ID-Align: RoPE-Conscious Position Remapping for Dynamic High-Resolution Adaptation in Vision-Language Models

Anonymous ACL submission

Abstract

Currently, a prevalent approach for enhancing Vision-Language Models (VLMs) performance is to encode both the high-resolution version and the thumbnail of an image simultaneously. While effective, this method generates a large number of image tokens. When combined with the widely used Rotary Position Embedding (RoPE), its long-term decay property hinders the interaction between high-resolution tokens and thumbnail tokens, as well as between text and image. To address these issues, we propose **ID-Align**, which alleviates these problems by reordering position IDs. In this method, high-resolution tokens inherit IDs from their corresponding thumbnail token while constraining the overexpansion of positional indices. Our experiments conducted within the LLaVA-Next framework demonstrate that ID-Align achieves significant improvements, including a 6.09% enhancement on MMBench's relation reasoning tasks and notable gains across multiple benchmarks.

1 Introduction

The swift advancement in large language models (LLMs) (Achiam et al., 2023; Cai et al., 2024; Yang et al., 2024; Liu et al., 2024a) has not only revolutionized natural language processing but also catalyzed the emergence of vision-language models (VLMs) (Liu et al., 2024d; Wu et al., 2024; Chen et al., 2024d; Li et al., 2023a; Wang et al., 2024). In the architecture of these advanced VLMs, visual encoders—such as Vision Transformers (ViTs) (Dosovitskiy, 2020) employing training objectives like CLIP (Radford et al., 2021) or SigLip (Zhai et al., 2023)—are primarily utilized to encode images. Subsequently, mechanisms such as Multi-Layer Perceptrons (MLPs) (Liu et al., 2024d) or Q-Former (Li et al., 2023a) are employed to fuse the encoded visual information with textual data. This fused multimodal information is then processed by the LLM, enabling comprehensive understanding and contextually relevant response generation across both visual and textual domains (Yin et al., 2023).

In the pursuit of developing more effective VLMs, researchers are undertaking multifaceted efforts, including curating higher-quality training datasets (Bai et al., 2024) and refining model architectures (Cha et al., 2024). Beyond these strategies, another approach explored to enhance model performance involves upscaling an input image to a higher resolution before encoding, while concurrently processing a low-resolution version as a thumbnail (Dai et al., 2024; Deitke et al., 2024; Wu et al., 2024; Chen et al., 2024c; Liu et al., 2024b). The image tokens derived from both the thumbnail and the high-resolution image are then concatenated and fed into the LLM. This technique is commonly referred to as dynamic high-resolution adaptation.

Despite its straightforwardness and effectiveness, this dynamic high-resolution adaptation method exhibits several critical shortcomings. Encoding high-resolution images inherently generates a large number of image tokens. Consequently, the application of Rotary Position Embedding (RoPE) (Su et al., 2024), a prevalent position encoding method, can pose specific challenges due to its characteristic **long-term decay property**, which posits that attention scores between query and key diminish as their relative distance increases. Although generally assumed to be valid, some researchers have contested this property (Barbero et al., 2024). Our further analysis reveals that, based purely on RoPE's mathematical formulation, its effective behavior (e.g., long-term decay, growth, or more complex patterns) can vary depending on the specific distributions of the query ($\mathbf{q}$) and key ($\mathbf{k}$) vectors. Furthermore, our empirical experiments confirm that, under the actual distributions of $\mathbf{q}$ and $\mathbf{k}$ observed in LLMs, RoPE indeed exhibits this long-term decay property.

This property may lead to:

- **Hinders image-text interaction:** The substantial increase in image embeddings resulting from high-resolution strategies can impede effective interaction between text and image embeddings. This issue is particularly pronounced for image embeddings whose sequential positions are distant from the text embeddings.
- **Loss of Multi-Resolution Correspondence:** A spatial correspondence should exist between high-resolution image tokens and their thumbnail counterparts, where two tokens are defined as corresponding if their encoded regions spatially overlap. However, RoPE's long-term decay property can disrupt this crucial relationship.

To address these issues, we propose **ID-Align**, a novel method that strategically rearranges the position IDs of image tokens. By assigning identical positional IDs to corresponding high-resolution and thumbnail image embeddings, ID-Align preserves their inter-resolution correspondence. This approach not only maintains the crucial relationship between high-resolution and thumbnail tokens but also mitigates the excessive inflation of position ID magnitudes that can arise from the large number of image embeddings in high-resolution strategies. Our experiments, conducted on the LLaVA-Next (Liu et al., 2024c) architecture, demonstrate that ID-Align significantly enhances model capabilities, particularly concerning fine-grained perception of global information. Our contributions can be summarized into the following two points:

- We analyze the mathematical properties of RoPE, demonstrating that its long-term decay property is contingent upon the specific distributions of $\mathbf{q}$ and $\mathbf{k}$ vectors. We further conduct empirical experiments showing that within LLMs, RoPE indeed imparts this long-term decay property to the model's attention mechanism.
- We first analyze the adverse effects of the long-term decay property of RoPE when increasing the number of image embeddings using the aforementioned super-resolution methods.
- On this basis, we introduce **ID-Align**, a technique for reorganizing position IDs. This method is aimed at maintaining the correspondence between image embeddings across different resolutions and mitigating the excessive growth of position IDs caused by dynamic adjustments to higher resolutions. Our experiments on the architecture and datasets of LLaVA-Next confirm the effectiveness of ID-Align.

2 Background & Related Work

2.1 Vision Language Model

Currently, the mainstream approach to build VLMs is to employ a projector to connect a pre-trained LLM with a visual encoder, thereby enabling the LLM to interpret visual information (Zhang et al., 2024a). For image inputs I_{image}, it is usual to first encode them using vision encoders such as SigLIP (Zhai et al., 2023) or CLIP (Radford et al., 2021) ViT (Dosovitskiy, 2020):

$$F_{image} = VE(I_{image}) \tag{1}$$

Subsequently, the projector processes the encoded image features F_{image}:

$$P_{image} = Projector(F_{image}, I_{text}) \tag{2}$$

where I_{text} represents the text input. In certain architectures, such as BLIP-2 (Li et al., 2023a), I_{text} also interacts with F_{image} at this stage. Following this, the LLM backbone processes I_{text} alongside P_{image}, generating the corresponding output:

$$Output = LLM(I_{text}, P_{image}) \tag{3}$$

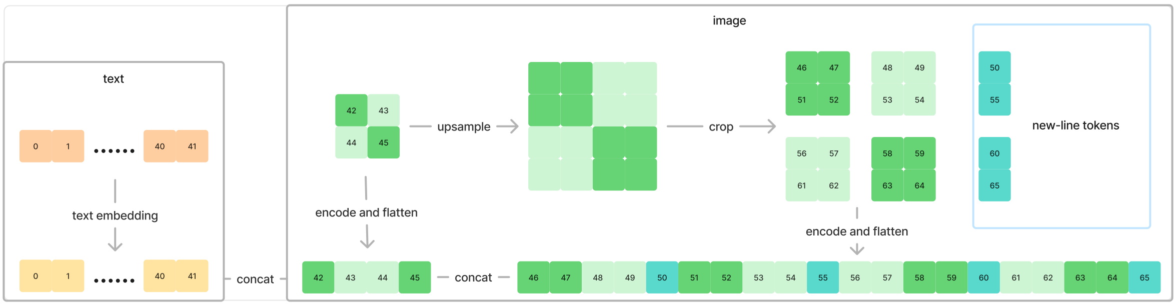
The architecture of the projector has many possible designs, and currently, a mainstream choice is to use a two-layer Multilayer Perceptron (MLP) to process F_{image} independently of I_{text}, as exemplified by the LLaVA architecture (Liu et al., 2024d):

