

# Evaluating the Gemini 2.5 Flash Model for Use in Dietary Monitoring

Anna-Margret Tamlin  
University of Waterloo  
Waterloo, ON, Canada  
atamlin@uwaterloo.ca

Yuhao Chen  
University of Waterloo  
Waterloo, ON, Canada  
yuhao.chen1@uwaterloo.ca

## Abstract

*Dietary monitoring is a complex yet highly impactful challenge within food computing, given its potential to transform the personalized management of metabolic and general health. Traditional 2D image-based assessment methods capture only static visual cues, offering limited information on eating behaviour. In contrast, video-based analysis provides greater temporal information, enabling the study of what is eaten, how it is eaten, and in what quantity. Prior research in this area introduced a baseline framework that used Vision–Language Models (VLMs) to analyze eating videos on a frame-by-frame basis. While this approach established the feasibility of using VLMs to interpret eating behaviours, it also revealed key limitations in model accuracy and contextual understanding. Building on this foundation, the present study evaluates the Gemini 2.5 Flash model relative to the previous framework across diverse eating scenarios and examines its performance with more specific prompts. These findings offer high-level insights into the potential of modern multimodal VLMs for dietary monitoring and pave the way for more accurate and practical approaches to video-based nutrient assessment.*

## 1. Introduction

Monitoring dietary intake remains a crucial aspect of healthcare and nutrition research, particularly for populations at increased risk of nutrition-related health issues, such as older adults [3]. Accurate monitoring is essential for preventing malnutrition, diabetes, and obesity, and for promoting healthy eating habits and nutritional well-being [3]. Achieving this goal requires reliable methods to identify foods consumed, estimate portion sizes, and evaluate nutritional composition.

Smartphones offer a convenient, relatively low-barrier platform for capturing high-quality recordings suitable for accurate analysis. During eating events, the user simply positions or holds the phone at a stable frontal angle during the meal, as seen in Fig. 1, so that both the face and the food

being lifted to the mouth remain visible. The setup, depicted in Fig. 2, records hand-to-mouth movements, chewing, and utensil use. Compared to staged, high-budget recordings and videos from unconventional viewpoints such as smart glasses, which can obscure what is being eaten, this approach enables flexible, unobtrusive data collection that captures natural eating behaviours



Figure 1. Illustration of a person holding their phone while eating, with the front camera capturing their eating behaviour.



Figure 2. Eating event recorded using the front-facing camera of a smartphone.

In a prior study, FoodVideoQA established a baseline

Table 1. Prompts and Corresponding Gemini 2.5 Flash Output

Food Item	Prompt	Output
7 Blackberries	<ol style="list-style-type: none"> <li>1. Identify the food(s) being eaten in this video. Please use the utensils as a good estimate of portion size and weight</li> <li>2. Please provide timestamps for each part of the video</li> <li>3. Provide a nutrition panel of food eaten during span of video (include the mass in g or volume in fl oz, calories, protein, carbs, fat) for the total food consumed</li> </ol>	<p>The food being eaten in the video is blackberries.</p> <ul style="list-style-type: none"> <li>• 0:00–0:02: Consumed 1 blackberry</li> <li>• 0:04–0:06: Consumed 1 blackberry</li> <li>• 0:09–0:11: Consumed 1 blackberry</li> <li>• 0:16–0:18: Consumed 1 blackberry</li> <li>• 0:21–0:23: Consumed 1 blackberry</li> <li>• 0:26–0:28: Consumed 1 blackberry</li> <li>• 0:31–0:34: Consumed 1 blackberry</li> </ul> <p>The blackberries appear to be of average to large size. An average-large blackberry weighs approximately 6g. <b>Total Blackberries Consumed: 7</b>  <b>Estimated Total Mass:</b> 6g/blackberry × 7 blackberries = 42g  Nutrition Panel for 42g of raw blackberries:</p> <ul style="list-style-type: none"> <li>• <b>Mass:</b> 42 g</li> <li>• <b>Calories:</b> 18 kcal</li> <li>• <b>Protein:</b> 0.6 g</li> <li>• <b>Carbohydrates:</b> 4.0 g</li> <li>• <b>Fat:</b> 0.2 g</li> </ul>

