# LLaVA-UHD: an LMM Perceiving any Aspect Ratio and High-Resolution Images

Zonghao Guo[3], Ruyi Xu[1], Yuan Yao[2⋆], Junbo Cui[1], Zanlin Ni[1], Chunjiang Ge[1], Tat-Seng Chua[2], Zhiyuan Liu[1], and Gao Huang[1⋆]

[1] Tsinghua University
[2] National University of Singapore
[3] University of Chinese Academy of Sciences
guozonghao96@outlook.com, yaoyuanthu@gmail.com

## A Additional Experiments

In this section, we provide additional experimental results of illustration on GPT-4V phases in processing images with changing resolutions.

**Detailed Illustration on GPT-4V Phases.** From the pilot experimental results in Fig. 1, we observe that the GPT-4V responses show a significant phase change with image resolutions. Here we provide detailed illustrations of the cause (best guess) from the perspective of visual encoding:

(1) In phase 1, since there is only one image slice, most answers are correct. More specifically, when dealing with input images under 512×512, if the images are resized to 512×512, the behavior will be the same within phase 1. However, since the behavior changes significantly within phase 1, we suspect that the input images are most likely to be padded into512×512, as shown in Fig. 2(a).

(2) In phase 2, answer 12 dominates the responses possibly due to the incomplete circles in each slice, as shown in Fig. 2(b).

(3) Phase 3 shows mixed answers of 9, 12 and 16. Among these responses, answer 16 can be well explained by the slice strategy in Fig. 2(c). Besides, we also notice that many abnormal phenomena in Fig. 1(b) cannot be perfectly explained yet, which we leave for future work.

## B Proofs

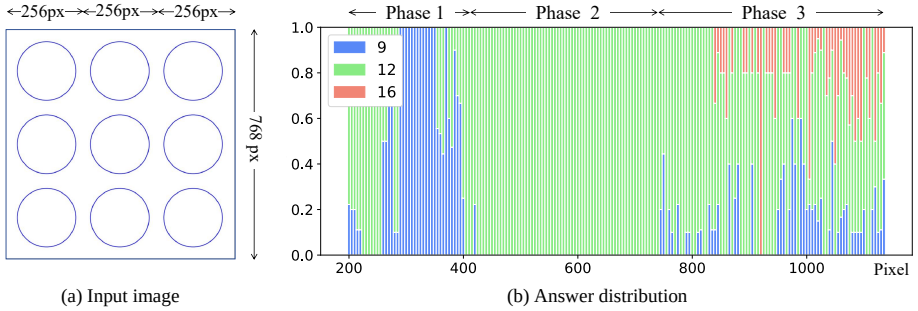
In this section, we provide proofs for the image partition strategy. We show that the slice resolution exhibits modest changes to the original resolution of ViT.

**Range of Slice Aspect Ratios.** The aspect ratio of the slice can be represented by:

$$\frac{W_v}{H_v} = \frac{W_I}{m} : \frac{H_I}{n},$$

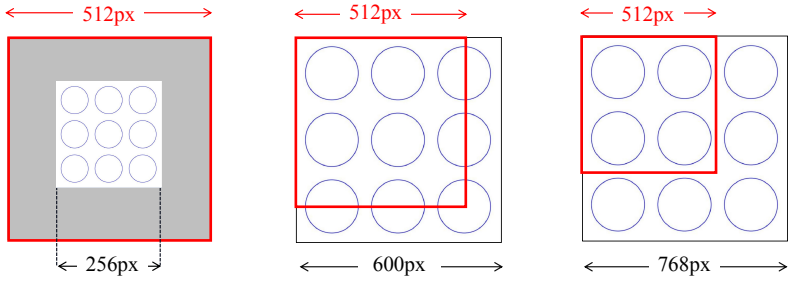
where $W_v$, $H_v$ are the width and height of the slice, $W_I$, $H_I$ are the sizes of the original image, and (m, n) is the best partition. Restricting the aspect ratio

⋆ Corresponding Authors




(a) Input image (b) Answer distribution

**Fig. 1:** Results on probing GPT-4V via continuously changing image resolutions.




Phase 1 (1 slice) Phase 2 (4 slices) Phase 3 (4 slices)

**Fig. 2:** Illustration on GPT-4V phases (best guess). Red square indicates a slice.

$r = \frac{W_v}{H_v} \in [\frac{1}{2}, 2]$ is equivalent to $|\log(\mathrm{r})| \leq |\log 2|$, which is also equivalent to $\left|\log\left(\frac{W_I}{H_I}\right) - \log(\frac{n}{m})\right| \leq |\log(2)|$. We need to prove:

$$\forall \frac{W_I}{H_I} \in [\frac{1}{6}, 6], N < 20$$

$$\exists (\mathrm{m},\, \mathrm{n}) \in \bar{\mathbb{C}}, \left|\log\left(\frac{W_I}{H_I}\right) - \log(\frac{n}{m})\right| \leq |\log(2)|,$$

which is equivalent to

$$\forall N < 20, (n_i, m_i) \in \bar{\mathbb{C}}$$

$$\exists (n_j, m_j) \in \bar{\mathbb{C}}, \left|\left(\log\left(\frac{n_i}{m_i}\right) - \log\left(\frac{n_j}{m_j}\right)\right)\right| \leq 2 \cdot |\log(2)|\,,$$

which can be verified by enumerating all possible factorizations of $\bar{\mathbb{C}} = \mathbb{C}_{N-1} \cup \mathbb{C}_N \cup \mathbb{C}_{N+1}$ for $N < 20$. The results show that the aspect ratio of each slice resides within $[\frac{1}{2}, 2]$.