$$P_{image} = MLP(F_{image}) \tag{4}$$

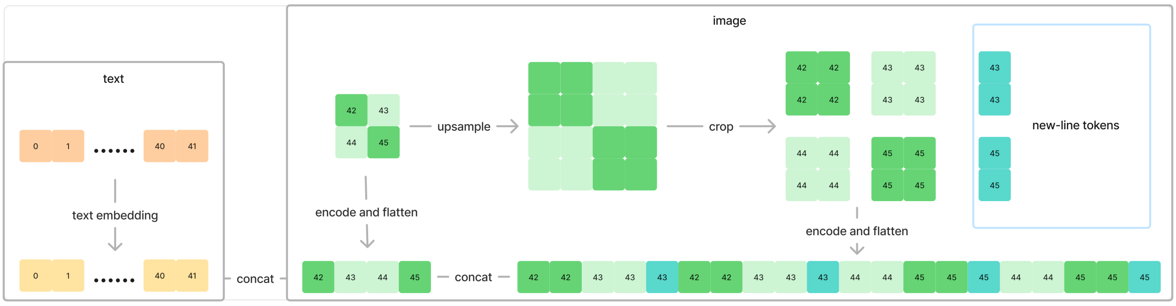
2.2 Dynamic High-resolution

While VLMs exhibit remarkable performance across diverse domains, they possess inherent limitations. These are sometimes characterized using the phrase 'VLMs are blind' (Rahmanzadehgervi et al., 2024), denoting their deficiencies in areas such as fine-grained perception and spatial understanding. One effective method is the dynamic high-resolution approach, the process of which is illustrated in Figure 2 and includes the following steps:

The current mainstream pipeline is as follows:



(a) the original method



(b) ID-Align

Figure 1: Intuitive presentation of the original high-resolution method and ID-Align.

- Set a Predefined Set of Resolutions. For instance, if the ViT used in a VLM is suitable for processing images of size (336, 336), this set of resolutions could be defined as [(672, 672), (336, 672), (672, 336), (1008, 336), (336, 1008)].
- Select Appropriate Resolution. Given an input image with dimensions (H_0, H_0), the most suitable resolution is selected from a set of predefined resolutions based on its aspect ratio.
- Adjust Input Image Resolution. For an input image with original resolution (H_0, W_0), two resolution adjustments are applied: first, super-resolving it from its original resolution to a selected higher resolution (H_h, W_h) to obtain a high-resolution image; and second, resizing it to a resolution (H_l, W_l) suitable for the ViT to serve as a thumbnail. The former process often preserves the original image's aspect ratio,filling the remaining regions with blank space, while the latter generally does not.
- Encode Image. ViT is used to encode the high-resolution image and its thumbnail separately. For the encoded features of the high-resolution image, an unpadding stage is typically required to remove the features corresponding to the padding regions. The resulting encoded features are then concatenated to obtain the final encoding.

This method is used by various leading VLMs (Zhu et al., 2025; Liu et al., 2024f; Deitke et al., 2024; Wu et al., 2024; Chen et al., 2024c; Liu et al., 2024b). When VLMs use a fixed-size ViT for encoding, to handle high-resolution images, the common approach is to divide the high-resolution image into patches or crops, encode each separately, and then rearrange the encoded results. Tokens, such as new-line tokens or separators, are also typically added at appropriate positions. This process can be seen in Figure 1a.

2.3 RoPE

The sequential nature of natural language is pivotal for understanding its semantics. However, the attention mechanism employed in the Transformer (Vaswani, 2017) architecture does not inherently capture this sequential information. Consequently, it is essential to incorporate positional encoding within the Transformer model to enable the processing of sequence-dependent information. For the query $\boldsymbol{q}$ with the position ID m and key $\boldsymbol{k}$ with

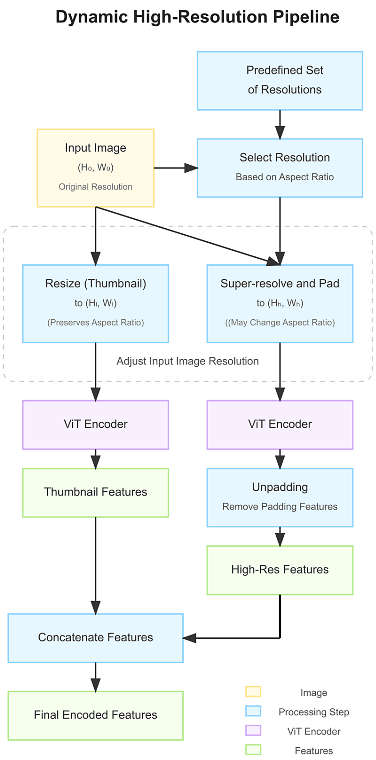


Figure 2: Flowchart of the Dynamic High-Resolution Method

the position ID n, positional encoding is applied to incorporate positional information into them:

$$\hat{\boldsymbol{q}} = PE(\boldsymbol{q}, m), \hat{\boldsymbol{k}} = PE(\boldsymbol{k}, n) \tag{5}$$

Positional encoding can be implemented in various ways (Gehring et al., 2017; Liu et al., 2020; Shaw et al., 2018; Dai, 2019; Raffel et al., 2020; He et al., 2020; Wang et al., 2019). Nowadays, in the choice of positional encoding methods, Rotary Position Embedding (RoPE) (Su et al., 2024) has become a prevalent encoding method. The implementation of RoPE is as follows:

$$RoPE(\mathbf{q}, m) = \mathcal{R}_m \mathbf{q} \tag{6}$$

where:

$$\mathcal{R}_m = \begin{pmatrix} A_0 & 0 & \cdots & 0 \\ 0 & A_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{d/2-1} \end{pmatrix} \tag{7}$$

$$A_i = \begin{pmatrix} \cos m\theta_i & -\sin m\theta_i \\ \sin m\theta_i & \cos m\theta_i \end{pmatrix} \tag{8}$$

$$\theta_i = \theta^{-\frac{2i}{d}} \tag{9}$$

Where d is the dimensionality of $\mathbf{q}$, θ is a hyperparameter, typically taking values ranging from 10^4 to 10^7.

RoPE exhibits several key characteristics:

- RoPE can be described as a form of absolute positional encoding because it uses the absolute positions of tokens during the encoding process. However, it also exhibits properties of relative positional encoding due to its mathematical property:

$$\begin{aligned} (\mathcal{R}_m \mathbf{q})^T (\mathcal{R}_n \mathbf{k}) &= \mathbf{q}^T \mathcal{R}_m^T \mathcal{R}_n \mathbf{k} \\ &= \mathbf{q}^T \mathcal{R}_{n-m} \mathbf{k} \end{aligned} \tag{10}$$

- RoPE exhibits a characteristic of long-range decay: for a query $\mathbf{q}$ at position m and a key $\mathbf{k}$ at position n, after encoding with RoPE, the dot product $(\mathcal{R}_m \mathbf{q})^T (\mathcal{R}_n \mathbf{k})$ generally decreases as the absolute value of $|m - n|$ increases. However, this property of RoPE is partially controversial, which we will discuss further in Section 3.1.

- The value of θ controls the positional encoding's sensitivity to positional differences. A smaller θ makes the model more sensitive to position changes, whereas a larger one facilitates the capture of long-range dependencies. Generally, the value of θ should increase as the training length increases (Men et al., 2024).

In the domain of VLMs, researchers are exploring modifications to RoPE to better accommodate multimodal features. Approaches such as CCA (Xing et al., 2025) and PyPE (Chen et al., 2025) aim to reconfigure position IDs from distinct angles, whereas V2PE (Ge et al., 2024) narrows the incremental scale of positional encodings specifically for image embeddings. Despite these advancements, none of these proposed methods sufficiently consider the prevalent application of super-resolution techniques—a critical aspect of the current technological landscape.

