Real-CalAgent: Degradation-Robust Agentic Calorie Estimation from Single Food Images

Longhao Sun[1], Yitian Ma[1], and Keiji Yanai[1]

The University of Electro-Communications, Tokyo, Japan
{sun-l, ma-y}@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract. Accurate calorie estimation from single food images is essential for daily dietary assessment, yet in practical applications, food images often suffer from complex mixed degradations that severely compromise existing methods. To address this issue, we propose Real-CalAgent, an agentic framework for calorie estimation from single food images with mixed types of degradation. Specifically, to remove degradation, we design Tool-Centric Greedy Recursive Image Restoration. This method employs a tool-centric greedy recursive strategy for tool scheduling, which enables more thorough degradation removal, thereby laying a solid foundation for calorie estimation. For the calorie estimation phase, we propose the Confidence-Aware Agentic Calorie Estimation method. This method uses MLLM to score the outputs from multiple calorie estimation methods according to the advantages and disadvantages of each calorie estimation method in the tool library. It then employs an MLP regressor to aggregate all the results, yielding the final calorie prediction. Extensive experiments demonstrate that our method achieves state-of-the-art (SOTA) performance in calorie estimation on both clean and degraded food images.

Keywords: Calorie estimation · Multimodal large language model · Image restoration

1 Introduction

The goal of single food image calorie estimation is to use one food image to estimate the total calories of all foods in that image. Accurate calorie estimation from food images is important for nutrition management, and health monitoring [14]. So far, researchers have developed many representative methods for calorie estimation from single food images. Examples include CalorieCLIP [8] and OmniFood [26]. They can achieve high accuracy on clean food images taken under good shooting conditions and with proper equipment.

Although the above methods have made great progress in calorie estimation accuracy, they are not suitable for food images with complex mixed degradations. All existing methods implicitly assume that the input food images are taken under good imaging conditions. However, in real scenarios, food images are often captured by mobile devices in uncontrolled environments and poor

shooting techniques. They frequently suffer from degradations such as motion blur, low light, and compression artifacts. These degradations can greatly affect the reliability of food-related vision tasks. For example, blur and noise may damage food boundaries. Low light may distort color and texture cues. Compression artifacts may harm fine-grained visual details. Therefore, when existing calorie estimation methods are directly applied to degraded food images, food type recognition, region segmentation, and depth estimation are all greatly affected. As a result, portion estimation and final calorie prediction may become unreliable. To our knowledge, there is currently no research on calorie estimation from degraded food images. Furthermore, even on clean images, individual calorie estimation methods are often tailored to specific food types or imaging conditions, limiting their general applicability. This motivates the need for a mechanism that can adaptively combine multiple estimators.

In recent years, multimodal large language models (MLLM) have shown strong abilities in visual understanding, reasoning, and tool use. They have also been applied to image restoration, for example, in methods like AgenticIR [32] and Q-Agent [30]. However, they all chose one tool for each type of degradation and only used it once throughout the entire image restoration process, which may lead to incomplete restoration under mixed or residual degradations. Moreover, some methods use MLLMs as a foundation model for estimating calories, such as FoodLMM [25] and CalorieLLaVA [19]. These methods provide new ideas for solving the task of calorie estimation from single food images. However, when faced with low-quality degraded image input, their accuracy will be greatly reduced, and relying solely on large models for calorie estimation can easily lead to hallucinations. Therefore, there is still a lot of room for improvement in accuracy.

Inspired by the above research, we propose Real-CalAgent, a multimodal agent framework based on MLLM for degraded food image calorie estimation. To reduce the impact of degradation on calorie estimation, we propose Tool-Centric Greedy Recursive Image Restoration to remove degradation. A greedy algorithm is employed to schedule degradation removal tools. After each degradation removal step, the MLLM is queried again to determine whether any degradation remains. This strategy effectively overcomes the issue of incomplete degradation removal. For calorie estimation, we adopt a confidence-aware agentic calorie estimation method. The agent invokes all available calorie estimation methods in our tool library, and the MLLM assigns confidence scores to the outputs of each method. Finally, an MLP regressor aggregates these results to produce the final calorie prediction. By integrating the respective strengths and compensating for the weaknesses of different methods, our method enhances both the accuracy and robustness of calorie estimation.

The main contributions of this paper are summarized as follows:

- We propose a novel calorie estimation agent to enhance the degradation robustness, Real-CalAgent, which is the first method that can accurately predict calories in degraded food images that contain complex mixed degradations.

