FoRAGe: High-CTR Food Image Synthesis with Retrieval-Augmented Diffusion Model

Jiaxu Feng*[1], Xinyu Gao*[1], Muqi Huang*[2], Kanjun Xu[1], Yun Xiong†[1], Kun Zhou[2], Chuan Li[2], Feng Shi[2]
[1] Shanghai Key Lab of Data Science, College of Computer Science and Artificial Intelligence, Fudan University
[2] Rajax Network Technology (ele.me), Alibaba Group * Equal contribution † Corresponding author

Paper Link

Author's Website

Introduction

- High Click-Through Rate (CTR) imagery has proven commercial value for food delivery platforms like Ele.me.
- Our investigations reveal a positive correlation between appropriate food backgrounds and subsequent user engagement.
- Inpainting new backgrounds does not guarantee high CTR, and fine-tuning diffusion models for this purpose is prohibitively expensive for the fast-paced online food delivery advertising sector.
- Consequently, there is a lack of cost-effective, transferable generation frameworks tailored to high-CTR food images.

Highlight

- We propose a novel pipeline named **FoRAGe** (**Fo**od image **R**etrieval-**A**ugmented **Ge**neration), which integrates ControlNet-based RAG and a multimodal CTR prediction model for high-CTR food image synthesis.
- We are the first to deploy a CTR-optimized AIGC framework for customer-facing commercial scenarios on food delivery platforms, which is also adaptable to other domains.
- We introduce a new task termed Hi-Fi (**Hi**gh-**Fi**delity) Background Replacement, defined as image-driven background replacement with stylistic consistency and little background variations, guided by a similar reference image.
- We have deployed the framework on the Ele.me food delivery platform. Results of online A/B tests and ablation studies demonstrate the pipeline's superior performance in real-world scenarios.

Architecture

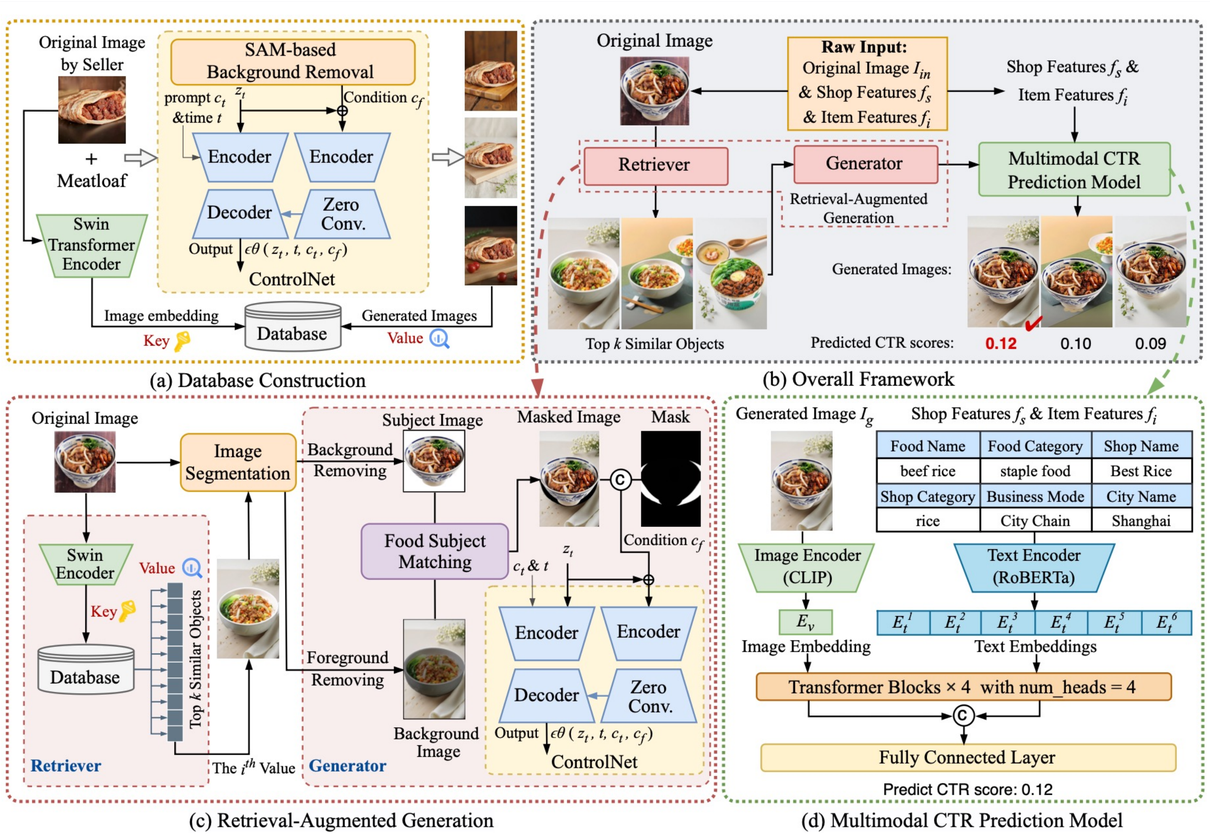

Figure 3: The architecture of our proposed FoRAGe pipeline. (a) is the process of database construction. (b) is the structure of the overall framework, which consists of (c), the retrieval-augmented generation module, and (d), the multimodal CTR prediction model. Notations: ⓒ indicates concatenation, and ⊕ means add.