**Expected Aspect Ratio.** We assume that the ratio of the original image is greater than 1 (i.e., $H_I > W_I$). The situation is the same for $H_I < W_I$. Assuming

that the sizes of the images are uniformly distributed for $N \in [0, 20]$, while the aspect ratio of the original images $\frac{W_I}{H_I} \in [1, 6]$, we have $P(W_I, W_H, n, m) = \frac{1}{20} \cdot \frac{1}{5}$. The expected aspect ratio can be obtained by:

$$\mathrm{E}(\frac{m \times W_I}{n \times H_I}) = \iint\limits_{\substack{\frac{W_I}{H_I} \in [1,6] \\ W_I \cdot H_I \in [0, 20s] \\ n, m = \arg\max S(\cdot)}} (\frac{m \times W_I}{n \times H_I}) \cdot P(W_I, H_I, n, m)\ dW_I dH_I,$$

where $s$ is the area of a standard resolution of ViT. After calculation, we obtain $\mathrm{E}(r) = 1.258$, $\mathrm{Var}(r) = 0.048$. The results show that the expected aspect ratio of the slices is 1:1.258, which is close to the standard pertaining setting of ViT. More commonly assuming that images are uniformly distributed between $[1, 3]$, and the aspect ratio is uniformly distributed between $[1, 2]$, we have $\mathrm{E}(r) = 1.147$, $\mathrm{Var}(r) = 0.011$, indicating even smaller changes.

**Range of Slice Area.** Let $n = \frac{W_I}{W_v} \times \frac{H_I}{H_v}$, which leads to $N = \lceil n \rceil$. We consider dividing the image into $\{N-1, N, N+1\}$ slices. Therefore, the maximum value of each slice $\mathrm{S_{max}} = \frac{n}{N-1}$ (when $N \neq 2$), and $\mathrm{S_{max}} = \frac{n}{N}$ (when $N = 2$). The minimum value $\mathrm{S_{min}} = \frac{n}{N+1}$. As $n$ approaches $3^-$, where $N = 3$, $\mathrm{S_{max}}$ achieves the maximum value of 1.5. Similarly, as $n$ approaches $1^+$, where $N = 2$, $\mathrm{S_{min}}$ achieves the minimum value of 0.33.

**Expected Slice Area.** Still assuming that the sizes of the images are uniformly distributed within $N \in [0, 20]$, while the aspect ratio of the images $\frac{W_I}{H_I} \in [\frac{1}{6}, 6]$. The expected area of slice can be obtained by:

$$\mathrm{E}(\frac{W_I \times H_I}{n \times m}) = \iint\limits_{\substack{\frac{W_I}{H_I} \in [1,6] \\ W_I \cdot H_I \in [0, 20s] \\ n, m = \arg\max S(\cdot)}} (\frac{W_I \times H_I}{n \times m}) \cdot P(W_I, H_I, n, m) dW_I dH_I.$$

After calculation, we obtain $\mathrm{E}(\frac{W_I \times H_I}{n \times m}) = 1.057$, $\mathrm{Var}(\frac{W_I \times H_I}{n \times m}) = 0.016$. This shows that our slice areas are relatively concentrated, similar to the original resolution of ViT.

## C Discussions

We provide discussions on limitations and potential negative impact of this work.

**Limitations and Future Work.** (1) Higher resolutions. In this work, we limit the resolution of LLaVA-UHD to maximum $672 \times 1008$. Although this resolution increases the standard LLaVA-1.5 resolution by 6 times, higher-resolution images such as 4K images and remote sensing images are still out of reach. In future, considering the promising efficiency and scalability, we will explore higher-resolution images and more challenging tasks such as small object detection and segmentation. (2) Joint slice encoding. Currently image slices are

independently encoded, with interactions only in LLMs. We plan to establish efficient connections between image slices via improved visual encoding strategies for fine-grained global information interaction.

**Potential Negative Impact.** In this work, we investigate the failure pattern and the underlying cause for GPT-4V and LLaVA-1.5. The mechanism can be potentially used for adversarial attacks on these models. It is worth noting that the goal of this work is to raise attention to the vulnerability of LMMs and provide a deeper understanding of the importance of visual encoding strategies. This work calls for further efforts to mitigate the revealed issues to ensure the robustness and safety of LMMs.