- We design a tool-centric greedy recursive image restoration method that can repeatedly select suitable tools according to image quality until the quality stops improving and no degradation persists, establishing a solid foundation for subsequent calorie estimation.
- We introduce a confidence-aware agentic calorie estimation method. It can obtain accurate calorie estimates based on multiple calorie estimation methods and confidence scores obtained by MLLM.

2 Related Work

2.1 Agent-Based Image Restoration

Recent agent-based image restoration methods use MLLMs to recognize image degradations and dynamically call restoration tools, such as AgenticIR [32] and Q-Agent [30]. Their performance is better than other non-agent all-in-one image restoration methods, such as SwinIR [9] and AdaIR [4]. AgenticIR decomposes restoration into degradation perception, tool scheduling, execution, reflection, and rescheduling. If the quality decreases after recovery, it will revert to the previous state and then re-plan the recovery order. Q-Agent further introduces degradation-type-centric greedy image restoration to improve restoration quality. However, these methods do not consider the fact that a single type of degradation may not be completely removed by applying only one tool or by using it only once. Moreover, a given degradation type may not appear only once during an image restoration process, and different degradations can also overlap and interact with each other. In this case, their strategy that arranges each degradation type in a linear order and removes each type only once tends to result in poor degradation removal and residual degradation.

2.2 Calorie Estimation from Single Food Images

Single food image calorie estimation aims to predict calories from one food image. Existing methods usually estimate food categories, RGB color information, portion size, and volume, and then infer calories using nutrition information. Recent methods such as Swin-Nutrition [16], DPF-Nutrition [7], CalorieCLIP [8], and OmniFood [26] have achieved promising results on clean food images. Most of these methods use techniques like object detection, semantic segmentation, and monocular depth estimation. They first determine the food volume and food type. Then they estimate the final calorie amount. Although they can only obtain good results on clean food images, the methods themselves are not suitable for degraded food images in practice. Moreover, these methods are typically optimized for specific food categories or imaging conditions, leading to inconsistent performance when applied to diverse food images. Recently, MLLM-based calorie estimation methods have also been explored, which involve providing MLLM images and corresponding prompt words to infer calories, such as FoodLMM [25] and CalorieLLaVA [19]. However, they still struggle with degraded food images, and they do not use chain-of-thought (CoT) to enhance the MLLM's reasoning ability.

3 Methodology

3.1 Overall Framework

We propose **Real-CalAgent**, an MLLM-based multimodal agent for calorie estimation from single degraded food images. The overall pipeline is shown in Figure 1. Given an input RGB food image I_0, our goal is to estimate its calorie content under poor imaging conditions. The proposed framework contains two stages: agentic image restoration and agentic calorie estimation.

In the first stage, MLLM detects possible degradations through CoT reasoning. Based on the detected degradation types, the agent retrieves relevant restoration tools and performs tool-centric greedy recursive image restoration. The restoration process stops when the image quality no longer improves and MLLM detects no obvious residual degradation. In the second stage, the restored image I is used for calorie estimation. The confidence-aware agentic calorie estimation is applied. Multiple calorie estimation methods are invoked by the agent, and their predictions are fused by an MLP regressor using MLLM-derived confidence scores. The final calorie estimation result is output by the MLP regressor.

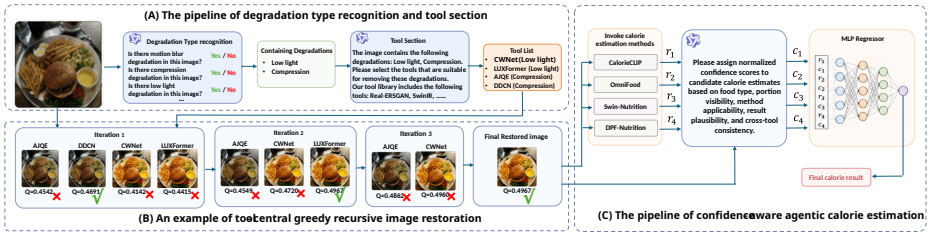


Fig. 1: The overall pipeline of our Real-CalAgent.

3.2 Tool-Centric Greedy Recursive Image Restoration

CoT-based Degradation Type Perception Recent studies, such as AgenticIR [32], have demonstrated that MLLMs are capable of identifying the types of degradations present in an image. However, they all rely on a generic prompt such as "What degradations does this image contain?" to query the MLLM, which greatly limits the model's ability to recognize degradation types. The resulting outputs are not only prone to hallucinations but also lack specificity and often miss certain degradation types.

