

Figure 1: (a) The correlation between discretization error difference (DED) and the quantization errors in 15th layer. (b) The correlation between DED and the entropy in 5th layer and (c) in 25th layer.

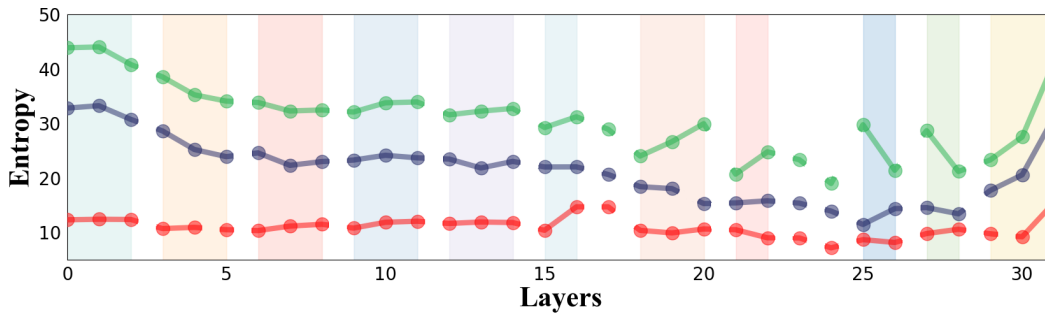



Figure 2: Visualization of block division for Q-VLM with different samples on LLaVA architectures.



**Q INT4 AWQ**  
Animals get their food by digesting other organisms. **But in the 1950s, scientists discovered that animals can make their own food.** Fromia monilis cells use chemosynthesis to make their food. **The answer is A.**

---

**Q INT4 Q-VLM**  
Today, many scientists classify organisms into six broad groups, called kingdoms. The table below shows some traits used to describe each kingdom. **Fromia monilis is an animal.** Animal cells cannot make their own food. **The answer is B.**

|                                | Bacteria         | Archaea          | Protists         | Fungi       | Animals | Plants |
|--------------------------------|------------------|------------------|------------------|-------------|---------|--------|
| How many cells do they have?   | one              | one              | one or many      | one or many | many    | many   |
| Do their cells have a nucleus? | no               | no               | yes              | yes         | yes     | yes    |
| Can their cells make food?     | some species can | some species can | some species can | no          | no      | yes    |

Figure 3: Visual reasoning examples from LLaVA-13B model. Q-VLM improves over the AWQ baseline for W4A4 quantization, reducing quantization errors and providing more reasonable answers. We color the text to show the correct or wrong responses.