation of culturally authentic food

*These authors contributed equally to this work.

This work was supported in part by NSF Awards 2243979 and 2318101

2160-7516/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPRW63382.2024.00377

Aspect	Description
Gastronomic Creativity	**Automation**: Diffusion models automate culinary creativity by utilizing data to generate innovative dishes and providing guidance to chefs on the plating and presentation of cuisine from various styles. **Scaling**: Diffusion models in culinary innovation enable rapid scaling, facilitating exploration of diverse cuisines and styles while swiftly adapting to consumer preferences.
Cuisine Awareness	**Vegan Diet**: A diffusion model for cuisine transfer can increase awareness of vegan diet by producing aesthetically pleasing and palatable vegan food dishes of popular meat-based dishes. **Global Cuisine Exploration**: Cuisine Transfer amplifies awareness of diverse culinary traditions from around the globe, empowering individuals to vividly visualize and eagerly explore an array of new dishes.
Marketing	**Marketing for stores**: Store owners can use the diffusion model for cuisine transfer to market their establishment as a fusion of technology and traditional culinary arts, creating a unique, adaptable, and localized dining experience.

Table 1. Different potential target aspects for usage of cuisine transfer model

cuisines.

- We curate an ingredient level detailed food image cuisine dataset with over 1600 unique samples.
- We propose three Cuisine Transfer specific metrics which accurately capture common failure scenarios of existing image editing works.

2. Background & Related Work

In the domain of culinary image generation, several studies have been conducted. Notably, the Adversarial Cross-Modal Embedding (ACME)[23] framework was introduced, aiming to learn a shared embedding space between cooking recipes and food images. This methodology employs the recipe embedding to generate a corresponding food image and utilizes the food image embedding to deduce the ingredients within the dish. The most advanced contribution in this field is FoodFusion[15], which employs a specially designed Latent Diffusion model to create realistic food images based on textual descriptions. LAIONFood was engineered by the same authors, employing prompt engineering techniques to curate realistic food images for training purposes. However, the FoodFusion model has not been made publicly available, rendering its evaluation and validation by the broader scientific community unfeasible. The Recipe Ingredients Dataset[14], comprises 39,774 training instances that include recipe ID, cuisine type, and ingredients, along with 9,944 test instances containing identical categories of data. Utilizing this dataset, we identified the prevalent ingredients across different cuisines, which served as the empirical foundation for developing our FoodCuisine Ingredient Prompt.

2.1. Diffusion Model

Rombach *et al*.[20] introduced latent diffusion models (LDMs) that operate within a latent space generated by an auto-encoder, facilitating both forward and reverse processes. These models integrate cross-attention mechanisms, significantly enhancing their efficacy in applications such as conditional image synthesis.

Additionally, there have been strides in enabling diffusion models to edit images through ControlNet[25] or spatial masking[6] along with the text, leading to image translation models[13, 27] that ascertain the mapping from conditioning images to target images. Stochastic Differential Editing (SDEdit)[16] introduces a diffusion model-based approach that creates realistic images from user inputs, without requiring additional training or loss functions. Text2Live[1] enables zero-shot, text-driven edits on images and videos by generating a semantically relevant edit layer using a training set derived from the input and a pre-trained CLIP model, allowing for localized adjustments without pre-trained generators or manual masks, across diverse objects and scenes. DiffEdit[5] introduces a novel approach to semantic image editing by leveraging text-conditioned diffusion models to automatically generate masks for regions to edit,

A notable instance of this approach is InstructPix2Pix[3], which facilitates instruction-based 2D image editing by conditioning the diffusion model via text. In our study, we have facilitated the pretrained latent diffusion models as a means to generate paired images. Furthermore, we have utilized InstructPix2Pix to facilitate the style transfer between various cuisines, effectively altering the visual representation of foods.

3. Method

InstructPix2Pix is designed as a two-stage process to generate training image pairs. First, they generate (original image, edit instruction, edited image) triplet training sample using a large language model, GPT 3 [4] for edit instruction and LDM [20] for the original image and edited image. Further, using the above triplet dataset, the authors train a diffusion model for the task of image-editing. In our work, we modified their technique for the task of cuisine style transfer.