To address this issue and further stimulate the degradation recognition capability of MLLM, we use CoT to decompose this complex task into multiple sub-tasks, as shown in Figure 1 (A). Specifically, we query the MLLM using a series of simple yes/no questions. The prompt follows the structure: "Is there $< deg >$ in this image?" where $< deg >$ represents a degradation type, such as

motion blur, compression artifacts, etc. The MLLM only needs to answer "yes" or "no". Once all degradation types have been queried, those types that receive a "yes" response are used as the result to guide the subsequent restoration tool selection.

Tool-Centric Greedy Recursive Strategy Previous image restoration agents have all performed restoration in a degradation-type-centric manner, where each degradation type is matched with a specific tool, and tools are applied in an order that maximizes image quality improvement at each step. However, they fail to consider that a single degradation type may not be completely removed by using only one corresponding tool or by applying it only once. Moreover, when multiple degradations are mixed, applying a degradation removal tool just once in a linear fashion is often insufficient for thorough restoration. To address this issue, we no longer arrange tool invocation orders based on degradation types. Instead, we propose a novel tool-centric greedy recursive strategy, detailed below.

First, we need to determine which image restoration tools are suitable. Based on the MLLM's response from the above section, we know the types of degradations present in the image. We then query the MLLM again with the following prompt: "The image contains the following degradations: $< img_deg >$. Our tool library includes the following tools: $< toolbox >$. Please select the tools that are applicable to remove these degradations." where $< img_deg >$ and $< toolbox >$ respectively represent the types of degradation contained in the image and the tools in our tool library. The MLLM then selects the corresponding tools for the identified degradation types. In this way, we can determine which tools to use, which can be denoted as $< tool >$.

Next, we perform image restoration using our tool-centric greedy recursive strategy. All tools in the list $< tool >$ selected by the MLLM are run once on the input low-quality food image. The result with the best perceptual quality is selected, and the tool used in the best result image in that iteration is then removed from the tool list $< tool >$. Starting from the best result image, the tools in the new tool list are applied to the current best result image, and the best result image is again selected as the starting point for the next iteration. This loop continues until the perceptual quality no longer improves. We have shown an example in Figure 1 (B). If the loop stops because the tool list $< tool >$ has been exhausted, the entire process is repeated. That is, we again ask the MLLM a series of "Is there $< deg >$ in this image?" questions. If all answers are no, the image restoration stops. If there is any yes answer, we continue with tool selection and our tool-centric greedy recursive strategy. To provide a more accurate description of this process, the specific pseudocode is presented in Algorithm 1.

Here, image perceptual quality is judged based on no-reference metrics. The no-reference metrics used are: CLIPIQA [21], CPBD [13], BRISQUE [12], and Hyper-IQA [18]. Let these metrics be denoted as $\{cl, cp, br, hy\}$, respectively.

Algorithm 1 Tool-Centric Greedy Recursive Image Restoration

Require: Low-quality image I_0, tool library $\mathcal{T}$, MLLM $\mathcal{M}$
Ensure: Restored image I

$I \leftarrow I_0$; $\mathcal{D} \leftarrow \textsc{DetectDegradations}(I, \mathcal{M})$
while $\mathcal{D} \neq \emptyset$ **do**
  $\mathcal{T}_r = \mathcal{T}_s \leftarrow \textsc{SelectTools}(\mathcal{D}, \mathcal{T}, \mathcal{M})$
  while $\mathcal{T}_s \neq \emptyset$ **do**
    $Q \leftarrow \textsc{EvalQuality}(I)$
    $(t^*, I^*) \leftarrow \arg\max_{t \in \mathcal{T}_s} \textsc{EvalQuality}(\textsc{ApplyTool}(t, I))$
    $Q^* \leftarrow \textsc{EvalQuality}(I^*)$
    if $Q^* \leq Q$ **then break**
    $I \leftarrow I^*$; $\mathcal{T}_s \leftarrow \mathcal{T}_s \setminus \{t^*\}$
  end while
  if $\mathcal{T}_r = \mathcal{T}_s$ **then break**
  $\mathcal{D} \leftarrow \textsc{DetectDegradations}(I, \mathcal{M})$
end while
return I

The final image quality can then be expressed by the following formula:

$$Q = \frac{cl + cp - br/100 + hy/100}{4}. \tag{1}$$

A higher quality score indicates better image quality.

