DeepTasteDM: A Real-Time Food Appearance Transformation AR System Using Diffusion Models

Karina Karasawa[1], Haruto Fujita[1], Kizashi Nakano[2], and Keiji Yanai[1]

[1] The University of Electro-Communications, Tokyo, Japan
karasawa-k@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp
[2] The University of Tokyo, Tokyo, Japan
kizashi@cyber.t.u-tokyo.ac.jp

Abstract. Cross-modal perception studies have shown that changing the visual appearance of food can change its perceived taste. DeepTaste, a prior augmented reality (AR) system, realized such gustatory manipulation by translating food images in real time with a GAN, but its translation flexibility and image quality were limited. Diffusion models can generate more realistic and diverse images, but their high computational cost makes them difficult to use in latency-sensitive AR applications. In this paper, we propose DeepTasteDM, a real-time food appearance transformation AR system based on diffusion models. DeepTasteDM builds on StreamV2V, a streaming video-to-video translation method, and transforms only the food region, which is segmented with YOLO26 and tracked with DeAOT, while LoRA adapts the diffusion model to the food domain. To reduce the computational cost, the system regenerates the food image only when the food mask changes substantially between frames and otherwise reuses the previously generated image with an affine transformation, achieving about 14 fps with a latency of about 130 ms. In the evaluation experiments, DeepTasteDM achieved a higher CLIP score and a lower FID than DeepTaste. In addition, reusing generated images with an affine transformation reduced flickering compared with StreamDiffusion, which transforms each frame independently.

Keywords: Augmented reality · Food image transformation · Diffusion models · Cross-modal perception

1 Introduction

Reproducing and presenting the taste of a specific dish, such as ramen, is difficult. Unlike visual information, for which presentation methods are well established, the taste of such a dish cannot be reproduced by simply combining basic tastes. On the other hand, taste perception is not determined by gustatory stimuli alone; it is strongly influenced by other senses such as vision and smell, and the same food can be perceived as tasting differently. Research on cross-modal perception has exploited this property to manipulate perceived taste by changing visual



Fig. 1: Overview of DeepTasteDM. A food sample of fried rice (top left, input image) is transformed into curry and rice (bottom left, output image), which is overlaid on the real food and presented to the user wearing an AR head-mounted display

information. Augmented reality (AR) is a promising means for this purpose, because it can overlay virtual content on real food and dynamically change its appearance while a person is eating.

DeepTaste [18] realized such gustatory manipulation by combining StarGAN [2] with AR, translating the appearance of food into that of another dish in real time. However, GAN-based translation is restricted to a fixed set of trained food categories, and its output is often a simple coloring of the input that lacks realistic texture. Replacing StarGAN with rapidly advancing diffusion models is expected to enable flexible, prompt-driven, and high-quality food transformation, which would broaden the applicability of research on gustatory manipulation. On the other hand, diffusion models are computationally expensive, and the resulting latency is a major obstacle to applying them to AR.

In this paper, we propose DeepTasteDM, a real-time food appearance transformation AR system that uses a diffusion model (Fig. 1). We adopt StreamV2V [12], a fast streaming video-to-video translation method, and transform only the food region instead of the entire camera image. To present food that looks closer to the real thing, we use a diffusion model capable of high-quality image generation, and to approach real-time transformation, we build on StreamV2V. Our purpose is to increase the immersion of the AR eating experience and thereby enhance the effect of gustatory manipulation through visual presentation. The main contributions of this paper are as follows:

- We build an AR system that transforms the appearance of real food with a diffusion model in real time, using text prompts to specify the target dish and LoRA to adapt the model to the food domain.
- We combine food-region segmentation and tracking with a frame-skipping strategy based on mask IoU and affine transformation, which reduces the number of diffusion inferences.
- Through an evaluation on 18 videos and eight target dishes, we show that the proposed system produces results that are closer to the target dishes and more realistic than those of the GAN-based DeepTaste, and we analyze the effect of reusing generated images.

2 Related Work

2.1 Food Appearance Transformation with AR

AR overlays virtual elements on the real world, allowing users to interact with computer-generated information and images through devices such as smartphones and AR glasses.