3 Analysis

3.1 On the long-range decay property of RoPE

In the RoPE paper(Su et al., 2024), the authors theoretically analyzed the long-range decay properties of RoPE:

$$\left| \sum_{i=0}^{d/2-1} \mathbf{q}_{[2i:2i+1]} \mathbf{k}_{[2i:2i+1]} e^{i(m-n)\theta_i} \right|$$
$$\leq \left(\max_i |h_{i+1} - h_i| \right) \sum_{i=0}^{d/2-1} |S_{i+1}| \quad (11)$$

where:

$$h_i = \mathbf{q}_{[2i:2i+1]} \mathbf{k}_{[2i:2i+1]} \quad (12)$$
$$S_j = \sum_{i=0}^{j-1} e^{i(m-n)\theta_i} \quad (13)$$

Since the value of $\frac{1}{d/2} \sum_{i=1}^{d/2} |S_i|$ is decreasing, the above formula indicates that the upper bound of $\mathbf{q}^T \mathcal{R}_{n-m} \mathbf{k}$ is decreasing as the relative distance $|m-n|$ increases.

They also plotted the $\mathbf{q}^T \mathcal{R}_{n-m} \mathbf{k}$ as a function of their relative distance, specifically for the case where $\mathbf{q}$ and $\mathbf{k}$ are all-one vectors, to illustrate RoPE's long-range decay properties.

Although the long-range decay property of RoPE is generally accepted, unlike positional encodings such as ALiBi(Press et al., 2021) that explicitly incorporate terms for long-range decay, some researchers have raised questions about this property, and the above inequality is not tight. Some researchers argue that if $\mathbf{q}$ and $\mathbf{k}$ are sampled from a standard multivariate normal distribution, the following formula holds:

$$\mathbb{E}_{\mathbf{q},\mathbf{k}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\mathbf{q}^\top \mathbf{R}_m \mathbf{k}] = 0 \quad \forall m \in \mathbb{Z} \quad (14)$$

leading them to conclude that RoPE does not possess the long-range decay property (Barbero et al., 2024).

However, their conclusions are based only on their rigorous assumptions. We point out that if $\mathbf{q} \sim \mathcal{N}(\mu_\mathbf{q}, \mathbf{I}), \mathbf{k} \sim \mathcal{N}(\mu_\mathbf{k}, \mathbf{I})$, the following formula holds:

$$\mathbb{E}[\mathbf{q}^\top \mathcal{R}_m \mathbf{k}] = {\mu_\mathbf{k}}^T \mathcal{R}_m \mu_\mathbf{q} \quad \forall m \in \mathbb{Z} \quad (15)$$

Furthermore, the trend of $\mathbb{E}[\mathbf{q}^\top \mathcal{R}_m \mathbf{k}]$ with respect to m is dependent on the value of $\mu_\mathbf{q}, \mu_\mathbf{k}$, and can be overall increasing or decreasing as m increases. More detailed results can be found in Appendix A.

Therefore, under the assumption of a normal distribution, we cannot prove that RoPE exhibits the property of long-range decay. However, deep neural networks possess a large number of parameters and are highly complex. During the training process, RoPE also influences model parameter updates, consequently affecting the activation values of query-key pairs. Thus, the simple assumption of a normal distribution is likely not representative of the actual situation.

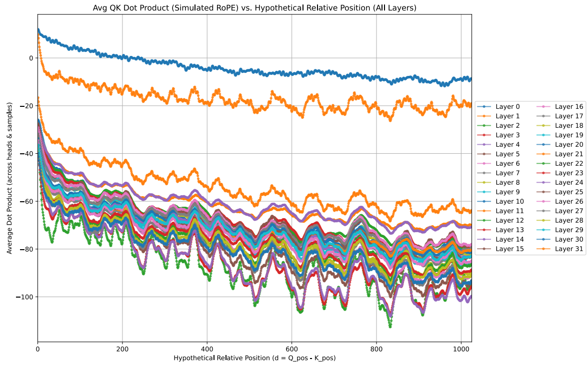


Figure 3: The Long-term Decay Property of RoPE. We randomly sampled 100 text data points from Wikitext and randomly selected 10 pairs of q-k from each layer of the Vicuna-7B model for computation.

To investigate whether RoPE exhibits a long-range decay property, we adopted an empirical approach. Specifically, we randomly sampled several data sequences from the WikiText (Merity et al., 2016) dataset. Then, for each layer, we randomly selected several q-k pairs before applying RoPE . By fixing these token pairs and progressively increasing their relative positions starting from 0, we measured the average inner product at each relative position. The results are shown in the Figure 3.

3.2 Problems with Previous Positional ID Arrangements

Having empirically demonstrated that RoPE indeed exhibits the long-range decay property in LLMs, we further analyze the issues inherent in previous positional encoding arrangements.

3.2.1 Disrupt the correspondence between thumbnail and high-resolution images.

Dynamic high-resolution methods employed by models such as LLaVA-Next simultaneously provide the LLM backbone with both high-resolution images and thumbnails. The high-resolution images furnish the model with fine-grained visual details, while the thumbnails offer global context. Similar to the introduction of RoPE in transformers for NLP to encourage attention mechanisms to focus on nearby tokens, images also exhibit local self-correlation. Consequently, during the interaction between high-resolution image tokens and thumbnail tokens, we aim for the high-resolution image tokens to attend more strongly to their corresponding thumbnail tokens. Here, two tokens are defined as corresponding if the image region encoded by the high-resolution token intersects with the image region encoded by the thumbnail token

However, the specific arrangement of position IDs in this dynamic high-resolution method, coupled with the long-term decay characteristic of RoPE, undermines this corresponding relationship. As shown in Figure 1a,

- For a token in the bottom-right corner of the high-resolution image, other tokens within the high-resolution region are relatively closer compared to their corresponding thumbnail tokens.

- For the token in the top-left corner of the high-resolution image, compared to its corresponding thumbnail token, its relative distance to the token in the bottom-right corner of the thumbnail is shorter.

As shown in Figure 4b, when computing the attention distribution from the red region of the high-resolution image towards the thumbnail, the attention can only focus on relevant information in shallow layers, while in deeper layers, attention is concentrated on unrelated areas.

3.2.2 Disrupts the interaction between text and image

Dynamic high-resolution methods produce a large number of image tokens. If a conventional position ID arrangement is used, this can result in excessive variation among the position IDs of image tokens corresponding to the same image. Assume a square image is input. In the dynamic high-resolution method, its width and height are scaled up to twice the original dimensions. Compared to approaches that do not use dynamic high resolution, the number of tokens increases by a factor of five, and consequently, the difference in positional encoding among image tokens also expands fivefold.

Effective acquisition of visual information during interaction with user instructions requires engaging with every image token. However, the distance between the top-left image token and the user instruction tokens is significant, causing the user instruction to attend more to the bottom-left corner of the image. The dynamic high-resolution method exacerbates this problem by increasing the difference in position IDs between the top-left and bottom-right tokens.

Furthermore, studies have shown that in VLMs, image tokens inherently receive less attention (Chen et al., 2024a). Coupled with RoPE's long-range decay characteristic, the excessive relative position between the top-left token and the user instruction tokens may lead to this part of the information being overlooked or neglected.

As shown in Figure 4d, when computing the attention distribution from 'each pair' towards the thumbnail, the attention is neither able to focus on the corresponding text in the image nor on the corresponding object.