3.3 Confidence-Aware Agentic Calorie Estimation

At present, single food image calorie estimation methods have different strengths, limitations, and applicable scenarios. Accuracy is related to the type of food and image status. To achieve accurate calorie estimation on diverse food images, we propose a confidence-aware agentic calorie estimation method. The pipeline is shown in Figure 1 (C).

After finishing degradation removal, the agent calls all calorie estimation tools in the tool library to estimate the calories of the restored food image I. Let r_i denote the calorie estimation result of the i-th tool, where i is the tool index and N is the number of calorie estimation tools in our tool library. We send the name and corresponding result of each calorie estimation method to the MLLM. The MLLM then assigns a confidence score c_i to each result based on factors such as image characteristics, food types, and method properties. The confidence scores are normalized to sum to 1. We use an example to show our prompts in Figure 2, the results provided by MLLM, and the corresponding scoring reasons.

In order to further improve the accuracy of calorie estimation, we introduce an MLP regressor to predict the final calorie. We convert raw calorie values into a compact log-space representation:

$$z_i = \log(r_i + 1). \tag{2}$$

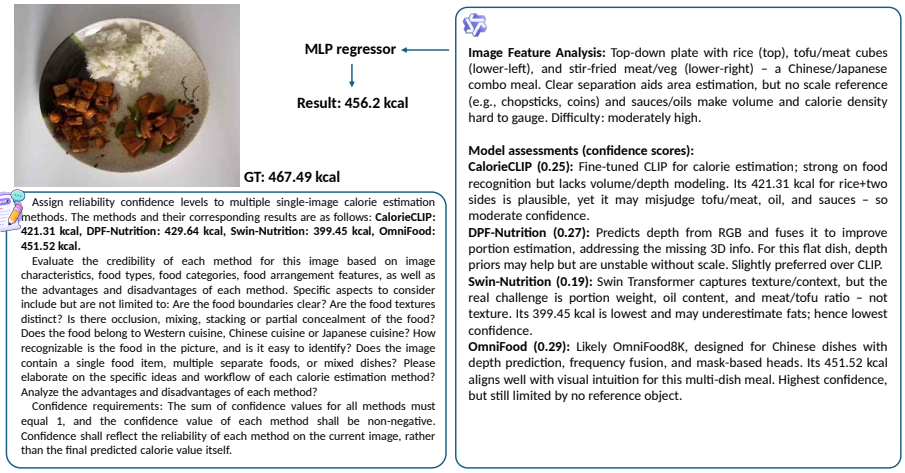


Fig. 2: A concrete example for our confidence-aware calorie estimation method.

This fixed log-space transformation compresses the dynamic range of calorie values. In order to more conveniently fit the true calorie value, we apply temperature smoothing to the confidence vector:

$$\bar{c}_i = \frac{\exp\left(\log(c_i + \epsilon)/T\right)}{\sum_{j=1}^{N} \exp\left(\log(c_j + \epsilon)/T\right)}, \tag{3}$$

where T is the temperature parameter and ϵ is a small constant for numerical stability. When $T > 1$, the confidence distribution becomes smoother while still satisfying $\sum_{i=1}^{N} \bar{c}_i = 1$. This reduces sensitivity to small absolute confidence differences while preserving their relative ordering. We can then obtain feature vectors:

$$\mathbf{x} = [z_1, ..., z_N, \bar{c}_1, ..., \bar{c}_N, z_1\bar{c}_1, ..., z_N\bar{c}_N], \tag{4}$$

where $N = 4$. By inputting $\mathbf{x}$ into the MLP, the calorie results in log space can be obtained:

$$\hat{z} = MLP(\mathbf{x}). \tag{5}$$

The final calorie estimation result can be obtained using the following formula:

$$\hat{y} = \exp(\hat{z}) - 1. \tag{6}$$

Let y denote the ground-truth (GT) calorie and z denote the log-space representation of y. The optimization objective can be expressed as:

$$\mathcal{L} = \text{SmoothL1}(\hat{z}, z) + \lambda_{rel} \frac{|\hat{y} - y|}{y + \epsilon}. \tag{7}$$

In this equation, the first term provides stable supervision in the log space. The second term directly optimizes the relative calorie estimation error, which

Table 1: The main tools in our tool library.

