Advancing General Multimodal Capability of Vision-language Models with Pyramid-descent Visual Position Encoding

Anonymous ACL submission

Abstract

Vision-language Models (VLMs) have shown remarkable capabilities in advancing general artificial intelligence, yet the irrational encoding of visual positions persists in inhibiting the models' comprehensive perception performance across different levels of granularity. In this work, we propose Pyramid-descent Visual Position Encoding (PyPE), a novel approach designed to enhance the perception of visual tokens within VLMs. By assigning visual position indexes from the periphery to the center and expanding the central receptive field incrementally, PyPE addresses the limitations of traditional raster-scan methods and mitigates the long-term decay effects induced by Rotary Position Embedding (RoPE). Our method reduces the relative distance between interrelated visual elements and instruction tokens, promoting a more rational allocation of attention weights and allowing for a multi-granularity perception of visual elements and countering the over-reliance on anchor tokens. Extensive experimental evaluations demonstrate that PyPE consistently improves the general capabilities of VLMs across various sizes. Code is available at `https://anonymous.4open.science/r/PyPE-34EE`.

1 Introduction

Large Language Models (LLMs) (Touvron et al., 2023; Brown et al., 2020) demonstrate significant universal capabilities that contribute to the pursuit of general artificial intelligence. However, language constitutes only one aspect of communication. Visual information plays a crucial role in augmenting and enhancing our understanding of the world. Consequently, there is a growing interest in the development of Vision-language Models (VLMs) (Chen et al., 2024c; Peng et al., 2023; Wang et al., 2024; Bai et al., 2023) that can process and integrate visual modality. To effectively leverage the powerful contextual understanding capabilities of LLMs, VLMs project visual information to the same dimensionality as textual embeddings through specific projection layers (Chen et al., 2023; Li et al., 2023b; Zhou et al., 2024), which are then directly embedded into the text sequence to form the input for the foundation LLMs, enabling cross-modal alignment and instruction-following learning using next-token prediction.

Despite their commendable progress, the typical processing of visual information does not align with the distribution patterns of visual elements. Since visual information is composed of fixed-sized patches obtained through raster scanning, patches located closer to the bottom right corner of the image are positioned nearer to the instruction tokens within the sequence. Due to the long-term decay from Rotary Position Embedding (RoPE) (Su et al., 2024), visual tokens closer to the instruction tokens will be more likely to receive higher attention weights, and vice versa. This is counter-intuitive, as the importance of visual information is not defined by the order of raster-scanning. Xing et al. (2024) observe a similar phenomenon by visualizing the attention information flow from instruction tokens to visual tokens in the first layer of the decoder. Consequently, they propose Concentric Causal Attention (CCA), which starts assigning the position indexes of images from the peripheral and ends in the center, to alleviate the long-term decay in RoPE and improve causal attention following 2D spatial locality of images. Although CCA is both intuitive and effective, its applicability is constrained by the assumption that all significant elements related to the instructions are situated at the center of the image. This assumption inherently results in a loss of detail, limiting its effectiveness in capturing comprehensive information.

To further investigate the impact of raster-scan and concentric PE on the fine-grained modeling of visual information, we extend the visualization to

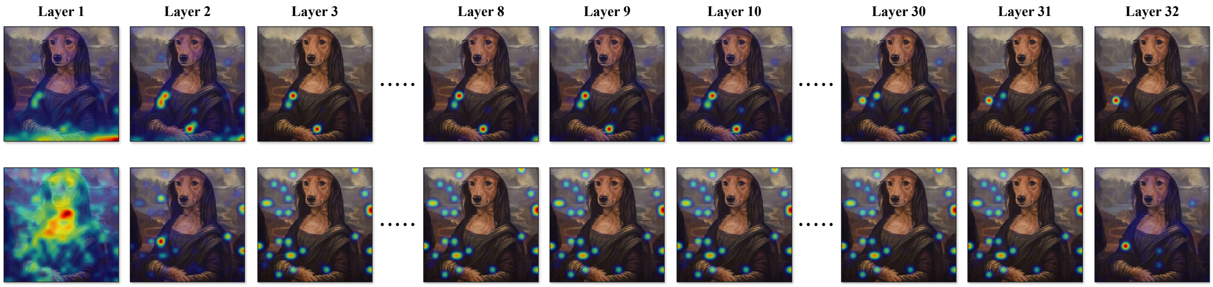


Figure 1: Layer-wise attention visualization of visual-to-instruction information flow. Displayed from top to bottom are the attention heatmaps from LLaVA-1.5-7B trained with raster-scan and concentric PE, respectively. The example is derived from LLaVA-Bench (Liu et al., 2024b) and the query is *"Describe this photo in detail"*.

