

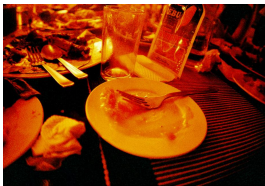
Supplementary Materials: Advancing Multimodal Large Language Models with Quantization-Aware Scale Learning for Efficient Adaptation

Anonymous Authors

1 MORE QUALITATIVE RESULTS

To explore the potential of QSLAW, Figure. 1 presents more qualitative results in multi-turn conversations. We compare QSLAW (left) with QLoRA (right) and it can be observed that QSLAW can provide a more detailed and correct response about different question form the same image. For example, QSLAW can describe the fork and

knife in the image which are overlooked by QLoRA. Furthermore, while both QSLAW and QLoRA can correctly identify the number of airplanes, QLoRA mistakenly interprets them as jet aircraft, whereas QSLAW correctly identifies them as commercial airplanes. We also provide some typical failure cases of QSLAW. In Figure. 2, we can observe that the angle from which a picture is captured



➤ Q1: What are the main objects on the table in the image?

The main objects on the table in the image are a plate of food, a glass of wine, a fork, and a knife.

➤ Q2: Describe the following image.

The image shows a close-up of a dining table with a plate of food, a fork, and a glass of wine. The plate has a piece of food on it, and the fork is placed next to the plate. The table is set for a meal, with the wine glass and the fork indicating that someone is about to enjoy their meal.

➤ Q3: What activity might have recently taken place around this table and what could be its aftermath?

A recent activity that might have taken place around this table is a meal, as evidenced by the presence of plates, cups, and utensils such as forks and knives. The aftermath of this activity could include cleaning up the table, disposing of any leftover food, and putting away the used utensils and cups. The presence of a wine glass and a bottle suggests that the meal might have been accompanied by wine, which could have been enjoyed during the meal. The aftermath of this activity would involve cleaning up the table, putting away the used utensils, and disposing of any leftover food.

➤ Q1: What are the main objects on the table in the image?

The main objects on the table in the image are plates of food, forks, and glasses.

➤ Q2: Describe the following image.

The image shows a table with a plate of food on it, including a piece of cake. There are also several glasses and utensils, such as forks and knives, placed on the table. The scene appears to be set in a restaurant or a dining area, and the lighting is dim, creating a cozy atmosphere.

➤ Q3: What activity might have recently taken place around this table and what could be its aftermath?

Given the presence of a plate of food, a fork, a knife, and a glass on the table, it is likely that someone has recently enjoyed a meal at this table. The aftermath of this activity could be the cleanup process, which may involve clearing the table of the used utensils and dishes, washing them, and putting them away. Additionally, the table may need to be wiped down and any remaining food scraps or crumbs removed. The cleanup process is crucial to maintaining a clean and hygienic environment, ensuring the well-being of the people using the space and preventing any potential health hazards.



➤ Q1: How many airplanes are visible in the image?

There are two airplanes visible in the image.

➤ Q2: What do you see happening in this image?

In the image, a large commercial airplane is flying over a body of water, with a group of birds flying nearby. The airplane is flying low, and it appears to be landing or taking off from an airport.

➤ Q3: What environmental impact might these planes' activity have on the birds?

The presence of several birds flying near the airplanes in the image suggests that the planes might be disturbing their natural habitat or flight patterns. Airplanes produce noise, vibrations, and emissions that can disrupt the behavior of birds and other wildlife. The noise from the planes can cause stress in birds, making them more susceptible to predators or other threats. Additionally, the emissions from the planes can contribute to air pollution, which can negatively affect the health and well-being of the birds and other living organisms in the area. To minimize the environmental impact, airports and airlines should implement measures such as noise reduction technologies, emissions control, and habitat preservation to minimize the disruption to the local ecosystem.

➤ Q1: How many airplanes are visible in the image?

There are two airplanes visible in the image.

➤ Q2: What do you see happening in this image?

In this image, a large jet airplane is flying low over the water, with a body of water in the background. The airplane appears to be landing or taking off, as it is either coming in for a landing or preparing to take off. There are also several birds flying near the airplane, adding to the dynamic scene.

➤ Q3: What environmental impact might these planes' activity have on the birds?

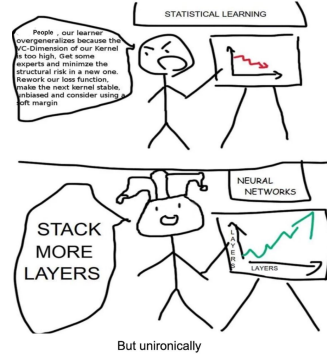
The presence of large airplanes flying over the water and the birds in the area might have several environmental impacts on the birds. Firstly, the noise generated by the airplanes can cause stress and disturbance to the birds, potentially affecting their breeding, feeding, and social behavior. Secondly, the airplanes' exhaust fumes and emissions can contribute to air pollution, which can negatively impact the birds' respiratory health and overall well-being. Lastly, the airplanes' activity might disrupt the birds' natural habitat, causing them to flee or alter their migration patterns. It is essential for aviation authorities and environmental organizations to consider these factors and implement measures to minimize the impact of airplane activity on wildlife and their habitats.