3.1. Dataset Generation

In order to encompass a wide range of culinary diversity, the study selected 20 distinct food dishes. Additionally, to complement the variety of dishes, 20 dis-

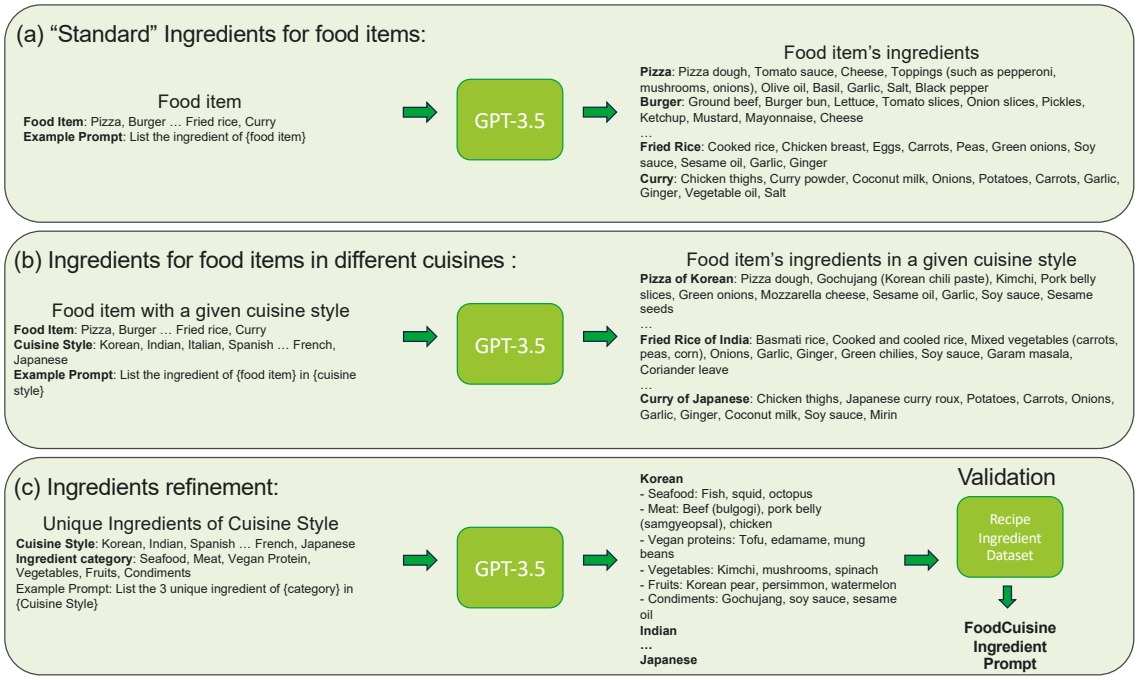

Figure 2. Prompt generation for ingredients of food items in distinct cuisine style: (a) Generating a comprehensive ingredient list of standard culinary preparation. (b) Generating cuisine-specific ingredients list of culinary items. (c) Refining the cuisine-specific ingredients list based on uniqueness

tinct cuisine styles were identified for inclusion. The food dishes and cuisines are described in Tab. 2. The adoption of these particular culinary styles was shaped by the Recipe Ingredient Dataset [14], which encompasses a diverse collection of food recipes, culinary traditions, and ingredients.

