The 29th Meeting on Image Recognition and Understanding

Real-CalAgent: Agentic Calorie Estimation from Single Real-World Food Images

LONGHAO SUN[1,a)] YITIAN MA[1,b)] KEIJI YANAI[1,c)]

Abstract

Accurate calorie estimation from single food images is important for dietary assessment in daily life. However, real-world food images often suffer from unknown and complex degradations that affect calorie prediction. To address this problem, we propose Real-CalAgent, a multimodal large language model (MLLM) agent framework for calorie estimation from single real-world food images. The proposed method first uses our agent to identify possible image degradations through chain-of-thought (CoT) and performs tool-centric greedy recursive image restoration. After restoration, the agent determines whether a reference object is available. If so, it performs our designed CoT-based calorie estimation via MLLM. Otherwise, it uses the novel confidence-aware calorie estimation method. Multiple calorie estimation tools are invoked, and the final result is selected based on confidence scores assigned by MLLM. This is the first method that can accurately predict calories for real-world food images.

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 dietary assessment, nutrition management, and health monitoring. So far, researchers have developed many representative methods for calorie estimation from single food images. Examples include CalorieCLIP [7] and OmniFood [22]. They can accurately determine the calories contained in food.

Although the above methods have made great progress in calorie estimation accuracy, they are not suitable for real-world food images with complex and unknown 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. 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 real-world 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.

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 [28] and Q-Agent [26]. Some methods now use MLLMs as a foundation for estimating calories, such as FoodLMM [21]. These methods provide new ideas for solving the task of calorie estimation from single real-world food images.

Inspired by this research, we propose Real-CalAgent, a multimodal agent framework based on an MLLM for real-world food image calorie estimation. The framework first uses CoT reasoning to perceive various degradations in the image. It then performs quality-driven greedy scheduling based on restoration tools. It iteratively selects the restoration result that best improves the no-reference image quality assessment score, until the image quality converges and no obvious residual degradation remains. After that, the agent moves to the calorie estimation stage. If a reference object for volume estimation exists in the image, it uses the reference-assisted method. Otherwise, it calls several calorie estimation tools. The MLLM uses CoT reasoning to understand the image and each calorie estimation method from multiple aspects. It then performs confidence scoring based on food type, portion visibility, occlusion and stacking relationships, and method applicability. Finally, it selects the result with the highest confidence as the calorie prediction.

The main contributions of this paper are summarized as follows:

- We propose a novel calorie estimation method, Real-CalAgent, which is the first method that can accurately predict calories in real-world food images that contain complex and unknown degradations.
- We design a novel tool-centric greedy recursive image restoration method. It can repeatedly select suitable tools according to image quality until the quality stops

[1] The University of Electro-Communications
[a)] sun-l@mm.inf.uec.ac.jp
[b)] ma-y@mm.inf.uec.ac.jp
[c)] yanai@cs.uec.ac.jp

improving and no degradation persists.

- We introduce novel CoT-based calorie estimation via MLLM and confidence-aware calorie estimation methods. They can obtain accurate calorie estimates respectively in cases with and without a reference object.

2. Related Works

2.1 Agent-based image restoration

Recent agent-based image restoration methods use MLLMs to recognize image degradations and dynamically call restoration tools. AgenticIR [28] decomposes restoration into degradation perception, tool scheduling, execution, reflection, and rescheduling. Q-Agent [26] further introduces quality-driven 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; different degradations can also overlap and interact with each other.

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, food regions, portion size, volume, or depth, and then infer calories using nutrition information. Recent methods such as Swin-Nutrition [13], DPF-Nutrition [6], CalorieCLIP [7], and OmniFood [22] 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. However, they can only obtain good results on clean food images, and the methods themselves are not suitable for real-world food images.

3. Methodology

3.1 Overall Framework

We propose **Real-CalAgent**, an MLLM-based multimodal agent for calorie estimation from a single real-world food image. Given an input RGB food image I_0, our goal is to estimate its calorie content under realistic imaging conditions, where image degradations and scale ambiguity often exist. The proposed framework contains two stages: agentic image restoration and agentic calorie estimation through MLLM and CoT.

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 is used for calorie estimation. The agent first checks whether reference objects are available. If so, CoT-based calorie estimation via MLLM is applied; otherwise, multiple single food calorie estimation tools are invoked, and the agent selects the final result through confidence scoring obtained by MLLM.