SOLLAMA [1] captures the label of a wine bottle and uses a Multimodal Large Language Model (MLLM) to display information retrieved from a matched wine database. ARChef [31], which supports cooking by recognizing ingredients and generating recipes with Gemini Flash 1.5, displays food labels, recipes, nutritional information such as calories and fat, and allergen information from an input food image. It can also serve as a voice-assistant chatbot, allowing users to check whether they are following the recipe while cooking. CalorieCaptor-Glass [20] estimates the actual size of food from the camera projection matrix to calculate and display its calorie content, and TrackSugAR [23] visualizes the amount of sugar contained in food to help users reduce their sugar intake.

Beyond displaying information, this technology has also been used to develop gustatory displays that change the appearance of actual food into that of different dishes. By changing the appearance of food, FoodAR can make low-calorie or low-sodium meals look like a user's favorite dishes, thereby keeping satisfaction, and can also diversify eating experiences using the same ingredients. Because these studies use cross-modal perception, in which the perceived taste changes under the influence of vision and other senses, larger gustatory manipulation effects are expected as image processing and generation techniques advance.

NetPot [3] uses a webcam and Skype to display three virtual food items as if they were dropped into one half of a hot pot, together with three avatars around it, enabling a more enjoyable remote hot-pot dining experience with people in distant locations. MetaCookie+ [21] is a pseudo-gustatory display that visually transforms a plain cookie printed with an AR marker into a chocolate or tea cookie and presents the corresponding smell with an olfactory display. CalibraTable [27] is an interactive tabletop system in which infrared LEDs are attached to a transparent plate. A webcam under the table recognizes their blinking patterns to locate the plate and the food, and virtual plates are projected around the food. The projective AR system [22] transforms single-color mask images

and projects them onto food, changing the appearance of food such as castella and potato chips, as well as that of plates, in real time. FoodChangeLens [19] uses a HoloLens and Conditional CycleGAN [35] to build an AR system that transforms food into 10 categories while taking the actual food shape. AR Food Changer [30] realizes accurate texture transformation by combining object extraction with Mask R-CNN and food contour extraction from binarized images. DeepTaste [18], on which our research is based, is a gustatory manipulation system that dynamically changes the appearance of food using StarGAN and AR. Its StarGAN model is trained on UECFOOD-100 [17] and additional images collected from X, and supports 10 food translations, such as from somen noodles to yakisoba and from white rice to curry. In this paper, we improve the gustatory manipulation effect of DeepTaste by realizing high-quality and flexible, prompt-driven food transformation with StreamV2V [12].

2.2 Food Image Generation

Image generation models such as Stable Diffusion [25] are trained on large-scale datasets and can generate diverse images. Although their training data include food images, these are only a small part of it, and the models are not specialized for the food. Therefore, many attempts have been made to generate higher-quality and more appealing food images with diffusion models. FoodFusion [16] achieves more realistic food image generation by training on LAION-Food, a dataset built by selecting high-quality food images from LAION-5B. CW-Food [33] combines private features extracted from images with intra-class features obtained from text and further uses LoRA, enabling realistic and diverse food image generation. As these studies show, generating high-quality images specialized for the food domain commonly involves curating training data or using LoRA [6], which trains only a small number of additional parameters. LoRA freezes the pretrained weights and trains only low-rank matrices, which enable efficient adaptation to a target domain while greatly reducing computational resources and memory consumption.

2.3 Fast Diffusion for Real-Time Generation

Diffusion models are superior to earlier generative models in that they can generate high-quality images, but their generation speed is slow because of the iterative denoising in the reverse diffusion process. To accelerate generation, many methods have been studied, such as Latent Consistency Models [14] and LD-DGAN [29] for images, and TurboDiffusion [34] and StreamDiffusion [10] for video. Building on these studies, we use StreamV2V [12], which enables real-time video-to-video translation guided by user prompts. StreamDiffusion minimizes waiting time with StreamBatch and enables fast processing. Building on it, StreamV2V processes video using a streaming method and supports an unlimited number of frames. It introduces an extended self-attention mechanism that refers to the keys and values of past frames and a feature bank that retains