Cuisines		**Food dishes**	
Italian	Spanish	Fried Rice	Burger
Mexican	Southern US	Sandwich	Pasta
Indian	Chinese	Soup Noodle	Pancake
French	Thai	Savoury Pie	Stew
Japanese	Greek	Fried Noodles	Pizza
Korean	Vietnamese	Rolls (e.g. Maki)	Burritos
Moroccan	British	Savory Waffle	Crepes
Filipino	Irish	Fried chicken	Lasagna
Jamaican	Russian	Barbecued meat	Curry
Brazilian	Vegan	French Fries	Salad

Table 2. List of food dishes and cuisine styles

3.1.1 Food Ingredient Prompt Generation

We provide a schematic representation of the methodology employed for prompt generation in our study in Fig. 2. In the development of our dataset, we employed OpenAI’s ChatGPT 3.5 for the generation of captions and ingredient lists for a variety of food dishes. This process commenced with the generation of a standardized ingredient list for 20 distinct food items. The establishment of a standard recipe for each item is of paramount importance as these recipes serve as input text prompts. These prompts are instrumental in generating images of the food items using a stable diffusion model, further elaborated in Section 3.1.2.

Additionally, we extended our dataset to include ingredients for food items adapted to different cuisine styles. By providing ChatGPT with specific prompts, such as listing the ingredients for ”a pizza in Korean style”, we received detailed ingredient lists reflective of cultural adaptations. For instance, the ingredient list for a Korean-style pizza generated by ChatGPT included ”Pizza dough, Korean chili paste, Kimchi, Pork belly slices, Green onions, Mozzarella Cheese, Sesame oil, Garlic, Soy sauce, and Sesame Seeds.” Through this approach, we amassed a collection of 400 unique food item and cuisine style combinations, each with its own tailored list of ingredients.

To ensure the diversity and authenticity of our dataset, we embarked on the final stage of compilation by identifying unique ingredients across 20 cuisine styles. This involved categorizing these ingredients into six categories: Seafood, Meat, Vegan, Protein, Vegetables, Fruits, and Condiments. Using ChatGPT, we pinpointed three unique ingredients for each cuisine style, which were then cross-checked against the Recipe Ingredient Dataset to verify their accuracy and relevance. Once validated, these unique ingredients were seamlessly integrated into our dataset, enriching the 400 distinct food item and cuisine style combinations. These comprehensive ingredient lists now serve as input prompts for generating images of cor-

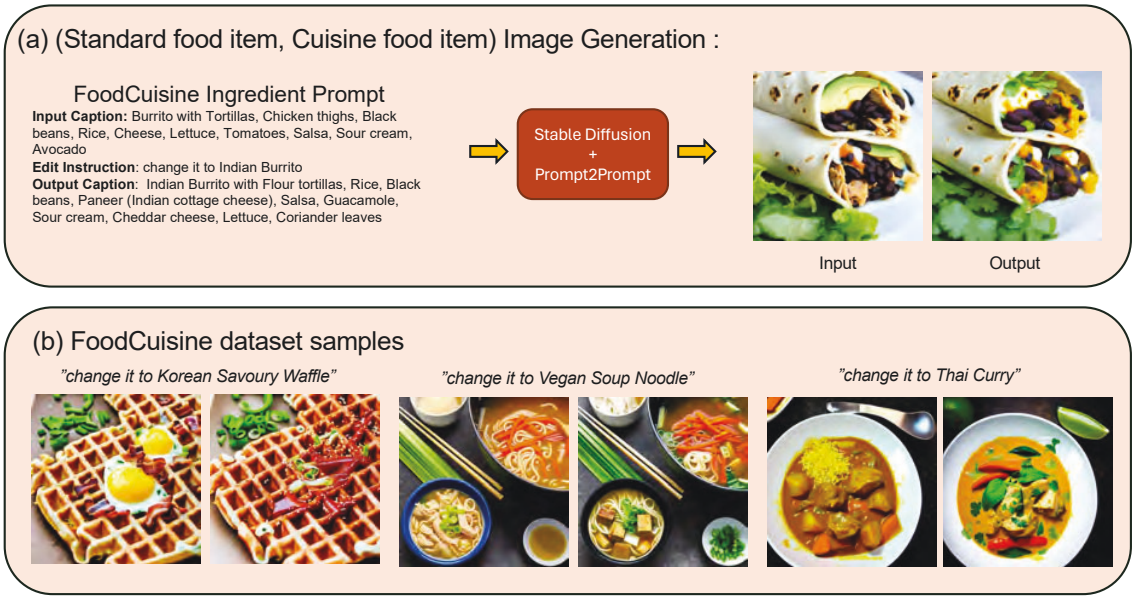

Figure 3. (a) Illustration of paired image generation of different cuisines. (b) FoodCuisine dataset examples

responding food items using a LDM. This approach not only increases the diversity of the dataset but also allows for a detailed exploration of the complex connection between culinary traditions and their cultural backgrounds.