all layers of the decoder. As illustrated in Figure 1, CCA demonstrates exceptional performance in the first layer, alleviating the long-term decay caused by RoPE in the raster-scan approach, thereby directing the model's attention to more significant areas. However, in the subsequent layers, both methods largely maintain the same attention patterns as observed in their respective third layers, with changes only occurring in the final layer. A similar phenomenon, namely "aggregation pattern", is observed in OPERA (Huang et al., 2024), where both LLMs and VLMs tend to generate new tokens by concentrating on a limited number of summary tokens (also referred to as anchor tokens (Wang et al., 2023)) rather than considering all preceding tokens. This tendency towards partial overtrust leads to the neglect of fine-grained image tokens, resulting in the generation that may be hallucinatory and do not accurately reflect the image content. Moreover, it has been demonstrated in OPERA that more hallucinations are generated when more summary tokens appear in the context.

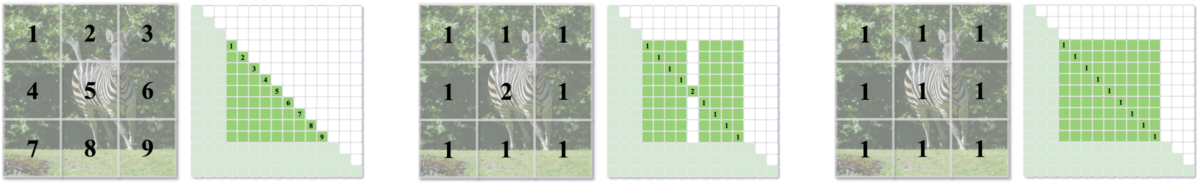
To this end, we present **Pyramid-descent Visual Position Encoding (PyPE)**, a novel position assignment approach for visual tokens, to alleviate the long-term decay induced by RoPE, avoid the "aggregation pattern" in the LLM, and ensure a comprehensive understanding of visual contents. PyPE reorganizes the flattened visual tokens into the 2D shape and assigns visual position indexes from the periphery to the center. This reduces the relative distance between interrelated visual elements, as well as the distance between significant visual elements and instruction tokens, thereby ensuring a more rational allocation of attention weights. Furthermore, to mitigate the impact of anchor tokens on the model's fine-grained perception of visual elements, we draw inspiration from Pyramid Vision Transformer (PVT) (Wang et al., 2021): consistently combining global and local receptive fields. PyPE gradually expands the central receptive field, *i.e.*, the central region of the position index matrix, at predetermined intervals of layers. Specifically, we expand the central region of the position index matrix by a circle every certain number of layers. Such expansion weakens the anchor tokens and enhances the model's ability to perceive visual elements at varying levels of granularity (more cases can be found in Section 5.4).

With extensive experiments on visual question answering and general multimodal benchmarks, PyPE consistently improves general perception capabilities across VLMs of different sizes. In a nutshell, the main contributions of this work are as follows: (I) We make an in-depth analysis of how position encoding affects visual perception in VLMs. (II) Our proposed PyPE effectively mitigates long-term decay and the "aggregation pattern", which helps better perceive visual elements at different granularities. (III) Extensive evaluations demonstrate the superior performance of PyPE, a simple yet effective method that applies to any VLMs.

2 Related Work

2.1 Vision-language Model

Recent advancements in VLMs have demonstrated impressive performance in processing multi-format information (Huang et al., 2023; Achiam et al., 2023). VLMs are typically built upon existing LLMs and incorporate visual information as input tokens by utilizing an additional vision encoder (e.g., CLIP) and a bridging connector (e.g., MLP). For instance, LLaVA (Liu et al., 2024a) employs an MLP to project visual tokens and aligns the feature dimensions with word embeddings, while BLIP-2 (Li et al., 2023b) utilizes a set of learnable query tokens to extract information in a query-



(a) Raster-scan PE. (b) Concentric PE. (c) All-One PE.