4 Methods

According to the calculation formula of RoPE, it can be observed that during inference, the relative distance between $\mathbf{q}$ and $\mathbf{k}$ is influenced not by their actual distance in the sequence, but by the difference in their position IDs. Simultaneously, as shown in Section 3.1, increasing the difference between the position IDs of $\mathbf{q}$ and $\mathbf{k}$ can enhance their attention coefficient, while decreasing it can reduce it. Therefore, we propose to alleviate the aforementioned issues by rearranging the position IDs. Our approach is as follows:

- For the tokens of thumbnails, we adopt the same position IDs as those used in the previously established approach.

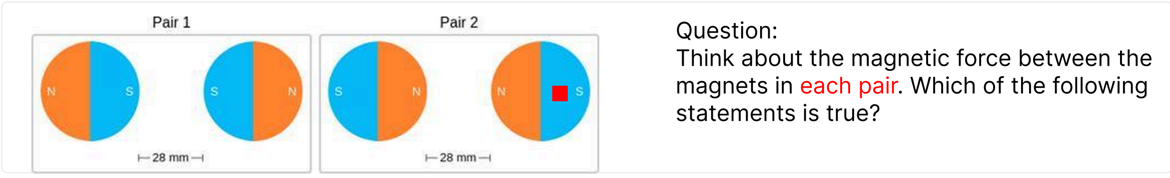
- For the tokens of high-resolution images, we assign them the same position ID as their corresponding thumbnail image tokens.

The difference between our method and the original approach can be seen in Figure 1. More details are available in Appendix B.

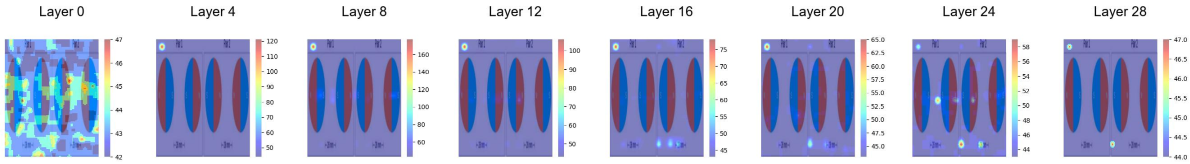
5 Experiments and Results

5.1 Experiments Setup

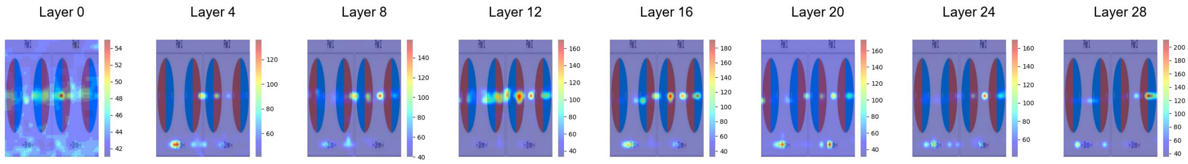
We adopted the LLaVA-Next architecture (Liu et al., 2024c). We used the Vicuna-1.5 7B (Zheng et al., 2023) as the LLM backbone and CLIP ViT-L/14 (336) (Radford et al., 2021) as the vision encoder. Alternatively, we used Qwen-2.5-7B-Instruct (Yang et al., 2024) as the backbone and SigLip 400M (Zhai et al., 2023) as the encoder. It is worth noting that the RoPE θ for the Qwen series models is 10^7, which is significantly larger than that of the Vicuna model (10^4). This indicates



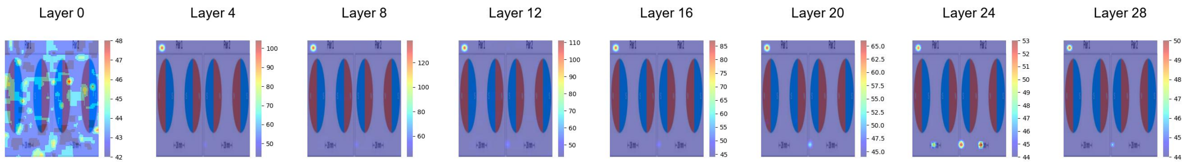
(a) Data example from MMBench, where the red square and red text are used for attention computation.



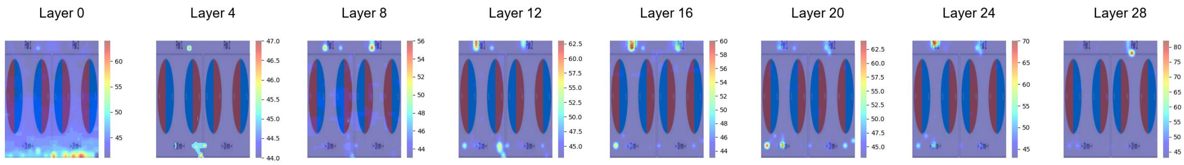
(b) Attention distribution for the red region (w/o ID-Align).



(c) Attention distribution for the red region (w ID-Align).



(d) Attention distribution for the red text (w/o ID-Align).



(e) Attention distribution for the red text (w ID-Align).

Figure 4: Attention distributions from the red region in the high-resolution image and the red text towards thumbnail tokens. Figure 4a shows the data example. Figures 4b and 4c depict the attention distribution from the red region, and figures 4d and 4e show the attention distribution from the red text.

that Qwen models are relatively less sensitive to changes in positional IDs. More details can be found in the Appendix C

5.2 Results and Analysis

From the perspective of attention distribution, in Figure 4c compared to 4b, the attention corresponding to the red region is no longer confined to certain unrelated areas but can focus on the magnets in the image. In Figure 4e compared to Figure 4d, the attention of 'each pair' can focus on the corresponding text portion in the thumbnail.

The primary experimental results are shown in Table 1. As can be observed from the table, the adoption of ID-Align has led to improvements in the model's performance metrics across various benchmarks. When using Vicuna and CLIP as pre-training models, there was a notable improvement across all benchmarks. These benchmarks cover a broad spectrum of capabilities, indicating the effectiveness of our approach. When employing Qwen2.5, which has a RoPE θ value of 10^7, and SigLIP as the base models, the performance gains were observed to decrease, and there was a decline in performance on several benchmarks. This observation aligns with our analysis, which suggests that these models are relatively insensitive to changes in positional encoding. However, after adopting ID-Align, the overall performance of the model showed an increasing trend.

To further investigate which specific capabilities

Model	MMBench$_{dev}$	MMStar	RealWorldQA	SEEDB2-Plus	POPE@ACC
Vicuna					
w/o ID-Align	66.58	36.61	58.43	51.38	87.97
w/ ID-Align	**68.21** (+1.63)	**38.32** (+1.71)	**59.18** (+0.75)	**51.56** (+0.18)	**88.66** (+0.69)
Qwen					
w/o ID-Align	78.14	**50.53**	**64.18**	61.00	**89.17**
w/ ID-Align	**78.48** (+0.34)	50.14 (-0.39)	63.79 (-0.39)	**62.06** (+1.06)	89.16 (-0.01)
	MME	**AI2D**	**VQAV2$_{val}$**	**SQA$_{img}$**	**Avg**
Vicuna					
w/o ID-Align	65.22	65.74	79.75	69.41	64.57
w/ ID-Align	**65.50** (+0.28)	**66.39** (+0.65)	**80.02** (+0.27)	**70.70** (+1.29)	**65.39** (+0.82)
Qwen					
w/o ID-Align	67.11	74.84	79.88	80.61	71.72
w/ ID-Align	**68.22** (+1.11)	**75.13** (+0.29)	**80.25** (+0.37)	**81.06** (+0.45)	**72.03** (+0.31)