multimodal framework that integrates VLMs, pose estimation, and object detection to assess food intake [4]. VLMs such as LLaVA-v1, LLaVA-v2, and BLIP-2 identify and provide descriptions of visible food items and associated utensils [4]. Pose estimation models such as DW-Pose track landmarks around the mouth, and object detection networks like GroundingDINO localize food items with bounding boxes to associate actions with foods on a frame-by-frame basis [4]. Eating behaviour is inferred by estimating mouth openness and calculating the distance between the mouth and the nearest food item [4]. Frames are then classified as “eating” or “not eating,” and patterns spanning 20 frames are used to determine when eating occurs [4]. The labelled frames generate a running log, using a series of prompts passed to the VLMs, which reports the food being consumed, associated macronutrients, and utensils used for each interval in real time [4]. While this model provides an automated approach to nutritional tracking with minimal model-specific training, it struggles to generalize across diverse foods, utensils, and environments [4].

To address these limitations, this study evaluates the Gemini 2.5 Flash model, which supports multimodal inputs and advanced reasoning to interpret real-world eating scenarios and temporal patterns such as bite sequences and utensil use [2]. The experiments outlined assess the model’s robustness across varying lighting conditions, camera an-

gles, and diverse foods, and evaluates whether it improves upon the baseline in producing coherent meal-level summaries and more reliable dietary estimates. By examining the Gemini 2.5 Flash model, this study demonstrates how newer and more advanced multimodal approaches can overcome limitations of prior frameworks, improving food intake monitoring and enabling scalable, practical methods for automated nutrient assessment.

Google’s Gemini 2.5 Flash model was selected for this study due to its native multimodal reasoning capabilities and significantly lower latency and low-cost profile compared to models like GPT-4o. This makes it a more feasible candidate for the high-frequency, frame-based analysis required in real-time dietary monitoring.

## 2. Methodology

### 2.1. Data Acquisition and Prompting

The Gemini 2.5 Flash model was used to process eating videos of approximately one to two minutes in duration. Each recording was preprocessed to ensure consistent frame rates and resolutions suitable for model analysis. Initially, the model was prompted to identify the foods present and estimate total macronutrient content, providing a general dietary overview but not capturing the timing of individual bites.

During the study, a prompt was added to capture the timestamps of each bite and the corresponding foods consumed during that time period, enhancing the detail and accuracy of the dietary logs. This incremental refinement also allowed the video to be more easily segmented and verified, ensuring that the model’s outputs accurately reflected individual eating events. The final prompting approach combined food identification with temporal annotation, producing structured summaries suitable for further nutritional analysis, as shown in Table 1.

## 2.2. Performance Metrics and Baselines

Model performance was evaluated by comparing the Gemini 2.5 Flash output with the FoodVideoQA framework, using ground truth measurements of food intake and macronutrient content as a reference. Each component of the dishes consumed was weighed before and after filming to determine precise caloric and macronutrient values. Mean absolute percentage error (MAPE) was then calculated for both models’ caloric estimations:

$$\text{MAPE} (\%) = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \quad (1)$$

where  $y_i$  denotes the ground truth value,  $\hat{y}_i$  the model’s estimation, and  $n$  the number of observations. To prevent distortion of quantitative results, only foods that were successfully recognized and assigned macronutrient information by both frameworks, as denoted in Table 2, were included in the MAPE calculations.

While MAPE provides a quantitative comparison of caloric estimation accuracy between the two methods, qualitative comparisons were also performed to evaluate food identification, description specificity, and the clarity of reported intake logs.