Figure 2: An overview of patch indexes and corresponding causal mask from raster-scan, concentric, and All-One position encoding on an example from COCO (Lin et al., 2014).

based manner. Building upon these foundational works, MM1 (McKinzie et al., 2025) has further investigated the significance of the number of visual tokens and image resolution, identifying them as the most critical factors, while finding that the type of connector has minimal impact. By effectively connecting visual and textual modalities, VLMs significantly enhance human-AI interaction and exhibit remarkable capabilities in understanding and generating multimodal content (Chen et al., 2024b; Peng et al., 2023; Chen et al., 2023; Wang et al., 2024; Hu et al., 2024; Xie et al., 2024).

2.2 Position Encoding for Transformers

Since transformer-based models contain no recurrence (Hochreiter, 1997) and convolution (Islam et al., 2020) structure, additional information about the relative or absolute position of the tokens in the input sequence is required. Therefore, the community has witnessed the development of various position encoding methods, e.g. sinusoidal (Vaswani, 2017), learnable (Dosovitskiy, 2020), relative (He et al., 2020; Shaw et al., 2018), and conditional (Chu et al., 2021) position encoding. Among these studies, RoPE (Su et al., 2024) is introduced to encode absolute and relative positional information, showing superiority in LLMs (Touvron et al., 2023; Achiam et al., 2023). The success of LLMs has led to the continued adoption of the effective RoPE scheme in VLMs for the unified encoding of positional information across sequences that incorporate multimodal features. However, it is important to note that visual information does not conform to the same sampling paradigm as language. The raster scanning is insufficient for modeling the spatial correlations among different patches. Consequently, numerous recent studies (Chu et al., 2024; Xing et al., 2024; Lu et al., 2024) have sought to explore improved solutions that extend RoPE to visual tasks. In this paper, we investigate a novel multi-granularity position assignment strategy to enhance the VLM's comprehension of visual information and improve the alignment between modalities.

3 Approach

3.1 Preliminaries

RoPE (Rotary Position Embedding) RoPE (Su et al., 2024) unifies both absolute and relative positional encodings, demonstrating a certain degree of extrapolation capability in LLMs and VLMs. Given the m-th query and n-th key vectors with a dimension D, denoted as $\mathbf{q}_m, \mathbf{k}_n \in \mathbb{R}^{|D|}$, RoPE multiplies a bias to the key or query vector in the complex vector space as follows:

$$f_q(\mathbf{q}_m, m) = e^{im\Theta}\mathbf{q}_m, \quad f_k(\mathbf{k}_n, n) = e^{in\Theta}\mathbf{k}_n \tag{1}$$

where $\Theta = \mathrm{Diag}(\theta_1, \cdots, \theta_{|D|/2})$ is the rotary frequency matrix, where $\theta_d = b^{-2d/|D|}$ and the rotary base $b = 10000$. In real space, for $l = |D|/2$, the rotary matrix $e^{im\Theta}$ can be expressed as:

$$\begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos m\theta_l & -\sin m\theta_l \\ 0 & 0 & \cdots & \sin m\theta_l & \cos m\theta_l \end{bmatrix} \tag{2}$$

The attention score using RoPE is calculated as follows:

$$\begin{aligned} A_n &= \mathrm{Re}(f_q(\mathbf{q}_m, m), f_k(\mathbf{k}_n, n)) \\ &= \mathrm{Re}(\mathbf{q}_m^\top e^{i(m-n)\Theta}\mathbf{k}_n) \end{aligned} \tag{3}$$

where $\mathrm{Re}(\cdot)$ is the real part of a complex number and $e^{i(m-n)\Theta} = (e^{im\Theta})^\top e^{in\Theta}$. As the relative distance $m - n$ increases, the attention score A_n correspondingly decreases due to long-term decay. This behavior aligns with the intuitive understanding that a pair of tokens separated by a significant relative distance should exhibit a weaker connection, and vice versa. However, a similar situation is observed in VLMs (Xing et al., 2024), which