Table 1: Performance on Different Benchmarks with and without ID-Align

Model	CP	FP-S	FP-C	AR	RR	LR
Vicuna						
w/o ID-Align	79.39	70.31	58.04	69.35	60.87	**36.44**
w/ ID-Align	**81.76** (+2.37)	**71.67** (+1.36)	**59.44** (+1.40)	**69.85** (+0.50)	**66.96** (+6.09)	34.75 (-1.69)
Qwen						
w/o ID-Align	**83.73**	**81.91**	71.26	**84.38**	75.83	56.65
w/ ID-Align	82.87 (-0.86)	**81.91** (+0.00)	**72.87** (+1.61)	83.33 (-1.05)	**77.72** (+1.89)	**59.54** (+2.89)

Table 2: The table presents the results on sub-metrics from the MMBench-Dev. Specifically, **CP** stands for Coarse Perception, **FP-C** represents Fine-grained Perception (cross-instance), **FP-S** denotes Fine-grained Perception (single-instance), **AR** refers to Attribute Reasoning, **LR** indicates Logical Reasoning, **RR** represents Relation Reasoning.

contributed most to the observed growth in benchmark performance, we have detailed the changes in various sub-metrics of MMbench, as shown in Table 2. We have also listed the subtasks of MMBench in Appendix D.3. As can be observed, when using Vinca as the LLM base, although all sub-indicators showed improvement, the most significant growth was seen in the RR metrics. Meanwhile, when employing qwen as the LLM backbone, it was the FP-C, RR, and LR metrics that maintained their growth. These metrics are all related to global information.

6 Conclusion

In this paper, we analyze the potential issues of the dynamic high-resolution strategies adopted by current VLMs. Based on our analysis, we propose **ID-Align**: a method that aligns the position IDs of high-resolution embeddings with their corresponding low-resolution embeddings, preserving their relationship and constraining excessive growth in position IDs. We conducted experiments on the LLaVA-Next architecture, demonstrating the effectiveness of our approach.

7 Limitation

Limitations of our work include: we did not investigate the performance of our method when combined with token compression techniques. We also did not examine the performance of our method when integrated with viT that inherently support dynamic resolution.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2024. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen and Long Xing. 2024. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. `https://github.com/xiaoachen98/Open-LLaVA-NeXT`.

Zhanpeng Chen, Mingxiao Li, Ziyang Chen, Nan Du, Xiaolong Li, and Yuexian Zou. 2025. Advancing general multimodal capability of vision-language models with pyramid-descent visual position encoding. *arXiv preprint arXiv:2501.10967*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Zihang Dai. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. 2024. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *ArXiv*, abs/2412.09616.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Xuanqing Liu, Hsiang-Fu Yu, Inderjit Dhillon, and Cho-Jui Hsieh. 2020. Learning to encode position for transformer with continuous dynamical model. In *International conference on machine learning*, pages 6327–6335. PMLR.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024e. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024f. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. 2024. Base of rope bounds context length. *arXiv preprint arXiv:2405.14591*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. 2019. Encoding word order in complex embeddings. *arXiv preprint arXiv:1912.12333*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2025. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37:92012–92035.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024b. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Long-term decay of RoPE

The proof of Equation (11) is as follows:
Let:

$$\begin{aligned} h_i &= \mathbf{q}_{[2i:2i+1]}\mathbf{k}_{[2i:2i+1]} \\ S_j &= \sum_{i=0}^{j-1} e^{i(m-n)\theta_i} \end{aligned} \tag{16}$$

Setting $h_{d/2} with = 0$ and $S_0 = 0$, with the Abel transformation, we have:

$$\begin{aligned} &\sum_{i=0}^{d/2-1} \mathbf{q}_{[2i:2i+1]}\mathbf{k}^*_{[2i:2i+1]} e^{i(m-n)\theta_i} \\ &= \sum_{i=0}^{d/2-1} h_i(S_{i+1} - S_i) \\ &= -\sum_{i=0}^{d/2-1} S_{i+1}(h_{i+1} - h_i). \end{aligned} \tag{17}$$

Thus,

$$\begin{aligned} &\left| \sum_{i=0}^{d/2-1} \mathbf{q}_{[2i:2i+1]}\mathbf{k}^*_{[2i:2i+1]} e^{i(m-n)\theta_i} \right| \\ &= \left| \sum_{i=0}^{d/2-1} S_{i+1}(h_{i+1} - h_i) \right| \\ &\leq \sum_{i=0}^{d/2-1} |S_{i+1}||(h_{i+1} - h_i)| \\ &\leq \left(\max_i |h_{i+1} - h_i| \right) \sum_{i=0}^{d/2-1} |S_{i+1}| \end{aligned} \tag{18}$$

The proof of Equation (15) is as follows:

Let $\mathbf{q} \sim \mathcal{N}(\mu_q, \mathbf{I})$ and $\mathbf{k} \sim \mathcal{N}(\mu_k, \mathbf{I})$ be independent random vectors

We use the law of total expectation, conditioning on $\mathbf{k}$:

$$\begin{aligned} \mathbb{E}[\mathbf{q}^\top \mathcal{R}_m \mathbf{k}] &= \mathbb{E}_{\mathbf{k}} \left[\mathbb{E}_{\mathbf{q}|\mathbf{k}}[\mathbf{q}^\top \mathcal{R}_m \mathbf{k} \mid \mathbf{k}] \right] \\ &= \mathbb{E}_{\mathbf{k}} \left[\mathbb{E}[\mathbf{q}^\top \mid \mathbf{k}] \mathcal{R}_m \mathbf{k} \right] \\ &= \mathbb{E}_{\mathbf{k}} \left[\mathbb{E}[\mathbf{q}^\top] \mathcal{R}_m \mathbf{k} \right] \\ &= \mathbb{E}_{\mathbf{k}} \left[\boldsymbol{\mu}_q^\top \mathcal{R}_m \mathbf{k} \right] \\ &= \boldsymbol{\mu}_q^\top \mathcal{R}_m \mathbb{E}_{\mathbf{k}}[\mathbf{k}] \\ &= \boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k. \end{aligned}$$

The i-th 2×2 block of $\mathcal{R}_m$, denoted $\mathcal{R}_m^{(i)}$, is given by:

$$\mathcal{R}_m^{(i)} = \begin{pmatrix} \cos(m\theta_i) & -\sin(m\theta_i) \\ \sin(m\theta_i) & \cos(m\theta_i) \end{pmatrix}$$

First, the product $\mathcal{R}_m \boldsymbol{\mu}_k$ results in a vector where the components corresponding to the i-th 2D block are:

$$\begin{aligned} (\mathcal{R}_m \boldsymbol{\mu}_k)_{2i-1} &= \mu_{k,2i-1} \cos(m\theta_i) - \mu_{k,2i} \sin(m\theta_i) \\ (\mathcal{R}_m \boldsymbol{\mu}_k)_{2i} &= \mu_{k,2i-1} \sin(m\theta_i) + \mu_{k,2i} \cos(m\theta_i) \end{aligned}$$

The dot product $\boldsymbol{\mu}_q^\top (\mathcal{R}_m \boldsymbol{\mu}_k)$ is then:

$$\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k = \sum_{j=1}^{d} \mu_{q,j} (\mathcal{R}_m \boldsymbol{\mu}_k)_j$$

Grouping the summation by the $d/2$ two-dimensional blocks:

$$\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k =$$
$$\sum_{i=1}^{d/2} [\mu_{q,2i-1}(\mathcal{R}_m \boldsymbol{\mu}_k)_{2i-1} + \mu_{q,2i}(\mathcal{R}_m \boldsymbol{\mu}_k)_{2i}] =$$
$$\sum_{i=1}^{d/2} [\mu_{q,2i-1}(\mu_{k,2i-1}\cos(m\theta_i) - \mu_{k,2i}\sin(m\theta_i))$$
$$+ \mu_{q,2i}(\mu_{k,2i-1}\sin(m\theta_i) + \mu_{k,2i}\cos(m\theta_i))]$$