In the augmentation of our dataset, subsequent to the acquisition of input and output caption pairs, we introduced an additional component termed as the "edit instruction." This component is crafted in the format "change it to CuisineStyle FoodItem," serving a pivotal role in the workflow of our Cuisine Transfer Model. The edit instruction is designed to function as an integral part of the model's editing prompt, which, when combined with the original image, facilitates the generation of the transformed output image.

3.1.2 Food Cuisine Paired Image Dataset

For image generation, we have adopted the image pair generation structure that is present in InstructPix2Pix and the overview of the Paired image generation of different cuisine is present in Fig. 3.

We used a text-to-image diffusion model [20] in our study to convert descriptive texts into corresponding images. However, ensuring consistency in generated images with slight text variations is challenging. To tackle this, we employed the Prompt-to-Prompt [9] technique, adjusting the process to enhance image similarity. This method utilizes an adjustable parameter, p, regulating image similarity by controlling the proportion of denoising steps sharing attention weights. We generated 50 sets of (input food image, cuisine transferred food image) pairs for each text description pair, refining them with a CLIP-based metric [7] to improve alignment between textual changes and visual modifications. This process enhances diversity and quality while strengthening data generation against method limitations.

In summary, we compiled the FoodCuisine Dataset, comprising 1,649 image pairs that were generated using the edit prompt "Change it to CuisineStyle FoodItem." Acknowledging the impact of edit prompt phrasing on the dataset, we foresaw the need to expand the range of prompts to ensure varied and precise results, which we explore in the next section.

3.1.3 Refining Edit Prompt

In our exploration of enhancing the Cuisine Transfer Model's effectiveness through linguistic variation, we initially adopted a standard edit prompt format: "Change it to CuisineStyle FoodItem." To investigate the impact of diverse linguistic expressions on the model's performance, we generated 20 different paraphrasings of this edit prompt using ChatGPT, aiming to introduce a wide array of linguistic inputs. The rationale was to determine if varying the phrasing of edit prompts could significantly influence the model's ability to accurately and creatively transform images according to specified cuisine styles.

However, upon evaluating the outcomes of these varied edit prompts, we observed minimal difference in the model's ability to execute the desired transformations effectively. This experiment suggested that the model's performance in interpreting and applying edit instructions is largely resilient to changes in the linguistic structure of the prompts. Consequently, this finding implies that efforts to optimize the model could be better focused on enhancing image quality