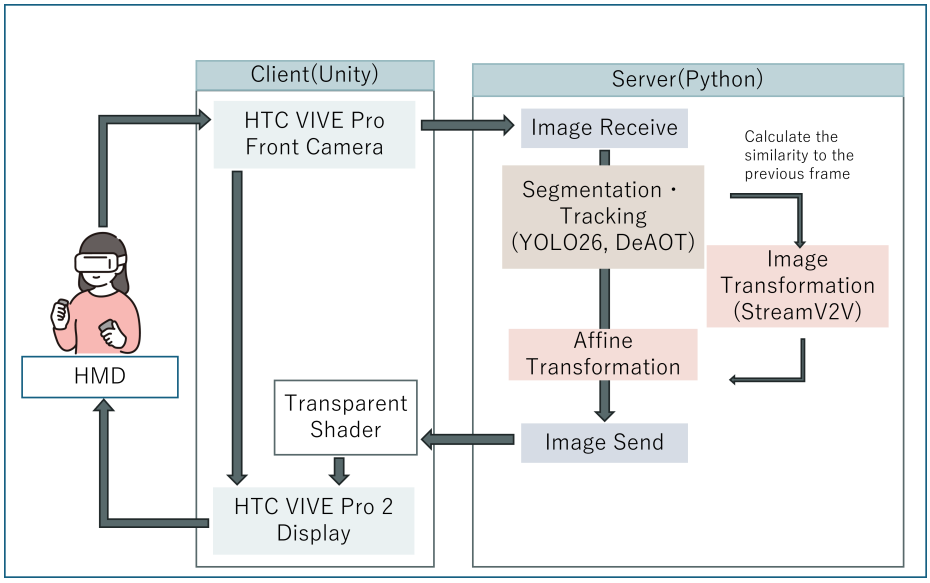


Fig. 2: Overview of the proposed system. The Unity client captures images from the front camera of the HTC VIVE Pro 2 and sends them to the Python server, which performs food-region segmentation and tracking (YOLO26, DeAOT), image transformation (StreamV2V), and affine transformation. The generated image is sent back and overlaid on the camera view with a transparent shader

information from past frames. As a result, it maintains temporal consistency while keeping high processing speed.

3 Proposed System

We propose a food appearance transformation AR system that runs in real time while maintaining the high expressive power of diffusion models. An overview of the system is shown in Fig. 2. The system is based on StreamV2V, a fast diffusion-based translation method. Instead of generating an image for every frame, we construct a pipeline that switches its processing according to the amount of motion between camera frames. That is, when the change in the user's viewpoint or in the food region is small, the time-consuming inference of the diffusion model is skipped and the transformation result of the previous frame is quickly warped by an affine transformation. To achieve both recognition and fast tracking of the food region, we combine segmentation by YOLO26 [7] with real-time tracking by DeAOT [32].

3.1 System Architecture

The system consists of a client implemented in Unity 2018.2.9f1 and a Flask server implemented in Python 3.10.16. First, the client crops a 512×512 region



(a) Input (b) Segmentation (c) Transformation (d) Affine transform

Fig. 3: Example of segmentation and transformation into curry

around the center of the front camera image of an HTC VIVE Pro 2 and sends it to the server through inter-process communication via shared memory, together with the text prompt and metadata. The server performs food-region segmentation, image transformation, and affine transformation for latency reduction, and returns the generated image to the client. The client overlays the received image on the front camera video with a transparent shader and presents it on the display of the head-mounted display.

3.2 Food Region Segmentation and Tracking

For natural food transformation, the background and the dishes should be kept as they are in the input image. Therefore, our system extracts the food region with YOLO26, and tracks and reuses the food-region mask with DeAOT, a video object segmentation model. Retaining the food-region mask in this way enables smooth image transformation. In our implementation, we replace everything outside the food region with a white background based on the mask, use this image as the input to StreamV2V, and generate a transformed image of the food only. Finally, only the food region of the transformed image is overlaid on the input image from the front camera. In this way, only the food is transformed while the background and the dish are left unchanged, enabling highly immersive AR presentation. Figure 3 shows an example of this process.

Excluding the User's Hands. The food segmentation model sometimes misrecognizes the user's hand or forearm reaching toward the dish as part of the food region (Fig. 4). In such cases, the transformed image is also overlaid on the hand, and because the affine transformation follows the shape change of the mask that includes the hand, the transformation result is unnaturally stretched. To solve this problem, our system estimates a person mask with a YOLO26 segmentation model trained on COCO [13] and subtracts it from the food mask, both immediately after food detection and on the final mask. Small holes inside the person mask are filled as part of the body. This prevents the hand and forearm from being transformed, so that only the food on the dish is transformed.