Additionally, certain recordings filmed with a lower-resolution laptop webcam were marked with an asterisk (\*) in Table 2 to allow for an ablation analysis, enabling comparison of model performance across different camera angles and video quality.

## 3. Results and Discussion

Tables 2 and 3 report the results of the Gemini 2.5 Flash model compared with the baseline FoodVideoQA framework. Across the evaluated meals, the Gemini 2.5 Flash model more accurately detected eating events, identified foods consumed, and summarized nutrition intake. Gemini achieved a MAPE of 61% for calorie estimates, substantially lower than the 310% observed for the previous framework. The Gemini model failed to identify food in only one of the eighteen videos, whereas the FoodVideoQA framework failed in six.

Both the Gemini 2.5 Flash and FoodVideoQA models tended to report larger serving sizes for the oatmeal and

Table 2. Recognition Results for Gemini 2.5 Flash and FoodVideoQA Framework

Food Item	Gemini	FoodVideoQA
Okonomiyaki*	✓	×
Popcorn*	×	×
Greek yogurt*	✓	×
Mandarin slices	✓	✓
Gyoza	✓	✓
Salmon and rice	✓	✓
Tonkatsu ramen	✓	✓
Dark chocolate	✓	✓
Beef, rice, broccoli	✓	✓
Fried chicken and radish	✓	×
Oatmeal	✓	✓
Shrimp dim sum	✓	✓
Blackberries	✓	✓
Cornflakes	✓	✓
Wonton soup	✓	×
Beef stir fry	✓	✓
Apple slices	✓	✓
Rice Pudding	✓	×

Table 3. Caloric Percent Error (PE) Comparison Between Gemini 2.5 Flash and FoodVideoQA Framework

Food Item	PE (%)	
	Gemini	FoodVideoQA
Mandarin slices	20.0	730.0
Gyoza	110.0	330.0
Salmon and rice	78.0	150.0
Tonkatsu ramen	9.6	130.0
Dark chocolate	43.0	200.0
Beef, rice, broccoli	41.0	50.0
Oatmeal	120.0	440.0
Shrimp Dim Sum	22.0	54.0
Cornflakes	190.0	650.0
Blackberries	13.0	530.0
Beef stir fry	25.0	260.0
Apple slices	56.0	230.0
<b>MAPE (%)</b>	<b>61.0</b>	<b>310.0</b>

cornflakes, likely due to the use of smaller-than-average spoons during consumption. This suggests that the frameworks assumed a standard-sized spoon, inflating the estimated volume per bite. Despite this shared limitation, Gemini’s estimates were closer to the actual values, resulting in substantially smaller errors (Table 3).

Qualitative comparisons further highlighted Gemini’s advantages. Table 4 depicts notable differences in outputs

Table 4. Qualitative Output Comparison: FoodVideoQA vs. Gemini 2.5 Flash

Model	Generated Output Response
<b>Base Framework (FoodVideoQA)</b>	<ul style="list-style-type: none"> <li>- processing frame100.jpg...</li> <li>Q: Identify only the food items visible</li> <li>A: noodles</li> <li>Q: Provide a list of cutlery/utensils</li> <li>A: ['chopstick']</li> <li>Q: Provide nutritional value (calories, ...)</li> <li>- Calories = 300</li> <li>- Fats = 10%</li> <li>- Protein = 15%</li> <li>- Carbohydrates = 55%</li> </ul>
<b>Gemini 2.5 Flash</b>	<ol style="list-style-type: none"> <li><b>1. Food(s) Identified:</b> Tonkatsu ramen with noodles, broth, and a soft-boiled egg.</li> <li><b>2. Timestamps &amp; Intake:</b> <b>Noodles (10g):</b> 0:00-0:15, 0:22-0:28, 0:44-0:58, 1:03-1:16; <b>Broth (7.5ml):</b> 0:32-0:41, 1:18-1:20; <b>Soft-boiled Egg (5g):</b> 0:48-0:51.</li> <li><b>3. Nutrition Panel (Total):</b> Ramen Noodles (~80g), Tonkatsu Broth (~30ml), Soft-boiled Egg (~5g). <b>Total:</b> 143 kcal (P: 4.4g — C: 26.1g — F: 2.8g).</li> </ol>