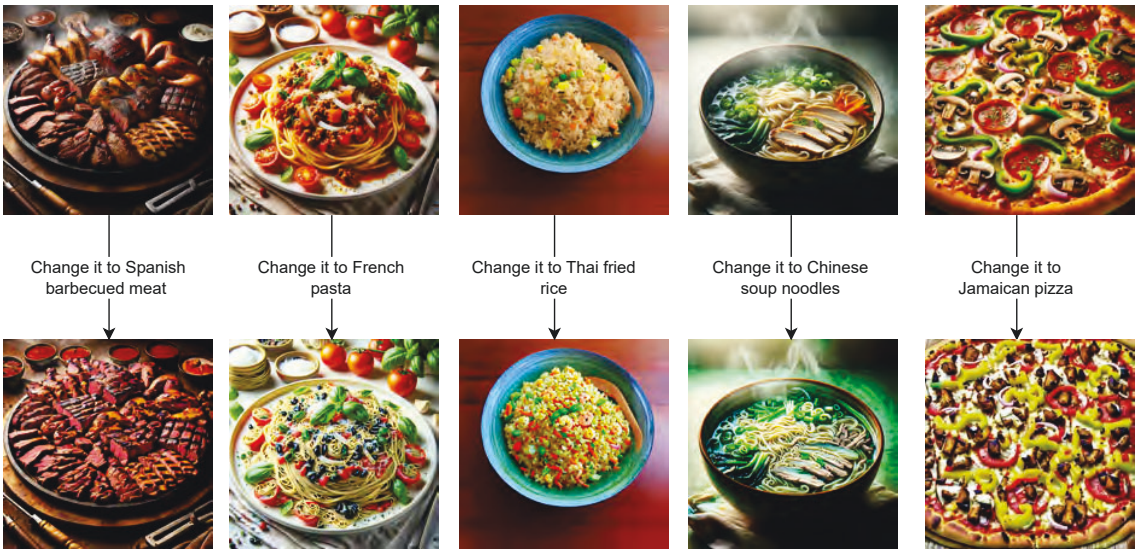

Figure 4. **Single Food dish** → **Single Cuisine**: Cuisine transferred images for different food dishes.

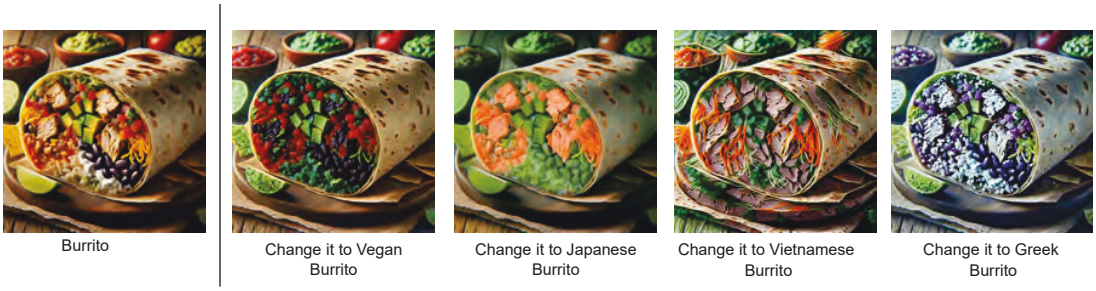

Figure 5. **Single Food dish** → **Multiple Cuisines**: Cuisine transferred burrito

and the accuracy of cuisine-specific details, rather than on expanding the diversity of prompt phrasing.

3.2. Cuisine Transfer Model Training

Building upon the methodologies outlined in [17], we opted to fine-tune the pre-trained InstructPix2Pix model using our Cuisine Transfer Pair Dataset rather than training from scratch with our data using a stable diffusion model. This fine-tuned model, referred to as the *CuisineTransfer* model, retains its general understanding while integrating specific task nuances. Fine-tuning enhances the model's adaptability, allowing it to evolve based on the characteristics of the new dataset. Additionally, we adopt the InstructPix2Pix approach of utilizing a classifier-free guidance mechanism [11] to condition on both the input image and edit instruction text. This results in two guidance scales: the Image Classifier-Free Guidance Scale (Image CFG) and the Text Classifier-Free Guidance Scale (Text CFG), which can be adjusted to balance the correspondence between the generated samples and the input image, as well as the correspondence with the edit instruction.

Implementation details: All our experiments are performed on a single NVIDIA A100 GPU with 80 GB of VRAM. While exploring different Text CFG and Image Image CFG scale values, we found 7.5 as text CFG, 1.5 as image CFG worked the best. We fixed 100 as the number of denoising steps of the diffusion model for all our experiments.

4. Results and Discussion

In this section, we analyze the performance of our proposed CuisineTransfer model both qualitatively and quantitatively. For qualitative metrics, we compare our CuisineTransfer model with SDEdit[16], Text2Live[1], DiffEdit[5] and InstructPix2Pix[3]. As for quantitative metrics, we have selected InstructPix2Pix[3] as our baseline model to gauge the impact of our domain specific dataset.