(a) Input (b) Without exclusion (c) With exclusion

Fig. 4: Effect of excluding the user's hand with the person mask (transformation from fried rice into curry). (a) Input image. (b) Without exclusion, the hand is misrecognized as part of the food region, and the transformed image is also overlaid on the hand. (c) With exclusion, only the food on the dish is transformed.



(a) Without LoRA (b) With LoRA

Fig. 5: Comparison of transformation into *hiyashi chuka* with and without LoRA

3.3 Domain Adaptation with LoRA

Because the base diffusion model is not specialized for food images, adapting it to the food domain is beneficial regardless of the type of dish. In particular training data mainly consist of Western dishes. Therefore the generated images tend to deviate more from the actual appearance when Japanese or Chinese dishes are targeted. In addition, the generation quality becomes unstable against blur and noise in the input images. Therefore we use LoRA to achieve stable food image transformation under diverse input conditions. As shown in Fig. 5, LoRA makes it possible to generate more natural food images. We used kohya-ss/sd-scripts[3] for training, and used 512×512 images from UECFood-256 [8] with non-food regions replaced by a white background as training data.

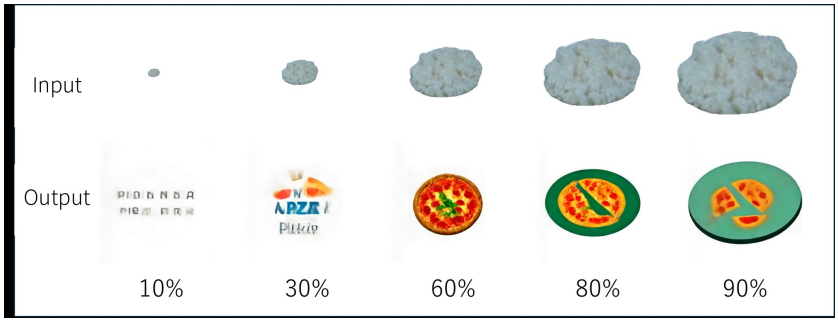


Fig. 6: Input images (top) and transformed images (bottom) for different proportions of the food region in the input image (10%, 30%, 60%, 80%, and 90%)

3.4 Size of the Food Region

We found that the proportion of the food region in the input image strongly affects the generation quality, as shown in Fig. 6. When the front camera moves away from the food, the food region becomes small relative to the input image of StreamV2V; in this case, the white background dominates the input and realistic images are not generated. When the food region is large, the generated images tend to be coarse. Based on this observation, our system fixes the proportion of the food region in the input image to 60%.

3.5 Skipping Image Generation with Affine Transformation

Because the inference cost of diffusion models is high, running inference on every frame causes large latency. On the other hand, the head motion of a user during a meal is relatively slow, and there are many scenes with little viewpoint change. In such cases, it is preferable to display the generation result of the previous frame and make it follow the food. Therefore we switch the processing based on the similarity of the food region between frames. Specifically, we compute the intersection over union (IoU) between the food-region masks of the previous and current frames. If the IoU falls below a threshold of 0.8, indicating a large change from the previous frame, a new image is generated with the diffusion model. Otherwise, when the change is small, the stored generated image is reused after being aligned to the current food region by a three-point affine transformation estimated from the contours of the food regions in the stored transformed image and the input image; when there are multiple food regions, they are matched by IoU and centroid. This strategy effectively skips the heavy generation process and ensures real-time performance.

[3] https://github.com/kohya-ss/sd-scripts

3.6 Parallel Thread Processing

If the sequence of image reception, segmentation, image transformation, affine transformation, and image sending is executed in a single thread, no other processing can be performed during a heavy image transformation, and the video on the head-mounted display (HMD) freezes. Even while the time-consuming image generation runs in the background, the client needs to keep following the food with the affine transformation to maintain smooth rendering. Therefore we build a parallel structure with four threads: an image reception thread, an image generation thread, an affine transformation thread, and an image sending thread. This reduces the waiting time between processes and lets the heavy generation task and the lightweight tracking task run in parallel, which reduces the rendering latency and realizes smooth video output.