3.2 Greedy iterative CoT image restoration

3.2.1 Degradation type recognition based on CoT

Recent studies such as AgenticIR [28] 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 MLLMs, we use CoT to decompose this complex task into multiple sub-tasks, as shown in Figure 1. 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 subsequent restoration tasks.

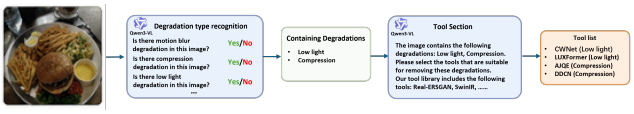


Fig. 1 The pipeline of degradation type recognition based on MLLM CoT

3.2.2 Tool-Centric Greedy Recursive Image Restoration

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 Image Restoration method.

First, we need to determine which image restoration tools are available. Based on the MLLM's response from the previous 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 suitable for removing 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 a 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 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 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 2. 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 "Does it contain $< deg >$?" 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 image restoration.



Fig. 2 An example of our tool-centric greedy recursive image restoration

Here, image quality is judged based on no-reference metrics. The no-reference metrics used are: CLIPIQA [17], CPBD [11], BRISQUE [10], and Hyper-IQA [15]. Let these metrics be denoted as $\{cl, cp, br, hy\}$, respectively. The final image quality can then be expressed by the following formula:

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

A higher quality score indicates better image quality.

3.3 CoT-based Calorie Estimation via MLLM

After the restoration of real-world food images is completed, we use agents and CoT for calorie estimation. We first query the MLLM: "Does the image contain a reference object that can be used to infer physical scale?" If the answer is yes, we use our CoT-based calorie estimation via MLLM. If the answer is no, we use our proposed confidence-aware calorie estimation method. Finally, we obtain the calorie estimation result. The pipeline has been shown in Figure 3.

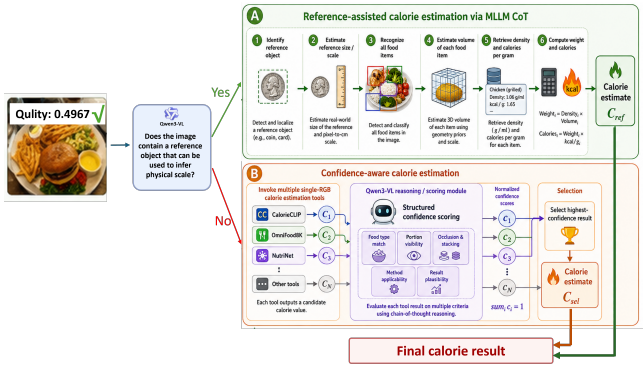


Fig. 3 The pipeline of CoT-based Calorie Estimation Agent

3.3.1 CoT-based calorie estimation via MLLM

Fridolfsson et al. [4] have demonstrated that large language models can achieve good calorie estimation results when a reference object is available. Inspired by their work, we propose to use only an MLLM to estimate calories in the presence of a reference object. However, the prompt they used is overly simple and direct. They simply ask the model to estimate the volume based on the reference object and then calculate the calories according to the food type. This fails to fully leverage the MLLM's prior knowledge and its ability to reason carefully.

To address this issue, we introduce CoT reasoning into MLLM-based calorie estimation and propose a novel method, CoT-based calorie estimation via MLLM. The core idea is to decompose the complex task of reference object identification, reference object size estimation, food volume computation, food type recognition, and final calorie calculation into multiple simpler sub-tasks.

Specifically, we decompose the prompt for calorie estimation into the following steps: (1) Please identify what the reference object is. (2) Based on your prior knowledge and the image, estimate the volume and size of the reference object. (3) Please carefully and thoroughly identify all food types in the image. (4) Based on the estimated volume and size of the reference object and the relative scale in the image, estimate the volume of each food item. (5) For each food item, look up the calories per gram and its density. (6) Calculate the weight of each food item based on its volume and density. We use these prompt words to sequentially query the MLLM and output the final calorie estimation result.

3.3.2 Confidence-aware 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 any food image as much as possible, we propose a confidence-aware calorie estimation method.

When the MLLM determines that the food image has no

Table 1 The main tools in our tool library.

Task	Label	Tools
Gaussian Denoising	DN	SCUNet [24], NAFNet [2]
Compression Artifact Removal	CA	AJQE [20], DDCN [5]
Deraining	DR	DRT [8], MPRNet [23]
Dehazing	DH	RIDCP [19], DehazeFormer [14]
Motion Deblurring	MDB	NAFNet, AdaRevD [9]
Defocus Deblurring	DDB	DRBNet [12], 3SDeblur [3]
Low-light Enhancement	LE	LUXFormer [27], CWNet [25]
Super-Resolution	SR	Real-ESRGAN [18]
Calorie Estimation	CE	CalorieCLIP [7], Swin-Nutrition [13], OmniFood [22], DPF-Nutrition [6]

reference object, the agent calls all calorie estimation tools in the tool library to estimate the calories of the restored food image. 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. Finally, we select the result with the highest confidence score as our final calorie estimation result.