Figure 1: More comparison examples between QSLAW and QLoRA in multi-turn conversations. Left is the results generated by QSLAW.



► Analyze the image in a comprehensive and detailed manner.

The image features a vintage suitcase with a brown leather exterior and a brown and white striped pattern. The suitcase is open, revealing its interior with a cardboard tag attached to it. The tag has a message written on it, possibly indicating the suitcase's origin or destination. The suitcase is placed on a table, and there are two other suitcases in the scene, one located to the left and the other to the right of the main suitcase. The arrangement of the suitcases suggests that they might be displayed for sale or showcasing.



► What is funny about this image? Describe it panel by panel.

The humor in this image comes from the juxtaposition of the man's facial expression and the content displayed on the two monitors. The man is making a face, possibly expressing frustration or annoyance, while the monitors show graphs and charts, which are typically associated with positive or informative content. This contrast creates a humorous situation where the man's emotions seem to be at odds with the content displayed on the screens. The viewer might find this amusing because it highlights the disconnect between the man's emotions and the context of the information being presented.

Figure 2: Failure cases of QSLAW.

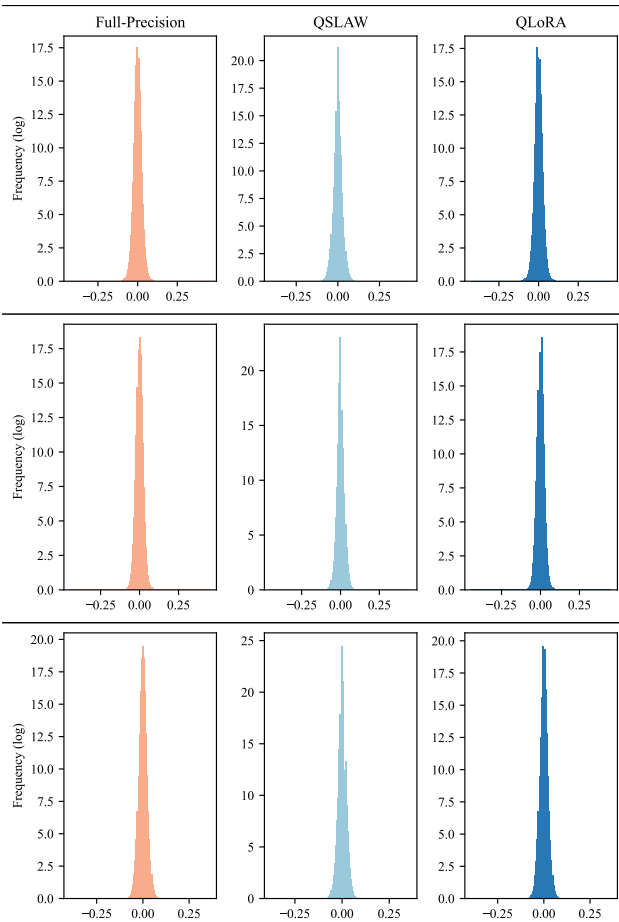


Figure 3: The weight distributions of 10th, 20th, and 30th layers are presented from top to bottom.

would influence the model's understanding and description of the image. This challenge also occurs in existing multimodal large language models. Meanwhile, QSLAW may focus on certain prominent yet non-essential details within the image like facial expression in the meme and provide an inaccurate explanation, which did not help in understanding the humor in the meme.

2 MORE VISUALIZATION

In this section, we provide some visualizations to gain deeper insights for QSLAW. We present visualizations of the weight distribution of the linear layer in Figure 3. NF4 of QLoRA equalize the quantity of values across all quantization bins from information-theoretically optimal to preserve original weight information, thus exhibiting a closer resemblance to the distribution of full-precision weights. However, LoRA struggles to effectively mitigate quantization errors caused by activation outliers which is shown in Figure 4, as it conducts fine-tuning in a coarse-grained, global manner and QSLAW can be effectively minimized quantization error towards downstream tasks by adjusting each weight group. We calculate quantization error according to $\|X - X_q\|$ where X and X_q represents output activation with full precision and quantized weight, respectively.

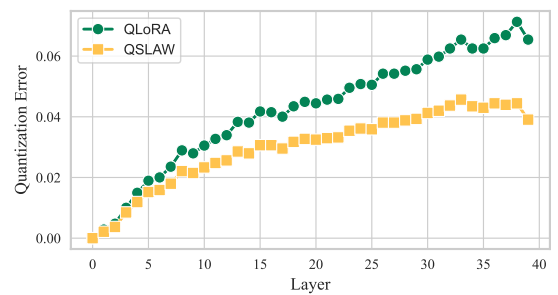


Figure 4: Quantization error of QLoRA and QSLAW.