With an HTC VIVE Pro 2, an Intel Core i7-9700K CPU, and an NVIDIA RTX 6000 Ada GPU, the system successfully overlays the generated images at about 14 fps with a latency of about 130 ms, as measured while running the system live on the head-mounted display.

4 Experiments

We conducted qualitative experiments and a quantitative evaluation that compares the proposed system with DeepTaste on the same input videos and with the same metrics, to verify the visual quality and temporal consistency of food transformation.

4.1 Experimental Setup

Data. We used 18 videos: 14 noodle videos (10 stock videos from Pexels[4] and 4 videos of somen noodles that we recorded) and 4 rice videos (2 recorded by us and 2 from Pexels). Videos in which the evaluation region was valid in fewer than 50% of the frames were excluded automatically, independently of the results of the methods. Each video was used in its entirety at its original frame rate, cropped to a square centered on the food region, and resized to 512×512; all methods received exactly the same preprocessed frames.

Target Dishes. The noodle videos were transformed into four noodle dishes (*yakisoba*, ramen, soba, and *hiyashi chuka*), and the rice videos into four rice dishes (curry, *chahan* fried rice, white rice, and *unadon*). These are the domains of DeepTaste's StarGAN for which our system has a corresponding LoRA. We used short prompts (*e.g.*, "yakisoba, japanese fried noodles, plate"), which also served as the target text for the CLIP score, and a dish-specific LoRA selected automatically from keywords in the prompt with a weight of 0.8. Note that some Pexels noodle videos contain ramen and the rice videos show white rice, so the source and target dishes are similar for the transformations into ramen and white rice, and their scores partly reflect the appearance of the input itself.

[4] https://www.pexels.com/

Compared Methods. We compared two methods: (1) *DeepTasteDM*, the proposed method with the same logic as the deployed server; and (2) *DeepTaste* [18], which, as in its deployed version, downsamples the input to 256×256, translates the whole frame with StarGAN, and upsamples the result to 512×512 with bicubic interpolation. For the proposed method, we used StreamV2V with "Jiali/stable-diffusion-1.5", LCM-LoRA [15], and a tiny autoencoder (TAESD) [9], with denoising steps $t \in \{5, 20\}$ and a guidance scale of 1.0. Food regions were extracted with YOLO26n-seg fine-tuned to a single "food" class, person masks with YOLO26n-seg trained on COCO, and the experiments were run on an NVIDIA RTX 6000 Ada GPU.

Metrics. All metrics were computed only within the evaluation region, defined as the food mask of the input frame minus the person mask, which is shared by all methods; frames in which this region contained fewer than 1,000 pixels were excluded. Image-level metrics were computed every third frame on crops in which pixels outside the evaluation region were set to white. We used the CLIP score [4] between the image and the target prompt with CLIP ViT-B/32 [24]; the CLIP directional similarity, the cosine similarity between the change in the image embedding from input to output and the change in the text embedding from the source ("noodles" or "rice") to the target text; FID [5] against real reference images of each target dish, computed with Inception-v3 features; and the warping error [11], the mean squared error within the evaluation region between output frame $t+1$ and output frame t warped by optical flow estimated with RAFT [28] on the input video, excluding occluded pixels by a forward–backward consistency check [26], as a measure of temporal consistency.

4.2 Qualitative Results

Figure 7 compares the transformation results of the conventional DeepTaste (StarGAN) and the proposed system for four dishes: *unadon* (grilled eel on rice), *hiyashi chuka* (cold ramen), ramen, and *yakisoba* (fried noodles). For these qualitative examples, we used the following more detailed prompts:

(a) "unadon, grilled eel with sweet soy glaze, lacquered bowl, sansho pepper, char-grilled marks"
(b) "hiyachu, hiyashi chuka, colorful toppings, shredded ham, cucumber, egg, cold noodles, ice cubes"
(c) "ramen, rich tonkotsu broth, soft boiled egg, nori, chashu pork, oil droplets on surface"
(d) "yakisoba, brown sauce glaze, red pickled ginger, aonori seaweed, steaming hot"

StarGAN transforms the entire input image, including the background such as the table and the dishes. In the results for *unadon* and *yakisoba*, the rice and noodles tend to be only simply colored. In contrast, the proposed system preserves the appearance of the table and the dishes, producing more realistic and