Task	Tools
Gaussian Denoising	SCUNet [28], NAFNet [2]
Compression Artifact Removal	AJQE [24], DDCN [6]
Deraining	DRT [10], MPRNet [27]
Dehazing	RIDCP [23], DehazeFormer [17]
Motion Deblurring	NAFNet, AdaRevD [11]
Defocus Deblurring	DRBNet [15], 3SDeblur [3]
Low-light Enhancement	LUXFormer [31], CWNet [29]
Super-Resolution	Real-ESRGAN [22]
Calorie Estimation	CalorieCLIP [8], Swin-Nutrition [16], OmniFood [26], DPF-Nutrition [7]

explicitly penalizes relative calorie estimation errors, encouraging predictions to remain proportional to the ground-truth calorie magnitude. We set $T = 1.3$ and $\lambda_{rel} = 0.5$. This MLP regressor network consists of six layers of fully connected neurons. The number of neurons in each layer is 12, 24, 36, 24, 12, 1. The 12 neurons in the first layer are used to receive the 12 elements of the input feature vector. The final neuron outputs the predicted log-space calorie value.

4 Experiment

4.1 Implementation Details

Table 1 shows the main tools in our tool library and the types of degradations we can handle. Moreover, in this paper, we select Qwen3-VL [1] as the MLLM. The specific version is Qwen3-VL-8B-Instruct.

During the training of the MLP regressor, the batch size is set to 128, and the learning rate is set to 5×10^{-4}. The training was conducted on an NVIDIA RTX A6000 GPU for 50 epochs.

For datasets, we construct two degraded datasets for calorie estimation, Deg-Nutrition5k and Deg-OmniFood8K, based on Nutrition5k [20] and Omni-Food8K [26], respectively. Since most original images are relatively clean, we apply a mixed degradation pipeline from Real-ESRGAN [22] that introduces blur, resizing, noise, and JPEG compression. To better approximate practical imaging conditions, we additionally include low-light and motion blur effects, with random order, intensity, and type for each image. All datasets come with official train and test splits. We adopt percentage mean absolute error (PMAE, %) as the evaluation metric. Given the ground-truth calorie value y_i and the predicted value $\hat{y}_i$ for the i-th sample, PMAE can be defined as:

$$\text{PMAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} y_i} \times 100\%. \tag{8}$$

where N denotes the number of test samples. A lower PMAE indicates more accurate calorie estimation.

Table 2: Quantitative comparison of calorie estimation in terms of PMAE (%, ↓). The best results are highlighted in **bold**.

Method	OmniFood8K	Deg-OmniFood8K	Nutrition5k	Deg-Nutrition5k
CalorieCLIP	10.9	18.7	15.6	21.6
Swin-Nutrition	11.7	22.1	16.2	25.9
DPF-Nutrition	12.5	23.6	14.7	23.1
FoodLMM	24.2	25.8	26.6	28.5
OmniFood	8.8	21.9	14.1	24.2
Real-CalAgent (Ours)	**8.2**	**9.8**	**11.9**	**14.1**

Table 3: Quantitative comparison of image restoration on the test set of Deg-Nutrition5k. The best results are highlighted in **bold**.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Q-Agent [30]	26.46	0.801	**0.185**	0.162
AdaIR [4]	25.75	0.817	0.191	0.141
AgenticIR [32]	26.12	0.797	0.197	0.189
SwinIR [9]	24.25	0.741	0.267	0.227
Real-CalAgent (Ours)	**26.78**	**0.831**	0.191	**0.138**

4.2 Quantitative Comparison of Calorie Estimation

The results are reported in Table 2. As shown in this table, Real-CalAgent achieves competitive performance on clean datasets and shows clear advantages on degraded datasets.

On clean datasets, our method achieves the lowest PMAE score. We use an MLLM to assign confidence scores, effectively filtering out unreasonable calorie values, which helps select more accurate results. We further fit the final calorie values through an MLP regressor, achieving the best results on both datasets, outperforming methods that rely on only a single calorie estimation approach.

On degraded datasets, we still achieve the best performance. Other calorie estimation methods do not account for degraded input images, resulting in poor robustness, and thus, their PMAE scores increase substantially. Benefiting from our tool-centric greedy recursive image restoration, we improve robustness against degraded images. Therefore, we still obtain the lowest PMAE score.