4. Experiment

4.1 Implementation details

Table 1 shows the main tools in our tool library and the types of degradations we can handle. It is worth noting that we do not use diffusion-based image restoration methods, because our ultimate goal is calorie estimation. Diffusion-based methods may introduce hallucinated objects and textures, which could interfere with calorie estimation. Moreover, in this paper, we select Qwen3-VL [1] as the MLLM. The specific version is Qwen3-VL-8B-Instruct.

4.2 Testing Dataset

We use OmniFood8K and Nutrition5k [16] to construct two degraded food image test datasets for calorie estimation without reference objects: Deg-OmniFood8K and Deg-Nutrition5k. Since most original images are captured under relatively clear conditions, we apply a mixed degradation pipeline to simulate real-world degradations. The pipeline is based on the Real-ESRGAN degradation process, which includes blur, resizing, noise, and JPEG compression. To better match practical food photography scenarios, we further add additional degradations such as low light and motion blur. For each image, one or multiple degradation types are randomly selected and applied. All compared methods are evaluated on the same degraded test images for a fair comparison.

4.3 Quantitative comparison of calorie estimation

We use the percentage mean absolute error (PMAE, %) as the metric, where lower values indicate better performance. 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. While existing methods suffer from large error increases under degraded imaging conditions, Real-CalAgent only increases from 8.7% to 10.6% on OmniFood8K and from 13.2% to 14.9% on Nutrition5k. This demonstrates that the proposed degradation-aware restoration and confidence-aware calorie estimation strategy effectively improve robustness for real-world degraded food images.

Table 2 Quantitative comparison of calorie estimation on clean and degraded food image test datasets in terms of PMAE (%, ↓). The best result in each column is 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)	**8.7**	**10.6**	**13.2**	**14.9**

Table 3 Quantitative comparison of image restoration on representative subsets of degraded food image datasets. The best results are highlighted in **bold**.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑
Q-Agent	26.91	0.824	**0.181**	0.657
AgenticIR	26.43	0.812	0.196	0.632
SwinIR	24.86	0.765	0.245	0.557
Real-CalAgent (Ours)	**27.48**	**0.839**	0.187	**0.661**

4.4 Quantitative comparison of image restoration

As shown in Table 3, Real-CalAgent achieves the best overall performance among all compared methods. SwinIR shows relatively limited performance because a single restoration model is difficult to handle mixed and unknown 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 CLIPIQA [17], demonstrating that the proposed tool-centric greedy recursive strategy can more effectively combine 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 improve image quality before calorie estimation.

5. Conclusion

In this paper, we propose Real-CalAgent, a multimodal agent framework for calorie estimation from a single real-world food image. The framework first performs degradation-aware image restoration and then estimates calories according to the availability of reference objects. When no reference object is available, multiple calorie estimation tools are invoked, and the final result is selected by MLLM-based confidence scoring.

Although Real-CalAgent improves the practicality of calorie estimation under degraded imaging conditions, it still has some limitations. This method runs relatively slowly, and its effectiveness is greatly affected by the MLLM version. In the future, reinforcement learning and other methods will be introduced to enable MLLM to generate trustworthy tool execution sequences and accelerate the running speed.

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. and Sun, J.: Simple baselines for image restoration, *European Conference on Computer Vision*, Springer, pp. 17–33 (2022).

[3] Chen, Y., Li, Z., Yao, B., Li, X., Wang, W. and Chen, Q.: 3SDeblur: Three-Stage Image Defocus Deblurring Guided by Residual Prior, *International Conference on Neural Information Processing*, Springer, pp. 77–91 (2025).

[4] Fridolfsson, J., Sjöberg, E., Thiwång, M. and Pettersson, S.: Performance evaluation of Three Large Language Models for Nutritional Content Estimation from Food Images, *Current Developments in Nutrition*, p. 107556 (2025).

[5] Guo, J. and Chao, H.: Building dual-domain representations for compression artifacts reduction, *European Conference on Computer Vision*, Springer, pp. 628–644 (2016).