(a) Unadon (b) Hiyashi chuka (c) Ramen (d) Yakisoba

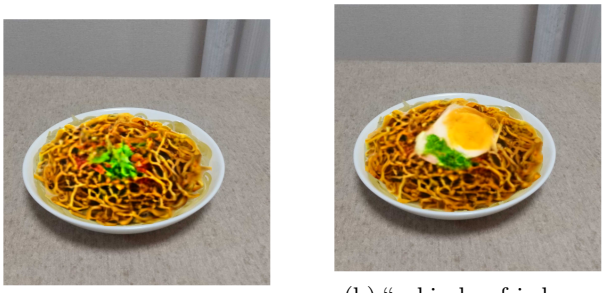
Fig. 7: Comparison of transformation results between DeepTaste [18] (top) and the proposed system (bottom)

immersive results. Moreover, ingredients appear, such as the eel of the *unadon* and the nori of the ramen, and the texture of the noodles is expressed. The transformation results are thus of high quality with clear details. These results confirm the effectiveness of the diffusion model and LoRA.

In addition, the proposed system can flexibly control the transformation results with text prompts. As shown in Fig. 8, adding "fried egg on top, runny yolk" to the prompt for yakisoba produces yakisoba topped with a fried egg. Unlike GAN-based methods, which can only transform food between fixed categories, the proposed system allows users to add details such as toppings according to their preferences.

4.3 Qualitative Temporal Consistency

Figure 9 compares the results with and without considering temporal consistency. We compare StreamDiffusion, which we adopted in an earlier version of the system, with the proposed system combining StreamV2V and affine transformation. With StreamDiffusion, which transforms each frame independently (Fig. 9(a)), the color of the curry roux changes drastically from frame to frame with slight camera motion, which is unsuitable for food transformation in AR. In contrast, the proposed method (Fig. 9(b)) keeps the generation result of the previous frame and makes it follow the food with an affine transformation when the change is small, yielding natural and temporally consistent results with little flickering. Reusing generated images with an affine transformation thus suppresses the flickering that arises in frame-by-frame diffusion-based translation with StreamDiffusion and greatly improves the visual stability of food transformation in AR.



(a) "yakisoba, brown sauce glaze, red pickled ginger, aonori seaweed, steaming hot"

(b) "yakisoba, fried egg on top, runny yolk, brown sauce glaze, red pickled ginger, aonori seaweed, steaming hot"

Fig. 8: Comparison of topping addition via text prompts

Table 1: Quantitative results averaged over the four noodle dishes and the four rice dishes. CLIP, CLIP-dir, and WE (warping error, $\times 10^{-3}$) are averaged over (video, dish) pairs; FID is the mean of the per-dish values. Bold: better of the two methods.

Group	Method	CLIP ↑	CLIP-dir ↑	FID ↓	WE ↓
Noodle (14 videos)	DeepTaste	0.511	0.006	228.4	7.28
	DeepTasteDM (ours)	**0.685**	**0.010**	**133.1**	16.65
Rice (4 videos)	DeepTaste	0.561	0.050	243.2	0.79
	DeepTasteDM (ours)	**0.642**	**0.055**	**165.0**	19.08

4.4 Quantitative Results

Table 1 summarizes the results for each group, and Tab. 2 shows the CLIP score and FID for each dish.

Comparison with DeepTaste. The proposed method achieved a higher CLIP score than DeepTaste for all eight dishes (differences of +0.14 to +0.21 for the noodle dishes and +0.07 to +0.12 for the rice dishes). The FID was also lower than that of DeepTaste for every dish. These results are consistent with the qualitative observation in Fig. 7 that DeepTaste mainly changes the color tone of the input, whereas the proposed method generates a realistic appearance of the target dish. Note that DeepTaste had a lower warping error than the proposed method. Because DeepTaste is GAN-based and applies almost the same color transformation to the whole frame, its results hardly differ between frames; we therefore consider that this metric does not meaningfully reflect differences in temporal consistency between DeepTaste and our system.