Our method includes the following four main components, where Similar Image Retrieval and Hi-Fi Background Replacement serve as the retrieval step and generation step, respectively, forming a Retrieval-Augmented Generation process.

- **Database Construction**: We first construct a retrieval database containing a large number of high-quality food images with visually enhanced backgrounds.
- **Similar Image Retrieval**: Given an input food image, we use an image encoder to extract its embedding and utilize an image retriever to search for the top-k semantically similar background images in the database.
- **Hi-Fi Background Replacement**: We extract the food subject from the input image and the background from the retrieved image, match them together, and use a pre-trained diffusion model to fill in the vacancy.
- **Multimodal CTR Prediction**: We train a multimodal CTR prediction model that analyzes both the visual and textual features to predict CTR and select the optimal image for display.

Deployment

The pipeline is deployed on the Ele.me platform with advertisements displayed in vertical poster formats across Ele.me APP, WeChat, and Alipay mini-programs. They target users on both the homepage and category pages. All experiment groups are allocated the same amount of traffic and run for the same duration, ensuring fair and statistically sound comparisons.

Qualitative Analysis

Figure 4: Examples showcasing original images, images with outpainting, and images generated using our pipeline. The CTR is displayed below each image, with higher values highlighted in red to indicate increased consumer engagement.

Quantitative Analysis

Table 1: Online A/B Test. All values in the table represent relative uplifts compared to the baseline internal outpainting model. ↑ indicates that higher is better.

Date	CTR ↑	CTCVR ↑	Cost ↑	RPM ↑
Day 1	+3.78%	+3.17%	+2.63%	+2.19%
Day 2	+1.66%	+5.42%	+0.36%	+0.30%
Day 3	+2.20%	-2.57%	+0.98%	+1.92%
Day 4	+4.74%	+3.07%	+3.53%	+4.92%
Overall	+3.10%	+2.08%	+1.87%	+2.32%

Table 3: Ablation Study on the Multimodal CTR Prediction Model (MCPM). All values in the table represent relative uplifts compared to the baseline internal outpainting model. The best results are in bold.

Top-k	Configuration	CTR ↑
Top-5	w/o MCPM	+10.25%
	w/ MCPM	**+22.59%**
Top-10	w/o MCPM	+4.87%
	w/ MCPM	**+39.70%**

Table 2: Ablation Study on Top-k of the Retrieval-augmented Generation Module. All values in the table represent relative uplifts compared to the baseline internal outpainting model. The best results are in bold.

Top-k	Top-1				Top-3			
Food Categories	Staple Food	Snacks & Barbecue	Desserts & Beverages	**Overall**	Staple Food	Snacks & Barbecue	Desserts & Beverages	**Overall**
CTR ↑	+8.80%	+2.81%	+12.77%	+6.54%	+15.83%	-0.02%	+27.20%	+10.21%
RPM ↑	+9.72%	+38.81%	+49.36%	+23.83%	+14.02%	+9.31%	+52.88%	+15.48%
Cost ↑	+11.28%	+29.14%	+32.45%	+20.17%	+8.61%	-14.44%	+20.19%	+1.23%
PPC ↑	**+0.85%**	+35.01%	+32.45%	+16.23%	-1.57%	+9.33%	+20.19%	+4.78%
Top-k	Top-5				Top-10			
Food Categories	Staple Food	Snacks & Barbecue	Desserts & Beverages	**Overall**	Staple Food	Snacks & Barbecue	Desserts & Beverages	**Overall**
CTR ↑	+45.98%	-0.58%	+1.92%	+22.59%	**+66.72%**	**+4.97%**	**+31.68%**	**+39.70%**
RPM ↑	+16.01%	+14.45%	+12.12%	+14.22%	**+67.12%**	**+61.23%**	**+83.48%**	**+66.24%**
Cost ↑	+15.09%	+0.11%	+10.00%	+8.96%	**+86.66%**	**+53.60%**	**+59.25%**	**+71.44%**
PPC ↑	-20.53%	+15.12%	+10.00%	-6.83%	+0.24%	**+53.60%**	**+39.34%**	**+19.00%**

$$\text{CTR}_r = \frac{\text{click}}{\text{exposure}} \qquad \text{CTCVR} = \frac{\text{order}}{\text{exposure}} \qquad \text{RPM} = \frac{\text{revenue}}{\text{exposure}} \times 1000 \qquad \text{PPC} = \frac{\text{cost}}{\text{click}} \qquad \text{uplift} = \frac{M_o}{M_b} - 1$$

- **Table 1**: Online A/B test proves the images in our constructed retrieval database possess more visually appealing backgrounds compared to the baseline outpainting images. This ensures that the images generated from our pipeline possess more appealing backgrounds as well.
- **Table 2**: Ablation study evaluates the impact of varying top-k values during the retrieval phase. A higher value of k indicates that more images are retrieved from the database, thus providing the CTR prediction model with a broader selection of candidates. Experiments show that CTR uplift increases with larger values of k.
- **Table 3**: Ablation study demonstrates the efficacy of the Multimodal CTR Prediction Model (MCPM). In the experiments without MCPM, one of the k generated images was selected at random. The result underscores the critical role of the CTR prediction model in enhancing image selection quality.