4.1. Qualitative Evaluation

We visualize the performance of our CuisineTransfer model using the input images generated by DALL·E 3 model [2] and images sourced online [18, 19]. We look at how our model performs when converting a

Figure 6. **Comparison with prior works**: Three food item changed to three different cuisine style (SDEdit[16], Text2Live[1], InstructPix2Pix[3], DiffEdit[5]) Text CFG: 7.5, image CFG: 1.5.

single food dish to a single cuisine, a single food dish to multiple cuisines, and finally compare our model's editing capabilities with prior works.

Single Food dish → Single Cuisine: We show the generated images of our Cuisine Transfer model for various food dishes in Fig. 4. We can see that the ingredients change align with the cuisine asked by the edit instruction. For instance, shrimp is prominent in Thai cuisine, which is added in Thai fried rice. Further, Jamaican pizza replaced mushrooms with jerk chicken, a common food ingredient in Jamaican cuisine.

Single Food dish → Multiple Cuisines: We convert an image of Burrito to Vegan, Japanese, Vietnamese and Greek cuisines, as shown in Fig. 5. We notice that our generated vegan burrito image removed the meat in burrito and substituted it with veggies and beans, Japanese burrito changed chicken to salmon, Vietnamese changed chicken to pork slices, and Greek cuisine has significant amount of feta cheese. All the changes are unique to the respective cuisines. However, there are some minor background abnormalities in the Vietnamese burrito, which we describe further in Sec. 4.2.2.

Comparisons with prior works: In Fig. 6, we compare the performance of cuisine transfer methods for three randomly chosen (dish, cuisine) pairs: fried noodles → Chinese fried noodles, crepes → Greek crepes, and burrito → vegan burrito. For the cuisine transfer, crepes → Greek crepes: SDEdit, Instructpix2pix and our model use the edit instruction, "change it to Greek crepes". However, Text2Live and DiffEdit requires both input and output captions of the images for the cuisine transfer task. Text2live uses "crepes" as input caption, and "greek crepes" as output caption. DiffEdit uses "a plate of crepes" as input caption, and "a plate of greek crepes" as output caption. We use similar instructions/captions for other food cuisine dish pairs. In the cuisine transfer of crepes, we observe that InstructPix2Pix and DiffEdit suffer from identifying background and the crepes. Specifically, the knife in the image is incorrectly converted to crepes. Another interesting observation is in the example of Vegan Burrito, where SDEdit and InstructPix2Pix return near identical image to the input image. We quantify this type of image conversion error in Sec. 4.2.2. Although, most of the prior works attempt to transfer cuisine of the dishes, they hardly match the ingredient prominent to that cuisine. For example, we can still see some meat in DiffEdit's vegan burrito, whereas in our CuisineTransfer model, we don't see any meat. Additionally, all prior work's Chinese noodles have no specific characteristic related to Chinese cuisine, but our model has realized a depic-

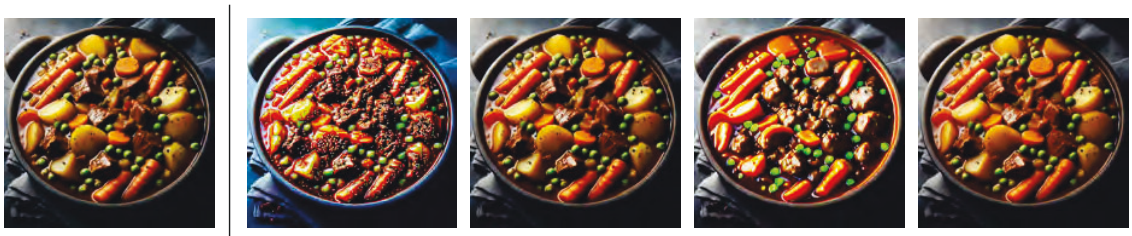

Figure 7. Comparison between Cuisine Transfer Model with InstructPix2Pix with different text CFG and image CFG value

Metric	InstructPix2Pix	CuisineTransfer (ours)
LPIPS (↓)	0.00252	**0.00232**
CLIP Image Similarity (↓)	0.9050	**0.7885**
CLIP Text-Image Direction Similarity (↑)	0.0761	**0.1320**

Table 3. Evaluation of baseline and our proposed method using image editing metrics

tion of shrimp chow mein, a notable dish in Chinese cuisine.