(a) Frame-by-frame transformation



(b) Transformation considering temporal consistency (ours)

Fig. 9: Comparison of transformation results with and without considering temporal consistency

Effect of Reusing Generated Images. In the proposed method, the diffusion model regenerated the image only in frames where the food mask changed substantially, and the generated image was reused with an affine transformation in the remaining frames. Between regenerations, the warping error of consecutive reused frames was as small as 8.1–14.6, showing that keeping the previous result and making it follow the food with an affine transformation yields a stable display. In contrast, it was 66.4–125.2 for pairs in which the next frame was regenerated; the discontinuity at the moment the reference image is replaced is the main source of the overall warping error of the proposed method, and smoothing this transition is left for future work.

Table 2: Per-dish CLIP score and FID. †: dishes similar to the source food. "Real" is the FID between two disjoint halves of the real reference images (a lower bound). Bold: better of the two methods

	Noodle (14 videos)				Rice (4 videos)			
Method	Yakisoba	Ramen†	Soba	Hiyashi	Curry	Chahan	Rice†	Unadon
CLIP score ↑								
DeepTaste	0.478	0.534	0.467	0.564	0.550	0.537	0.584	0.572
DeepTasteDM (ours)	**0.691**	**0.682**	**0.661**	**0.707**	**0.615**	**0.654**	**0.654**	**0.647**
FID ↓								
DeepTaste	259.3	176.3	253.9	224.2	214.2	342.5	192.5	223.5
DeepTasteDM (ours)	**162.9**	**100.9**	**126.9**	**141.6**	**145.3**	**170.1**	**153.9**	**190.7**
Real	11.2	7.4	47.2	16.5	17.3	18.0	38.2	34.9

4.5 Limitations

The exclusion by the person mask still has weaknesses. In third-person videos, the whole torso is detected as a person, so food lifted in front of the torso is also removed from the food mask. In the intended first-person view through the head-mounted display, only the user's own hands and forearms are visible, so this effect is small; however, in extreme close-ups where only fingertips are visible, the hand is not detected as a person and is not excluded. Our evaluation also has limitations. The rice group contains only four videos, and the Pexels videos are not captured from the viewpoint of an AR user. The evaluation region is defined with the same food and person segmentation models used in our system, and frames and videos in which it cannot be defined are excluded. The CLIP score and FID are semantic and statistical proxies, not human judgments of naturalness or of the effect on taste; in particular, FID is more sensitive to photorealism than to dish identity. The real reference images were taken from the data pool used to collect LoRA training data, and the training images could be excluded only for *yakisoba*; moreover, DeepTaste's StarGAN and our LoRAs were trained on different data with different methods. Finally, the quantitative evaluation was performed offline on prepared videos and does not evaluate the experience on an actual head-mounted display.

5 Conclusion

In this paper, we proposed DeepTasteDM, a real-time food transformation AR system using a diffusion model. By introducing a diffusion model and LoRA, the system enables high-quality image generation and flexible, text-prompt-driven food transformation, which were difficult with conventional GANs. In an evaluation on 18 videos and eight target dishes, the proposed system outperformed DeepTaste in both the CLIP score and FID. We also showed that reusing generated images with an affine transformation suppresses the flickering observed with StreamDiffusion, and that subtracting the person mask prevents the user's hands from being transformed. Challenges remain in the excessive removal of food by the person mask, in generation quality, and in the discontinuity at regeneration. In future work, we will introduce a segmentation model trained to distinguish hands and chopsticks from food, perform additional LoRA training to improve the quality of the generated images, and smooth the transition when the reference image is regenerated. We also plan to conduct user studies to evaluate the effect of the proposed system on taste perception.

Acknowledgements This work was supported by JST CRONOS JPMJCS24K4 and JSPS KAKENHI 26K23875.

References

1. Armandi, V., Loretti, A., Stacchio, L., Cascarano, P., Marfia, G.: Multi-modal large language model driven augmented reality situated visualization: the case of