4.3 Quantitative Comparison of Image Restoration

As shown in Table 3, Real-CalAgent achieves the best overall performance among all compared methods. SwinIR and AdaIR show relatively limited performance because a single restoration model is difficult to handle mixed and random degradations. Agent-based methods, including Q-Agent and AgenticIR, obtain better results by adaptively selecting restoration operations. Compared with them, Real-CalAgent further improves PSNR, SSIM, and DISTS [5], demonstrating that the proposed tool-centric greedy recursive strategy can more effectively use

restoration tools for degraded food images. Although Q-Agent achieves a slightly lower LPIPS score, the difference is marginal. Overall, these results verify that the proposed restoration stage can better remove degradations and improve image quality before calorie estimation.

4.4 Qualitative Comparison of Image Restoration

To more intuitively evaluate the performance of each method, visual results are presented in Figure 3. As shown in the figure, it can be seen that our method is the most thorough in removing degradation. SwinIR, an all-in-one image restoration method for non-agents, leaves visible residual degradation under low-light and mixed degradation conditions. Because it is not designed to handle all degradation types considered in this work. Other agent-based image restoration methods, AgenticIR and Q-Agent, have more comprehensive functions and have a certain enhancement effect on low light, but they are not thorough. In the figure of the second line, SwinIR has some residual motion blur and noise, while Q-Agent has less residual noise but still has residual motion blur. And our method for image restoration is relatively thorough, with almost no residual degradation.



Fig. 3: Visual results of different image restoration methods

5 Ablation Study

5.1 The Effectiveness of MLP Regressor and Confidence

To verify the effectiveness of our proposed MLP regressor and confidence score, we designed four different scenarios, and the experimental results are shown in Table 4.

As shown in the table, directly averaging the calorie estimates from all methods yields the worst overall performance, because it lacks the ability to filter out unreliable outliers. Selecting the calorie value with the highest confidence as the final output also overall underperforms compared to Ours and the variant that uses only the MLP without confidence. This is because the MLP can further

Table 4: Quantitative result of the MLP regressor and confidence ablation study in terms of PMAE (%, ↓).

Variant	OmniFood8K	Deg-OmniFood8K	Nutrition5k	Deg-Nutrition5k
Highest-confidence Selection	8.7	14.1	13.7	16.4
Average all calorie results	11.0	21.6	15.2	23.7
w/ MLP but w/o confidence	9.1	13.3	14.7	15.2
w/o temperature smoothing	8.6	11.4	13.1	14.6
Real-CalAgent (Ours)	8.2	9.8	11.9	14.1

eliminate systematic biases, leading to more accurate results. Without temperature smoothing, performance consistently decreases, suggesting that directly using raw confidence magnitudes makes the fusion model overly sensitive to small confidence differences. Our method still achieves the best performance. This demonstrates the effectiveness of our MLP regressor, its temperature smoothing, and the confidence design.

In addition, even without the MLP regressor, our performance is better than other calorie estimation methods shown in Table 2, which also indicates that the confidence scored by MLLM is effective and reasonable.

5.2 The Effectiveness of the Tool-Centric Greedy Recursive Strategy

To evaluate the effectiveness of the tool-centric greedy recursive strategy and investigate the impact of different tool invocation orders, we compare our method with four alternative restoration strategies: 1) Reverse-Order Restoration: images are restored sequentially in reverse order based on the sequence in which degradation is applied; 2) Random Restoration: the restoration tool order is randomly selected from the degradation types perceived by the MLLM; 3) Rollbackable Restoration: A rollbackable search strategy used by AgenticIR [32]; 4) Q-Agent Restoration: a degradation-type-centric greedy strategy is used to determine the restoration order. The specific implementation method is that our tool invocation order adopts the different strategies mentioned above, while other parts of our framework remain unchanged. The results are reported in Table 5.

Random Restoration, which randomly generates the tool invocation order, cannot produce a targeted restoration sequence and thus yields the poorest performance. Reverse-Order Restoration does not adaptively select the degradation removal tool according to the current image quality, so it performs only marginally better than random. Rollbackable Restoration and Q-Agent Restoration overlook the fact that repeated use of a single tool or interleaved use of multiple tools can remove degradation more thoroughly, leading to inferior results compared to ours. In summary, our strategy attains competitive performance and effectively achieves image restoration.

Table 5: Quantitative comparison of different restoration methods on the test set of Deg-Nutrition5k. The best results are highlighted in **bold**.