Rearranging terms within the sum based on

$$\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k =$$
$$\sum_{i=1}^{d/2} [\mu_{q,2i-1}(\mathcal{R}_m \boldsymbol{\mu}_k)_{2i-1} + \mu_{q,2i}(\mathcal{R}_m \boldsymbol{\mu}_k)_{2i}] =$$
$$\sum_{i=1}^{d/2} [\mu_{q,2i-1}(\mu_{k,2i-1}\cos(m\theta_i) - \mu_{k,2i}\sin(m\theta_i))$$
$$+\mu_{q,2i}(\mu_{k,2i-1}\sin(m\theta_i) + \mu_{k,2i}\cos(m\theta_i))]$$

To simplify notation, let:

$$A_i = \mu_{q,2i-1}\mu_{k,2i-1} + \mu_{q,2i}\mu_{k,2i}$$
$$B_i = \mu_{q,2i}\mu_{k,2i-1} - \mu_{q,2i-1}\mu_{k,2i}$$

The expression then becomes:

$$\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k = \sum_{i=1}^{d/2}(A_i \cos(m\theta_i) + B_i \sin(m\theta_i))$$

From this expression, we cannot derive the trend of $\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k$ as m changes. Next, we will demonstrate experimentally that $\boldsymbol{\mu}_q^\top \mathcal{R}_m \boldsymbol{\mu}_k$ exhibits different trends with respect to m depending on the values of $\boldsymbol{\mu}_q$ and $\boldsymbol{\mu}_k$.

For each component of $\mathbf{q}$and $\mathbf{k}$, we sampled from normal distributions with the same mean and a standard deviation of 1. Different mean values were set for $\mathbf{q}$ and $\mathbf{k}$ in each experimental run. Then, we set the relative distance between them to different values and calculated their attention scores. For each choice of mean value, we simulated 1000 times and averaged the results at each relative position. We experimented with two values of θ, 10^4 and 10^7. The results are shown in Table 5. The experiments reveal that different values of μ_q and μ_k influence the long-term properties of RoPE, and a small value of θ increases the positional sensitivity of dot-product attention.

B Method details

Through the reorganization of position IDs, the "distance" between thumbnail tokens and their corresponding high-resolution tokens is reduced. This adjustment not only brings related embeddings closer in terms of positional encoding but also effectively restricts the growth of position IDs. Consequently, this approach prevents the issue of position IDs increasing by thousands when processing a single image, which could otherwise lead to exceeding the maximum position ID values encountered during training.

Our algorithm process is shown in Algorithm 1. In practice, assuming that the 2D feature map obtained after encoding the thumbnail with ViT has dimensions (H_0, W_0) , and the feature map obtained after encoding the entire high-resolution image has dimensions (H_1, W_1) , for simplicity, we assume that the positional id of the first token in the thumbnail image is 0. We first generate a 1D tensor ranging from 0 to $H_0 * W_0$ then reshape it to (H_0, W_0) , and use interpolation to resize the reshaped tensor to (H_1, W_1), rounding the values to intenger. After flattening both parts, they are concatenated to form our positional IDs.

C Experiments Setup

All experiments were conducted using eight A800 GPUs.

C.1 Parameter Settings

As for the hyperparameter settings, we adopted the configurations from Open-LLaVA-Next (Chen and Xing, 2024). We will also list these hyperparameters below.

Listing 1: The script for the LLaVA-Next pretrain phase, using Vicuna and CLIP as the LLM backbone and visual encoder, respectively.

```
nnodes=1
num_gpus=8
deepspeed --num_nodes ${nnodes} --
    num_gpus ${num_gpus} --master_port
    =10270 llava/train/train_mem.py \
    --deepspeed ./scripts/zero2.json \
    --model_name_or_path ${MODEL_PATH} \
    --version plain \
    --data_path ${DATA_PATH} \
    --image_folder ${IMAGE_FOLDER} \
    --vision_tower ${VISION_TOWER} \
    --mm_projector_type mlp2x_gelu \
    --tune_mm_mlp_adapter True \
    --unfreeze_mm_vision_tower False \
    --mm_vision_select_layer -2 \
    --mm_use_im_start_end False \
    --mm_use_im_patch_token False \
```

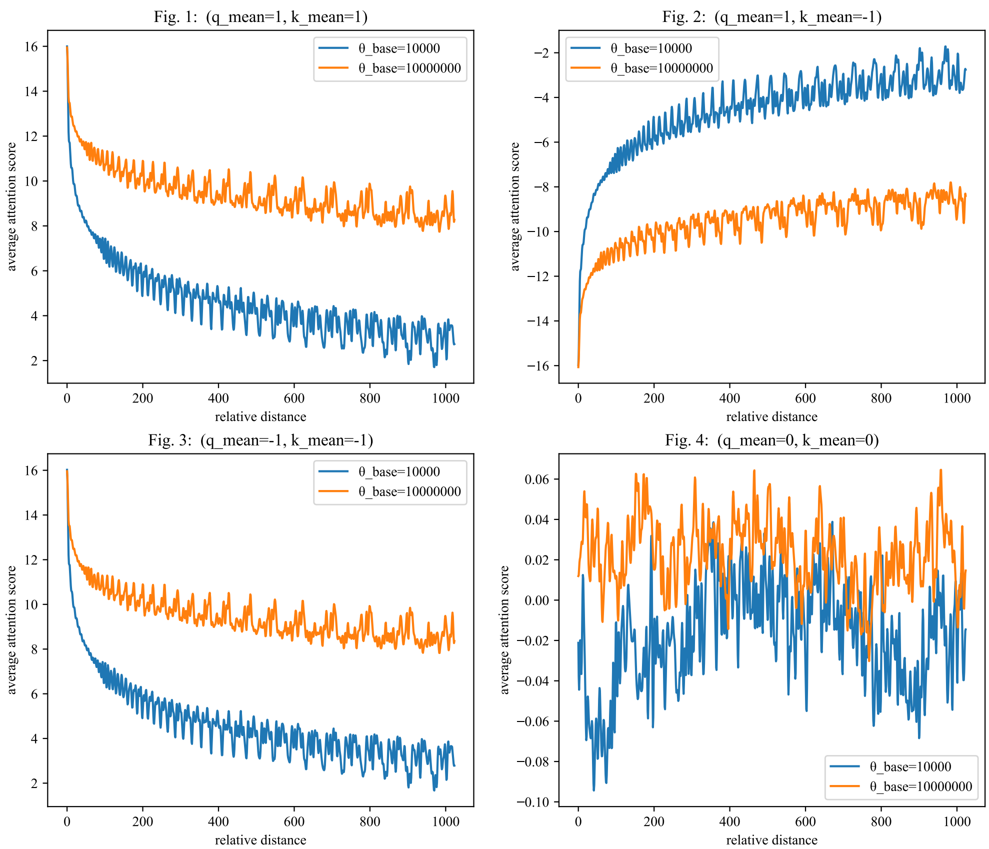


Figure 5: Simulation of RoPE's Long-term Properties under Different μ_q and μ_k

Algorithm 1 ID-Align with RoPE

Require:
- E_{text}: Sequence of text embeddings
- E_{low}: Sequence of thumbnail embeddings
- E_{high}: Sequence of high-resolution image embeddings
- $\mathcal{M} : E_{\text{high}} \to E_{\text{low}}$: Return the E_{low} corresponding to E_{high}

Ensure:

$max\_pid \leftarrow 0$
$E_{\text{merged}} \leftarrow \text{Concat}(E_{\text{text}}, E_{\text{low}}, E_{\text{high}})$
for each embedding $e_i \in E_{\text{merged}}$ **do**
    if $e_i \in E_{\text{text}} \cup E_{\text{low}}$ **then**
        $\text{pos\_id}(e_i) \leftarrow max\_pid$
        $max\_pid \leftarrow max\_pid + 1$
    else if $e_i \in E_{\text{high}}$ **then**
        $\text{pos\_id}(e_i) \leftarrow \text{pos\_id}(\mathcal{M}(e_i))$
        $max\_pid \leftarrow \max(max\_pid, \mathcal{M}(e_i) + 1)$
    end if
end for

function ApplyRotaryEncoding(E_{merged})
    for each $e_i \in E_{\text{merged}}$ **do**
        $e_i \leftarrow \text{RoPE}(e_i, \text{pos\_id}(e_i))$
    end for
    return E_{merged}
end function

```
    --mm_patch_merge_type spatial_unpad
        \
    --image_aspect_ratio anyres \
    --group_by_modality_length False \
    --bf16 True \
    --output_dir ./checkpoints/${
        RUN_NAME} \
    --num_train_epochs 1 \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 4 \
    --gradient_accumulation_steps 4 \
    --evaluation_strategy "no" \
    --image_grid_pinpoints "[(336, 672),
        (672, 336), (672, 672), (1008,
        336), (336, 1008)]" \
    --use_id_align True \
    --save_strategy "steps" \
    --save_steps 24000 \
    --save_total_limit 1 \
    --learning_rate 1e-3 \
    --weight_decay 0. \
    --warmup_ratio 0.03 \
    --lr_scheduler_type "cosine" \
    --logging_steps 1 \
    --tf32 True \
    --model_max_length 4096 \
    --gradient_checkpointing True \
    --dataloader_num_workers 4 \
    --lazy_preprocess True \
    --report_to None \
    --run_name ${RUN_NAME}
```

Listing 2: The script for the LLaVA-Next finetune phase, using Vicuna and CLIP as the LLM backbone and visual encoder, respectively.

```
nnodes=1
num_gpus=8

deepspeed --num_nodes ${nnodes} --
    num_gpus ${num_gpus} --master_port
    =10271 llava/train/train_mem.py \
    --deepspeed ./scripts/zero3.json \
    --model_name_or_path ${MODEL_PATH} \
    --version v1 \
    --data_path ${DATA_PATH} \
    --image_folder ${IMAGE_FOLDER} \
    --pretrain_mm_mlp_adapter ./
        checkpoints/${BASE_RUN_NAME}/
        mm_projector.bin \
    --unfreeze_mm_vision_tower True \
    --mm_vision_tower_lr 2e-6 \
    --vision_tower ${VISION_TOWER} \
    --mm_projector_type mlp2x_gelu \
    --mm_vision_select_layer -2 \
    --mm_use_im_start_end False \
    --use_id_align True \
    --mm_use_im_patch_token False \
    --group_by_modality_length True \
    --image_aspect_ratio anyres \
    --mm_patch_merge_type spatial_unpad
        \
    --bf16 True \
    --image_grid_pinpoints "[(336, 672),
        (672, 336), (672, 672), (1008,
        336), (336, 1008)]" \
    --output_dir ./checkpoints/${
        RUN_NAME} \
    --num_train_epochs 1 \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 4 \
    --gradient_accumulation_steps 2 \
    --evaluation_strategy "no" \
    --save_strategy "steps" \
    --save_steps 1000 \
    --save_total_limit 1 \
    --learning_rate 2e-5 \
    --weight_decay 0. \
    --warmup_ratio 0.03 \
    --lr_scheduler_type "cosine" \
    --logging_steps 1 \
    --tf32 True \
    --model_max_length 4096 \
    --gradient_checkpointing True \
    --dataloader_num_workers 4 \
    --lazy_preprocess True \
    --report_to none \
    --run_name ${RUN_NAME}
```

Listing 3: The script for the LLaVA-Next pre-train phase, using Qwen and SigLIP as the LLM backbone and visual encoder, respectively.

```
nnodes=1
num_gpus=8
deepspeed --num_nodes ${nnodes} --
    num_gpus ${num_gpus} --master_port
    =10270 llava/train/train_mem.py \
```

```
    --deepspeed ./scripts/zero2.json \
    --model_name_or_path ${MODEL_PATH} \
    --version plain \
    --data_path ${DATA_PATH} \
    --image_folder ${IMAGE_FOLDER} \
    --vision_tower ${VISION_TOWER} \
    --mm_projector_type mlp2x_gelu \
    --tune_mm_mlp_adapter True \
    --unfreeze_mm_vision_tower False \
    --mm_vision_select_layer -2 \
    --mm_use_im_start_end False \
    --mm_use_im_patch_token False \
    --mm_patch_merge_type spatial_unpad \
    --image_aspect_ratio anyres \
    --group_by_modality_length False \
    --bf16 True \
    --output_dir ./checkpoints/${RUN_NAME} \
    --num_train_epochs 1 \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 4 \
    --gradient_accumulation_steps 4 \
    --evaluation_strategy "no" \
    --image_grid_pinpoints "[(384, 768), (768, 384), (768, 768), (1152, 384), (384, 1152)]" \
    --use_id_align True \
    --save_strategy "steps" \
    --save_steps 24000 \
    --save_total_limit 1 \
    --learning_rate 1e-3 \
    --weight_decay 0. \
    --warmup_ratio 0.03 \
    --lr_scheduler_type "cosine" \
    --logging_steps 1 \
    --tf32 True \
    --model_max_length 32768 \
    --gradient_checkpointing True \
    --dataloader_num_workers 4 \
    --lazy_preprocess True \
    --report_to none \
    --run_name ${RUN_NAME}
```

Listing 4: The script for the LLaVA-Next finetune phase, using Qwen and SigLIP as the LLM backbone and visual encoder, respectively

```
nnodes=1
num_gpus=8
deepspeed --num_nodes ${nnodes} --num_gpus ${num_gpus} --master_port=10271 llava/train/train_mem.py \
    --deepspeed ./scripts/zero3.json \
    --model_name_or_path ${MODEL_PATH} \
    --version ${PROMPT_VERSION} \
    --data_path ${DATA_PATH} \
    --image_folder ${IMAGE_FOLDER} \
    --pretrain_mm_mlp_adapter ./checkpoints/${BASE_RUN_NAME}/mm_projector.bin \
    --unfreeze_mm_vision_tower True \
    --mm_vision_tower_lr 2e-6 \
    --vision_tower ${VISION_TOWER} \
    --mm_projector_type mlp2x_gelu \
    --mm_vision_select_layer -2 \
    --mm_use_im_start_end False \
    --use_id_align True \
    --mm_use_im_patch_token False \
    --group_by_modality_length True \
    --image_aspect_ratio anyres \
    --mm_patch_merge_type spatial_unpad \
    --bf16 True \
    --image_grid_pinpoints "[(384, 768), (768, 384), (768, 768), (1152, 384), (384, 1152)]" \
    --output_dir ./checkpoints/${RUN_NAME} \
    --num_train_epochs 1 \
    --per_device_train_batch_size 8 \
    --per_device_eval_batch_size 4 \
    --gradient_accumulation_steps 2 \
    --evaluation_strategy "no" \
    --save_strategy "steps" \
    --save_steps 1000 \
    --save_total_limit 1 \
    --learning_rate 2e-5 \
    --weight_decay 0. \
    --warmup_ratio 0.03 \
    --lr_scheduler_type "cosine" \
    --logging_steps 1 \
    --tf32 True \
    --model_max_length 32768 \
    --gradient_checkpointing True \
    --dataloader_num_workers 4 \
    --lazy_preprocess True \
    --report_to none \
    --run_name ${RUN_NAME}
```

C.2 Benchmarks

Focusing on the overall and various hierarchical capabilities of models, we primarily adopted three benchmarks—MMBench (Liu et al., 2024e), MME (Yin et al., 2023), and MMStar (Chen et al., 2024b). Additionally, SeedBench-2-Plus (Li et al., 2024) and AI2D (Kembhavi et al., 2016) were utilized to assess the models' capability in processing rich text images such as charts, maps, and web pages. RealWorldQA was employed to evaluate the models' effectiveness in handling real-world images, whereas POPE (Li et al., 2023b) was used to examine the phenomenon of model hallucinations. To evaluate the model's performance on QA tasks, we will utilize the VQAv2 (Goyal et al., 2017) and ScienceQA (Lu et al., 2022) datasets. We utilized LMMS-Eval (Zhang et al., 2024b) for the evaluation of our model. The decision to utilize ID-Align can be controlled by setting the value of use-id-align.

