Supplementary Material
Multi-modal Large Language Models are Effective Vision Learners

A. Additional Results of Token-Level Visualization

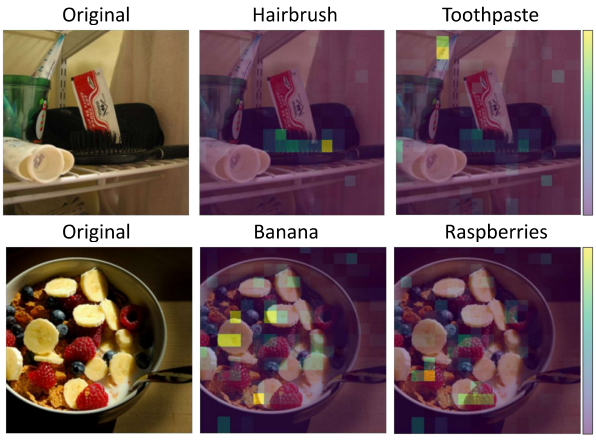

Figure 1. Additional results of token-level visualization. We overlay a sample image with attention maps between selected text tokens in generated response and vision tokens.

B. Full LLM-Generated Response for Example Images

For the example in Fig.1 in the main manuscript, the generated response is *The image features a dining table with two white plates of food, each containing a serving of broccoli. The table also has two forks and two knives placed beside the plates. Additionally, there is a piece of bread on the table, and a cup is located near the top left corner of the table*. For the example in Fig.4, the generated response is *The image displays four plastic containers filled with various types of food, placed on a table. The containers hold a variety of items, including meat, vegetables, fruit, and bread. Some of the specific items include broccoli, which can be seen in multiple containers, and oranges, which are present in one of the containers*.