Strategy	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Reverse-Order Restoration	25.14	0.673	0.247	0.218
Random Restoration	24.62	0.623	0.289	0.254
Rollbackable Restoration	26.34	0.805	0.194	0.186
Q-Agent Restoration	26.56	0.812	**0.185**	0.157
Real-CalAgent (Ours)	**26.78**	**0.831**	0.191	**0.138**

Table 6: Ablation study on different image restoration methods for calorie estimation on degraded food image datasets in terms of PMAE (%, ↓).

Dataset	SwinIR	AdaIR	AgenticIR	Q-Agent	Real-CalAgent (Ours)
Deg-OmniFood8K	19.1	13.2	12.7	11.8	**9.8**
Deg-Nutrition5k	21.8	18.3	18.6	16.5	**14.1**

5.3 Impact of Image Restoration on Calorie Estimation

To verify whether image restoration directly benefits calorie estimation, we conduct an ablation study on degraded food image datasets. As shown in Table 6, SwinIR yields the worst calorie estimation performance on degraded food images, because it handles only a limited range of degradation types and leaves substantial residual degradations that severely compromise estimation accuracy. Other methods also fail to remove degradations completely, which negatively affects calorie estimation methods that are sensitive to RGB color information, leading to higher PMAE than ours. Our image restoration method achieves the best performance, as it removes degradations to the greatest extent and restores image fidelity most effectively. Overall, our image restoration method proves effective for calorie estimation and consistently outperforms other image restoration methods.

Furthermore, we also conduct experiments that integrate other calorie estimation methods into our image restoration method, as shown in Table 7. On clean datasets, restoration produces only minor performance changes, since most images are already of relatively high quality and unnecessary restoration is largely avoided. On degraded images, restoration consistently improves all methods, preserving essential food appearance, structure, and portion cues. Notably, even without restoration, our method still performs best on degraded images because of the MLLM confidence scores and MLP regressor. Overall, our restoration module is universally effective across other methods, and image restoration is helpful for estimating the calories in degraded food images.

5.4 The Effectiveness of the Multi-Metric Quality Criterion

To validate the multi-metric quality criterion in Equation (1), we progressively incorporate CLIPIQA [21], CPBD [13], BRISQUE [12], and Hyper-IQA [18]

Table 7: Effectiveness of our image restoration as a preprocessing module for different calorie estimation methods in terms of PMAE (%, ↓). "Rest." denotes our image restoration method.

Method	OmniFood8K		Deg-OmniFood8K	
	w/o Rest.	+ Rest.	w/o Rest.	+ Rest.
CalorieCLIP	10.9	10.9	18.7	12.7
Swin-Nutrition	11.7	11.6	22.1	14.6
DPF-Nutrition	12.5	12.3	23.6	15.4
FoodLMM	24.2	24.2	25.8	22.7
OmniFood	8.8	8.8	21.9	11.2
Ours	8.2	8.2	15.4	9.8

Table 8: Ablation study of the quality metrics used in Equation (1) on Deg-Nutrition5k.

CLIPIQA	CPBD	BRISQUE	Hyper-IQA	PSNR↑	LPIPS↓	DISTS↓	PMAE↓
✓	×	×	×	26.18	0.215	0.171	16.4
✓	✓	×	×	26.39	0.203	0.157	15.2
✓	✓	✓	×	26.61	**0.187**	0.147	14.5
✓	✓	✓	✓	**26.78**	0.191	**0.138**	**14.1**

and evaluate their effects on both image restoration and downstream calorie estimation.

As shown in Table 8, progressively incorporating complementary no-reference image quality metrics generally improves both restoration quality and downstream calorie estimation accuracy. Using CLIPIQA alone yields a PSNR of 26.18 dB and a PMAE of 16.4%. Adding CPBD and BRISQUE further improves the restoration results and reduces PMAE to 14.5%, indicating that different quality metrics provide complementary cues for selecting more reliable restoration outputs. With all four metrics, our method achieves the best PSNR (26.78), DISTS (0.138), and PMAE (14.1%). Although the three-metric variant obtains a slightly lower LPIPS, the full criterion achieves the best overall performance, validating the effectiveness of the multi-metric quality assessment used in Equation (1).

6 Conclusion

In this paper, we propose Real-CalAgent, the first multimodal agent framework specifically designed for calorie estimation from degraded food images. The framework first performs tool-centric greedy recursive image restoration and then uses confidence-aware agentic calorie estimation. Our method delivers strong performance on both clean and degraded images, overcoming the problem of poor accuracy of current calorie methods on degraded images.