generated by the two models. The previous framework often labelled foods broadly, for instance, identifying the ramen video as “noodles.” Gemini was more precise, identifying individual components in greater detail: “tonkatsu ramen with noodles, broth, and a soft-boiled egg.” The previous model also reported nutritional values for a single frame that were inconsistent or implausible, often showing hundreds of calories or macronutrient percentages that did not sum to 100%. In contrast, Gemini provided macronutrient values consistent with the total calorie estimates [1].

The Gemini 2.5 Flash model more accurately identified which frames contained eating events and then aggregated this information to summarize intake timestamps over each dish, as shown in Table 1. Although the previous framework tracked eating on a frame-by-frame basis, interpreting the resulting log entries required manual aggregation of hundreds of frames, making it difficult to determine total consumption or the exact timing of individual bites, and providing no clear breakdown of when specific foods were eaten. With the addition of a prompt to identify the timing of each bite and the corresponding food consumed, Gemini generated summaries that were well-structured and comprehensive. This enhancement improved the usability and interpretability of the meal summaries.

As shown in Table 2, device and filming conditions significantly influenced performance for both models. Higher-resolution smartphone recordings, captured vertically to include both the food and face, generally resulted in lower prediction errors. In contrast, lower-resolution laptop webcam recordings (marked with an asterisk), captured horizontally and often cutting off parts of the food or face, as in

Fig. 3, reduced accuracy for both models to varying degrees. While the food items were not identical across devices, this preliminary cross-device feasibility check assesses the models’ sensitivity to camera angle and video quality, a critical factor for real-world deployment where standardized hardware cannot be guaranteed. Notably, while the baseline framework failed to recognize any foods in the webcam videos, Gemini remained generally capable of identification, albeit with lower accuracy than in smartphone-captured trials.



Figure 3. Eating event captured with a laptop webcam, showing a wide aspect ratio that cuts off part of the yogurt and face.

## 4. Conclusion

This study establishes Gemini 2.5 Flash as a more practical foundation for automated dietary monitoring than prior VLM-based frameworks. By enabling accurate food identification and the coherent aggregation of temporal data, it overcomes the fragmented, frame-by-frame limitations of the FoodVideoQA baseline. While portion estimation remains sensitive to utensil size, Gemini’s improved reasoning under suboptimal lighting and varied hardware makes it a viable candidate for scalable, real-world nutrient assessment. This study underscores the potential for modern multimodal VLMs to provide consistent, automated nutritional insights across diverse eating environments.

### 4.1. Limitations

Accurate portion size estimation remains a primary challenge. Both models tend to overestimate servings when small utensils are used, indicating a reliance on standard size assumptions. Beyond utensil bias, failure modes, such as the missed recognition of popcorn, suggest challenges with high-texture items and motion blur during consumption. Future research should integrate contextual cues (e.g., utensil dimensions and camera distance) and frame-level binary classification (eating versus not eating) to further refine caloric and macronutrient scaling.

## References

- [1] Iain Campbell. Macronutrients, minerals, vitamins and energy. *Anaesthesia & Intensive Care Medicine*, 18(3):141–146, 2017. 4

- [2] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind, 2025. [2](#)
- [3] Wilma Leslie and Catherine Hankey. Aging, nutritional status and health. *Healthcare*, 3(3):648–658, 2015. [1](#)
- [4] Krish Shah, Siddharth Viswanath, Pengcheng Xi, Alexander Wong, and Yuhao Chen. Foodvideoqa: A novel baseline framework for dietary monitoring. In *Proceedings of Conference on CVPRW 2025*, Waterloo, ON, Canada; Ottawa, ON, Canada, 2025. University of Waterloo; National Research Council Canada. [2](#)