[6] Han, Y., Cheng, Q., Wu, W. and Huang, Z.: Dpf-nutrition: Food nutrition estimation via depth prediction and fusion, *Foods*, Vol. 12, No. 23, p. 4293 (2023).

[7] Haplo, L.: CalorieCLIP: Accurate Food Calorie Estimation from Images, Hugging Face model repository (2024). Available at: `https://huggingface.co/jc-builds/CalorieCLIP`.

[8] Liang, Y., Anwar, S. and Liu, Y.: Drt: A lightweight single image deraining recursive transformer, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 589–598 (2022).

[9] Mao, X., Li, Q. and Wang, Y.: Adarevd: Adaptive patch exiting reversible decoder pushes the limit of image deblurring, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25681–25690 (2024).

[10] Mittal, A., Moorthy, A. K. and Bovik, A. C.: No-reference image quality assessment in the spatial domain, *IEEE Transactions on Image Processing*, Vol. 21, No. 12, pp. 4695–4708 (2012).

[11] Narvekar, N. D. and Karam, L. J.: A no-reference image blur metric based on the cumulative probability of blur detection (CPBD), *IEEE Transactions on Image Processing*, Vol. 20, No. 9, pp. 2678–2683 (2011).

[12] Ruan, L., Chen, B., Li, J. and Lam, M.: Learning to deblur using light field generated and real defocus images, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16304–16313 (2022).

[13] Shao, W., Hou, S., Jia, W. and Zheng, Y.: Rapid non-destructive analysis of food nutrient content using swin-nutrition, *Foods*, Vol. 11, No. 21, p. 3429 (2022).

[14] Song, Y., He, Z., Qian, H. and Du, X.: Vision transformers for single image dehazing, *IEEE Transactions on Image Processing*, Vol. 32, pp. 1927–1941 (2023).

[15] Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J. and Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network, *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 3667–3676 (2020).

[16] Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T. and Sim, J.: Nutrition5k: Towards automatic nutritional understanding of generic food, *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 8903–8911 (2021).

[17] Wang, J., Chan, K. C. and Loy, C. C.: Exploring clip for assessing the look and feel of images, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 2, pp. 2555–2563 (2023).

[18] Wang, X., Xie, L., Dong, C. and Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914 (2021).

[19] Wu, R.-Q., Duan, Z.-P., Guo, C.-L., Chai, Z. and Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22282–22291 (2023).

[20] Yang, J., Xing, Q., Xu, M. and Qiao, M.: Uncover Treasures in DCT: Advancing JPEG Quality Enhancement by Exploiting Latent Correlations, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17598–17607 (2025).

[21] Yin, Y., Qi, H., Zhu, B., Chen, J., Jiang, Y.-G. and Ngo, C.-W.: Foodlmm: A versatile food assistant using large multimodal model, *IEEE Transactions on Multimedia* (2025).

[22] Yu, D., Min, W., Jiang, Q., Lin, X., Jin, X. and Jiang, S.: OmniFood8K: Single-Image Nutrition Estimation via Hierarchical Frequency-Aligned Fusion, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41562–41572 (2026).

[23] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H. and Shao, L.: Multi-stage progressive image restoration, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14821–14831 (2021).

[24] Zhang, K., Li, Y., Liang, J., Cao, J., Zhang, Y., Tang, H., Fan, D.-P., Timofte, R. and Gool, L. V.: Practical blind image denoising via Swin-Conv-UNet and data synthesis, *Machine Intelligence Research*, Vol. 20, No. 6, pp. 822–836 (2023).

[25] Zhang, T., Liu, P., Lu, Y., Cai, M., Zhang, Z., Zhang, Z. and Zhou, Q.: Cwnet: Causal wavelet network for low-light image enhancement, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8789–8799 (2025).

[26] Zhou, Y., Cao, J., Zhang, Z., Wen, F., Jiang, Y., Jia, J., Liu, X., Min, X. and Zhai, G.: Q-agent: quality-driven chain-of-thought image restoration agent through robust multimodal large language model, *arXiv preprint arXiv:2504.07148* (2025).

[27] Zhou, Y., Gao, X., Guo, J., Li, G., Wang, L. and Liu, J.: LUXFormer: Low-light image enhancement via joint spatial-frequency illumination modeling, *Journal of King Saud University Computer and Information Sciences*, Vol. 37, No. 8, p. 251 (2025).

[28] Zhu, K., Gu, J., You, Z., Qiao, Y. and Dong, C.: An intelligent agentic system for complex image restoration problems, *International Conference on Learning Representations*, Vol. 2025, pp. 57985–58013 (2025).