wine recognition. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 1313–1322 (February 2025) 3
2. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8789–8797 (2018) 2
3. Foley-Fisher, Z., Tsao, V., Wang, J., Fels, S.: Netpot: easy meal enjoyment for distant diners. In: International Conference on Entertainment Computing. pp. 446–448. Springer (2010) 3
4. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7514–7528 (2021) 10
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017) 10
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. International Conference on Learning Representations (ICLR) (2022) 4
7. Jocher, G., Qiu, J., et al.: Ultralytics YOLO26: Unified real-time end-to-end vision models. arXiv preprint arXiv:2606.03748 (2026) 5
8. Kawano, Y., Yanai, K.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV) (2014) 7
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 10
10. Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., Keutzer, K.: StreamDiffusion: A pipeline-level solution for real-time interactive generation. arXiv preprint arXiv:2312.12491 (2023) 4
11. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proc. European Conference on Computer Vision (ECCV). pp. 170–185 (2018) 10
12. Liang, F., Kodaira, A., Xu, C., Tomizuka, M., Keutzer, K., Marculescu, D.: Looking backward: Streaming video-to-video translation with feature banks. In: Proc. International Conference on Learning Representations (ICLR) (2025) 2, 4
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proc. European Conference on Computer Vision (ECCV). pp. 740–755 (2014) 6
14. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023) 4
15. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: LCM-LoRA: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556 (2023) 10
16. Markham, O., Chen, Y., Tai, C.e.A., Wong, A.: FoodFusion: A latent diffusion model for realistic food image generation. arXiv preprint arXiv:2312.03540 (2023) 4
17. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: Proc. IEEE International Conference on Multimedia and Expo (ICME). pp. 25–30 (2012) 4

18. Nakano, K., Horita, D., Sakata, N., Kiyokawa, K., Yanai, K., Narumi, T.: DeepTaste: Augmented reality gustatory manipulation with GAN-based real-time food-to-food translation. In: Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 212–223 (2019) 2, 4, 10, 11
19. Naritomi, S., Tanno, R., Ege, T., Yanai, K.: FoodChangeLens: CNN-based food transformation on HoloLens. In: Proc. IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp. 197–199 (2018) 4
20. Naritomi, S., Yanai, K.: Caloriecaptorglass: Food calorie estimation based on actual size using hololens and deep learning. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 818–819. IEEE (2020) 3
21. Narumi, T., Nishizaka, S., Kajinami, T., Tanikawa, T., Hirose, M.: Meta cookie+: an illusion-based gustatory display. In: International Conference on Virtual and Mixed Reality. pp. 260–269. Springer (2011) 3
22. Nishizawa, M., Jiang, W., Okajima, K.: Projective-ar system for customizing the appearance and taste of food. In: Proceedings of the 2016 workshop on Multimodal Virtual and Augmented Reality. pp. 1–6 (2016) 3
23. Plecher, D.A., Eichhorn, C., Steinmetz, C., Klinker, G.: Tracksugar. In: International Conference on Human-Computer Interaction. pp. 442–459. Springer (2020) 3
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. International Conference on Machine Learning (ICML). pp. 8748–8763 (2021) 10
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022) 4
26. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: Proc. German Conference on Pattern Recognition (GCPR). pp. 26–36 (2016) 10
27. Sakurai, S., Narumi, T., Ban, Y., Tanikawa, T., Hirose, M.: Calibratable: Tabletop system for influencing eating behavior. In: SIGGRAPH Asia 2015 Emerging Technologies, pp. 1–3 (2015) 3
28. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: Proc. European Conference on Computer Vision (ECCV). pp. 402–419 (2020) 10
29. Trinh, L.T., Hamagami, T.: Latent denoising diffusion GAN: Faster sampling, higher image quality. IEEE Access **12**, 78161–78172 (2024) 4
30. Ueda, J., Okajima, K.: AR food changer using deep learning and cross-modal effects. In: Proc. IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) (2019) 4
31. Vir, R., Madinei, P.: Archef: An ios-based augmented reality cooking assistant powered by multimodal gemini llm. arXiv preprint arXiv:2412.00627 (2024) 3
32. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 36324–36336 (2022) 5
33. Yu, D., Min, W., Jin, X., Jiang, Q., Jin, Y., Jiang, S.: Diverse and high-quality food image generation from only food names. ACM Transactions on Multimedia Computing, Communications and Applications **21**(5), 1–22 (2025) 4
34. Zhang, J., Zheng, K., Jiang, K., Wang, H., Stoica, I., Gonzalez, J.E., Chen, J., Zhu, J.: TurboDiffusion: Accelerating video diffusion models by 100-200 times. arXiv preprint arXiv:2512.16093 (2025) 4

35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 2242–2251 (2017) 4