Although our method enhances robustness to degradation, a drawback is the lack of a dedicated dataset of food images collected under degraded conditions in

real-world scenarios. Therefore, our current evaluation mainly relies on synthetically degraded datasets. In the future, collecting and annotating such datasets will be an important direction.

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

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision. pp. 17–33. Springer (2022)
3. Chen, Y., Li, Z., Yao, B., Li, X., Wang, W., Chen, Q.: 3sdeblur: Three-stage image defocus deblurring guided by residual prior. In: International Conference on Neural Information Processing. pp. 77–91. Springer (2025)
4. Cui, Y., Zamir, S.W., Khan, S., Knoll, A., Shah, M., Khan, F.: Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In: International Conference on Learning Representations. vol. 2025, pp. 101306–101327 (2025)
5. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2020)
6. Guo, J., Chao, H.: Building dual-domain representations for compression artifacts reduction. In: European Conference on Computer Vision. pp. 628–644. Springer (2016)
7. Han, Y., Cheng, Q., Wu, W., Huang, Z.: Dpf-nutrition: Food nutrition estimation via depth prediction and fusion. Foods **12**(23), 4293 (2023)
8. Haplo, L.: Calorieclip: Accurate food calorie estimation from images (2024), `https://huggingface.co/jc-builds/CalorieCLIP`
9. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
10. Liang, Y., Anwar, S., Liu, Y.: Drt: A lightweight single image deraining recursive transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 589–598 (2022)
11. Mao, X., Li, Q., Wang, Y.: Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25681–25690 (2024)
12. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
13. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Transactions on Image Processing **20**(9), 2678–2683 (2011)
14. Queipo-Alvarez, P., Munoz-Organero, M., Abdul Wahab, A., Mohd Rahim, M.S.: Food calorie estimation using computer vision: A comparative review of models and approaches. Malaysian Journal of Medicine & Health Sciences **21**, 194 (2025)

15. Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16304–16313 (2022)
16. Shao, W., Hou, S., Jia, W., Zheng, Y.: Rapid non-destructive analysis of food nutrient content using swin-nutrition. Foods **11**(21), 3429 (2022)
17. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)
18. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2020)
19. Tanabe, H., Yanai, K.: Caloriellava: Image-based calorie estimation with multimodal large language models. In: International Conference on Pattern Recognition. pp. 63–75. Springer (2024)
20. Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., Sim, J.: Nutrition5k: Towards automatic nutritional understanding of generic food. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8903–8911 (2021)
21. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
22. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1905–1914 (2021)
23. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22282–22291 (2023)
24. Yang, J., Xing, Q., Xu, M., Qiao, M.: Uncover treasures in dct: Advancing jpeg quality enhancement by exploiting latent correlations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17598–17607 (2025)
25. Yin, Y., Qi, H., Zhu, B., Chen, J., Jiang, Y.G., Ngo, C.W.: Foodlmm: A versatile food assistant using large multi-modal model. IEEE Transactions on Multimedia (2025)
26. Yu, D., Min, W., Jiang, Q., Lin, X., Jin, X., Jiang, S.: Omnifood8k: Single-image nutrition estimation via hierarchical frequency-aligned fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41562–41572 (2026)
27. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14821–14831 (2021)
28. Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.P., Timofte, R., Gool, L.V.: Practical blind image denoising via swin-conv-unet and data synthesis. Machine Intelligence Research **20**(6), 822–836 (2023)
29. Zhang, T., Liu, P., Lu, Y., Cai, M., Zhang, Z., Zhang, Z., Zhou, Q.: Cwnet: Causal wavelet network for low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8789–8799 (2025)
30. Zhou, Y., Cao, J., Zhang, Z., Wen, F., Jiang, Y., Jia, J., Liu, X., Min, X., Zhai, G.: Q-agent: quality-driven chain-of-thought image restoration agent through robust multimodal large language model. arXiv preprint arXiv:2504.07148 (2025)

31. Zhou, Y., Gao, X., Guo, J., Li, G., Wang, L., Liu, J.: Luxformer: Low-light image enhancement via joint spatial-frequency illumination modeling. Journal of King Saud University Computer and Information Sciences **37**(8), 251 (2025)
32. Zhu, K., Gu, J., You, Z., Qiao, Y., Dong, C.: An intelligent agentic system for complex image restoration problems. In: International Conference on Learning Representations. vol. 2025, pp. 57985–58013 (2025)