4.2. Quantitative evaluation

In this section, we quantitatively analyze the performance of our proposed CuisineTransfer model using existing image editing metrics and our proposed task-specific metrics. We compute quantitative metrics averaged on a test data of over 80 image, edit instruction pairs (approximately 5% of our training dataset) with 4 randomly chosen dishes: burritos, fried rice, fried noodle, and pasta.

In our quantitative evaluation, we conducted a comparative analysis between the InstructPix2Pix model and our CuisineTransferModel. Given that our model is an extension of InstructPix2Pix, comparing these two models is particularly pertinent. To ensure a fair and consistent comparison, we standardized the configuration settings for both the Image CFG and the Text CFG at fixed values of 7.5 and 1.5, respectively. These parameters were determined after conducting qualitative assessments of images processed by InstructPix2Pix, where it was observed that images did not exhibit significant change, as illustrated in Fig. 7. This setup allowed us to rigorously evaluate the performance enhancements introduced by our CuisineTransferModel over the baseline InstructPix2Pix framework.

4.2.1 Image Editing Metrics

We follow the steps of Wang *et al*. [24] for evaluation and use LPIPS [26] and CLIP similarity metrics [7, 10] for the task of Cuisine Transfer. LPIPS (lower is better) and CLIP Image similarity (lower is better) metric measure the similarity between the input image and generated image, while CLIP Text-Image Direction Similarity (higher is better), measures the agreement between the change in images and text captions.

The results in Tab. 3 show that our proposed model performs better than the baseline model, instructpix2pix. Our model improves CLIP Text-Image direction similarity scores significantly by 73% compared to the baseline, indicating our generated images align well with the input edit instruction. However, it's worth highlighting that the changes in the image similarity metrics are not conclusive, as the values are not significantly different, with a 12% change for CLIP image similarity and a 7% change for LPIPS. Therefore, we explore task-specific metrics in the next section.

4.2.2 Task Specific Metrics

We start by analyzing in how many generated images the ingredient changed compared to the original image. We call the ratio of ingredient changes to the total number of samples as Ingredient Change Ratio. On further analysis, we notice some common failure patterns as illustrated in Fig. 8. We propose three metrics to quantify them, namely, Morphed Images Ratio, Colour Shift Images Ratio and Cuisine-Ingredient Agreement Ratio for the task of Cuisine Transfer.

Morphed Images Ratio: We define morphed food images as those that have structural abnormalities in food items. Examples of these abnormalities include incorrect shapes of food items, food items morphed together in ways that don't make sense, like a fruit that partially looks like a vegetable, plates and background content wholly or partially changing into cuisine ingredients, etc. We define Morphed Images Ratio as the ratio of the number of generated images that are morphed to that of the total number of generated im-

Metric	InstructPix2Pix	CuisineTransfer (ours)
Ingredient Change Ratio (↑)	58/80	**78/80**
Color Shifted Images Ratio (↓)	19/22	**0/2**
Cuisine-Ingredient Agreement Ratio (↑)	24/58	**67/78**
Morphed Images Ratio (↓)	41/80	**11/80**

Table 4. Evaluation of proposed Cuisine Transfer specific metric

Figure 8. Different Failure cases. Morphed images: plate changed to burrito; Color shift: Fried rice has green tint; Cuisine ingredient agreement failure: no Filipino specific protein substitution.

ages in the test data.

Color Shifted Images Ratio: We define color shift as those images that have no visible changes in food ingredients but have changes only in the color of the image. Thus, Color Shifted Images Ratio is the ratio of number of generated images that are color shifted and have no ingredient change to that of the number of generated images in the test data which have no ingredient change.