D More Results and Analysis

D.1 Learning Curve

In this section, we also plot the learning curve. From these curves, it can be observed that after applying ID-Align, the training loss is slightly lower during the latter half of the training phase compared to when not using ID-Align. Additionally, the gradient norm is notably lower, indicating that the model is closer to achieving convergence. This effect is especially pronounced on Vinuca. These plots were generated using a sliding average window with a window length of 100.

D.1.1 Vicuna

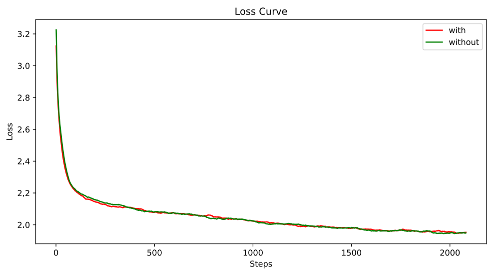


Figure 6: Pretrain Loss

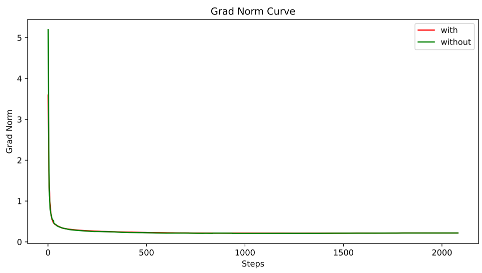


Figure 7: Pretrain Grad Norm

Figure 8: Finetune Loss

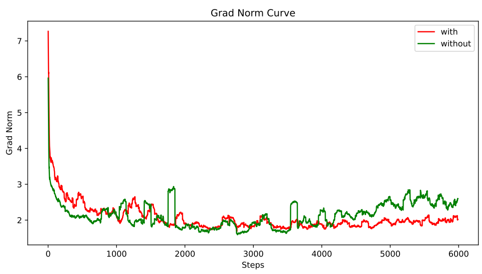


Figure 9: Finetune Grad Norm

D.1.2 Qwen

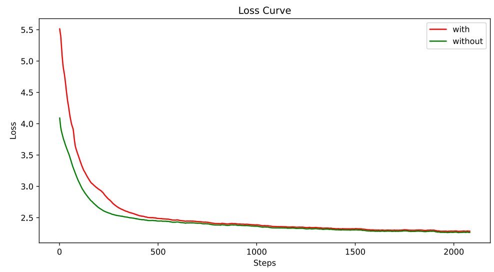


Figure 10: Pretrain Loss

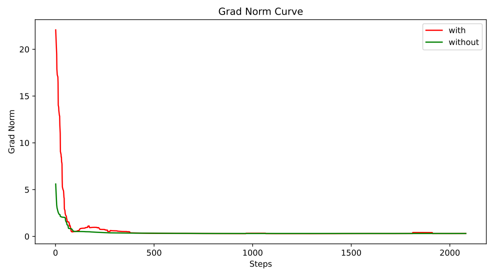


Figure 11: Pretrain Grad Norm

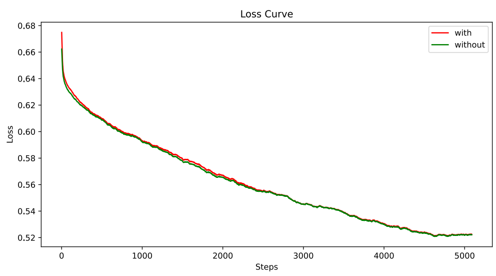


Figure 12: Finetune Loss

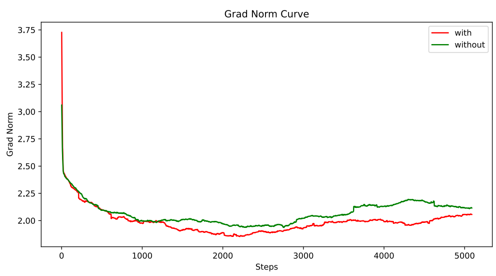


Figure 13: Finetune Grad Norm

D.2 Compare with Other Methods

We also compared our method with MRoPE(Wang et al., 2024) and V2PE(Ge et al., 2024). Our method is not in competition with these methods;

rather, it is compatible with them. The focus of these methods is on positional encodings within a single image, whereas our method addresses the correspondence between thumbnail and high-resolution images. Therefore, these methods can be combined.

Due to the limitation of computing resource, we only experimented with the Qwen-2.5-0.5B model and SigLip. These experiments only involved adjusting these two methods to suit high-resolution scenarios, without combining them with ID-Align. For V2PE, we set $\epsilon = 0.5$, which is the best result reported in their paper for conventional benchmarks. For MROPE, we treated the thumbnail and high-resolution images as separate images. Since the hidden state dimension of the 0.5B model is small, we set $MRoPE \quad section = [8, 12, 12]$, which is a proportionally scaled version of $MRoPE \quad section = [16, 24, 24]$ used in the Qwen-2.5-VL-3B model. The results are shown in Table 3.

Compared to V2PE and MRoPE, our method shows significant improvement in metrics that measure the overall capability of the model (MME, MMBench, MMStar). In metrics that measure specific capabilities, such as PoPE and AI2D, our method may not perform as well as V2PE or MRoPE, which could be related to the characteristics of their methods and the benchmark data distribution. Overall, in the context of dynamic high-resolution, our method is superior.

D.3 MMBench Leaf Tasks

Coarse Perception:

- Image Style
- Image Topic
- Image Scene
- Image Mood
- Image Quality

Fine-grained Perception (Single-instance):

- Attribute Recognition
- Celebrity Recognition
- Object Localization
- OCR

Fine-grained Perception (Cross-instance):

- Spatial Relationship
- Attribute Comparison
- Action Recognition

Attribute Reasoning:

- Physical Property Reasoning
- Function Reasoning
- Identity Reasoning

Relation Reasoning:

- Social Relation
- Nature Relation
- Physical Relation

Logic Reasoning:

- Future Prediction
- Structuralized Image-text Understanding

	MMB	MMStar	RWQA	SEEDB	POPE	MME-C	MME-P	AI2D	VQAV2	SQA	avg
V2PE	56.28	37.43	51.76	**48.09**	87.82	30.85	63.86	**57.87**	**65.62**	60.83	56.04
MRoPE	55.44	38.28	53.73	46.73	**88.41**	30.01	62.81	57.77	64.78	59.89	55.79
ID-Align	**57.68**	**39.74**	**55.03**	47.56	87.50	**31.03**	**64.03**	56.96	64.86	**60.88**	**56.53**

Table 3: Comparison with MRoPE and V2PE