Interleaved-Modal Chain-of-Thought

Supplementary Material

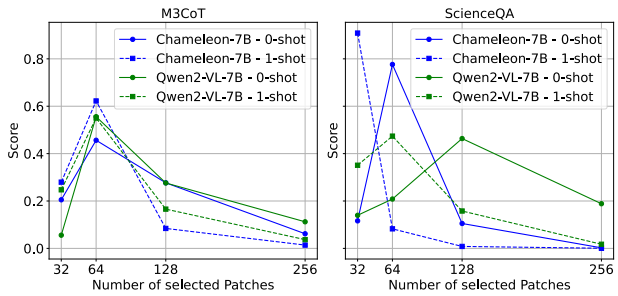

Figure 4. The results of ICoT across validation sets of two datasets on both Chameleon and Qwen2-VL, with the number of selected patches set to 32, 64, 128, and 256. The reported scores are normalized for simplicity.

8. Analysis on the Selected Patches

Intuitively, the performance of ICoT is sensitive to the number of selected patches. If ADS selects a large number of patches every time, the selected patches will be dispersed, resulting in more noise introduced and higher computation costs. In contrast, only a few selected patches perhaps failed to contain enough fine-grained visual information. It is non-trivial to determine the exact number of patches selected by ADS, as fine-grained information in an image is not always the same size. Therefore, in Figure 4, we empirically set the number of patches selected by ADS n to 32, 64, 128, and 256 at a coarse-grained level and illustrate their performance variance across two benchmarks [3]. Observed results indicate that setting n too large or too small is not good for VLMs, and ICoT achieves relatively better performance when n is set to 64.

9. Performance on General Benchmark

1-shot	Flickr30k (CIDEr ↑)	OKVQA (VQA-ACC ↑)
Chameleon	22.3	26.2
+ICoT	23.6	28.2

Table 5. Evaluation on general benchmarks

ICoT is a plug-and-play prompting method designed for complex multimodal reasoning, while the performance of ICoT on tasks requiring weak reasoning ability is still unknown. To explore whether ICoT causes degradation, we evaluate ICoT on captioning and VQA in Tab.5. Results indicate the advantages of ICoT.

[3] LLaVA-W only contains a test set.

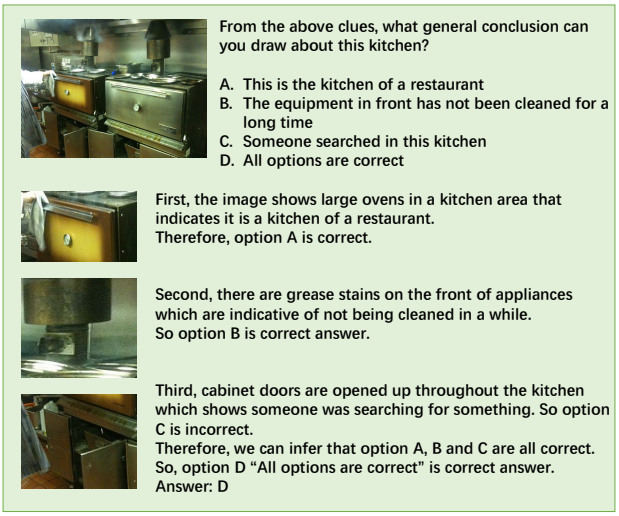

Figure 5. The case of demonstration with Fine-grained Visual Information (FVI), which is used in 1-shot ICoT.

10. Detail Declaration

In Fig. 5, we provide a case to illustrate the FVI in 1-shot ICoT. In Algorithm 1, the stopping criterion is the maximum generation length or generating the special token of "end of sequence".