Cuisine-Ingredient Agreement Ratio: We say a generated image has the cuisine-ingredient agreement property, when the ingredients in the images match with the given cuisine. We define, Cuisine-Ingredient Agreement Ratio as the ratio of generated images that have ingredients change and the cuisine-ingredient agreement property to that of the total number of images that have ingredient change in the test data.

We compute the value of the above metrics manually by looking at each image in the test data. We showcase the results of our proposed metrics for the baseline model and our proposed CuisineTransfer model in Tab. 4. We observe that our proposed model changes ingredient in almost 97% of samples (measured by Ingredient Change Ratio), while the baseline model, InstructPix2Pix changes only in 72%. Further, our model matches the ingredient changes with the target cuisine for almost 85% of ingredient changed images, compared to 41% in insturctpix2pix. The failure case of color shift is non-existent for our method in the test data, while InstructPix2Pix exhibit color shifted generated images. Finally, the number of generated images that are morphed is much lower for our model compared to the baseline model.

5. Discussion and Limitations

In our exploration, we conducted a zero-shot evaluation of the Cuisine Transfer Model on food items not encountered during training, specifically tacos and poke bowls. This exercise revealed the model's capacity to modify ingredients in these novel dishes, albeit with a caveat: achieving precise and reliable edits necessitates a more comprehensive dataset for each specific food item. Despite this, we are optimistic about the model's potential to revolutionize culinary creativity and to significantly reduce marketing expenses for the food industry, thereby exerting a profound influence on the culinary community.

An area identified for improvement involves conducting more rigorous and comprehensive human subjective studies while computing the task specific metrics for robust evaluation. Such research could offer invaluable insights into enhancing the model's accuracy and user satisfaction, a task we aim to address in our future work.

6. Conclusion

In this work, we presented the CuisineTransfer Model, a novel framework based on diffusion model principles specifically tailored for altering the composition of food items in accordance with given culinary style queries. Given a style change as query, using the diffusion model, Cuisine Transfer method can edit the ingredients of the food item. We have thoroughly designed to create Food Cuisine Ingredient prompts that could capture ingredient of cuisine style, leverage these prompts to generate and validate the training image pair, and trained diffusion model. Our model underwent quantitative evaluation through both image editing metrics and the three task-specific cuisine transfer metrics we introduced, in addition to a qualitative assessment. These evaluations showed the superiority of our model compared to other existing prior work on cuisine transfer task.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 1, 2, 5, 6

[2] Jason Betker, Greg Goh, Li Jing, Tim Brooks, Jianyu Wang, Liang Li, Lucy Ouyang, Jie Zhuang, Jason Lee, Yuxuan Guo, Waseem Manassra, Prafulla Dhariwal, Chenxi Chu, and Yong Jiao. Improving image generation with better captions. *OpenAI Blog*, 2023. 5

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2, 5, 6

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023. 1, 2, 5, 6

[6] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 4, 7

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020. 1

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 7

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[14] Kaggle. Recipe ingredients dataset. https://www.kaggle.com/datasets/kaggle/recipe-ingredients-dataset, 2023. [Online; accessed 11-March-2024]. 2, 3

[15] Olivia Markham, Yuhao Chen, Chi en Amy Tai, and Alexander Wong. Foodfusion: A latent diffusion model for realistic food image generation, 2023. 2

[16] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2, 5, 6

[17] Sayak Paul. Instruction-tuning stable diffusion with instructpix2pix. *Hugging Face Blog*, 2023. https://huggingface.co/blog/instruction-tuning-sd. 5

[18] Pixabay Contributor. Pancakes crepes filled pancake. https://pixabay.com/photos/pancakes-crepes-filled-pancake-577386/, 2024. Accessed: 2024-03-16. 5

[19] Pixabay Contributor. Noodles fried noodles plate. https://pixabay.com/photos/noodles-fried-noodles-plate-5617583/, 2024. Accessed: 2024-03-16. 5

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[21] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[22] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 1

[23] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11572–11581, 2019. 2

[24] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv preprint arXiv:2305.18047*, 2023. 7

[25] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2