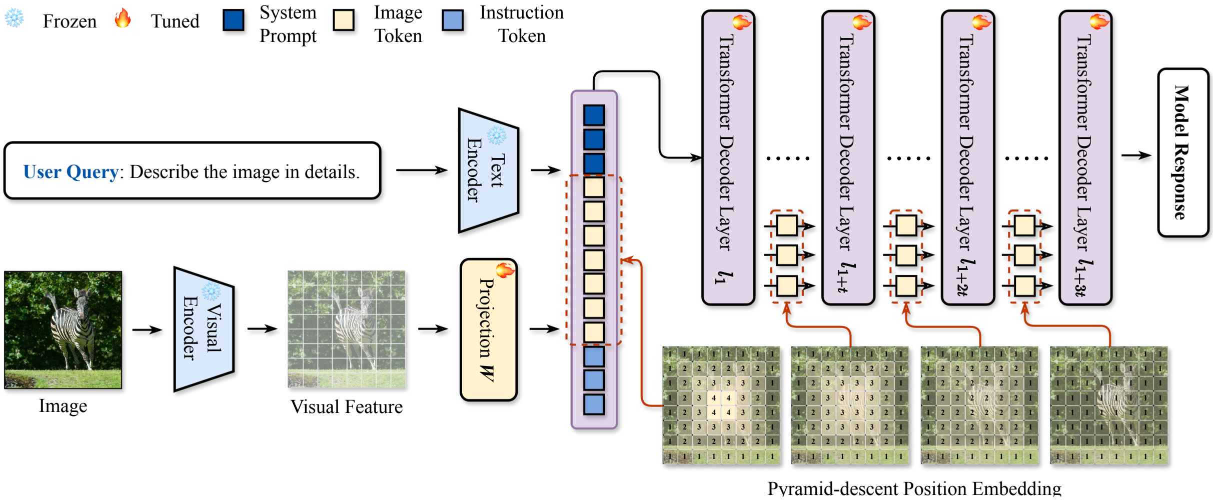

Frozen
Tuned
System Prompt
Image Token
Instruction Token
User Query: Describe the image in details.
Text Encoder
Image
Visual Encoder
Visual Feature
Projection W
Transformer Decoder Layer l_1
Transformer Decoder Layer l_{1+t}
Transformer Decoder Layer l_{1+2t}
Transformer Decoder Layer l_{1+3t}
Model Response
Pyramid-descent Position Embedding

Figure 3: **An overview of the proposed PyPE.** We first reorganize the visual tokens from their vanilla flattened 1D sequence form into the 2D format. Subsequently, we assign visual position indexes from the periphery to the center and expand the central receptive field incrementally across the layers with an interval of t.

can lead to the model lacking attention to patches that are relatively far from the instruction token obtained through raster scanning.

All-One Position Encoding To further explore the impact of visual position encoding on the model's perception of visual elements, we propose All-One Position Encoding: directly setting the relative distance between all image tokens and instruction tokens to 1. By doing so, the relative distances from all image tokens to the instruction token become equal, thereby excluding the influence of relative position decay introduced by RoPE. As a result, all patches are treated equally.

As indicated in Table 1, All-One PE performs weaker than the baselines in perception but keeps competitive in coarse-grained perception tasks on different sizes of models. This suggests that even when assigning the same positional weight to all image tokens, the VLM still possesses certain perception capabilities and performs better than raster-scan and concentric in coarse-grained situations. This is more pronounced on LLaVA-1.5-13B because larger models have stronger sequence modeling and feature capturing capabilities, which correspondingly bridge the gap in fine-grained abilities between All-One PE and other methods.

3.2 Pyramid-descent Visual Position Encoding

Though presenting competitive coarse-grained perception capabilities, All-One PE still falls short in fine-grained perception. Using identical position weights hampers the model's ability to differentiate the significance of image tokens, while the positional priors introduced by raster scanning conflict with general cognitive principles.

Similar challenges were also present in the early development of Vision Transformer (ViT) (Dosovitskiy, 2020). Due to the columnar structure of ViT, which uses coarse image patches as input, it is difficult to apply it directly to pixel-level dense predictions such as object detection and segmentation. This difficulty arises because its output feature map is single-scale and low-resolution. To address these issues, Wang et al. (2021) proposed the Pyramid Vision Transformer (PVT). They utilize fine-grained image patches as input to learn high-resolution representations and introduce a progressive shrinking pyramid to reduce the sequence length of the Transformer as the network deepens, significantly lowering the computational cost. Moreover, compared to CNNs, PVT consistently produces a global receptive field, ensuring a holistic perception of visual elements and benefiting its performance in detection and segmentation tasks.

In light of this, we propose the **Pyramid-descent Visual Position Encoding (PyPE)**, a simple yet effective position assignment strategy for visual tokens in VLMs. As shown in Figure 3, we first reorganize the visual tokens from their vanilla flattened 1D sequence form into the 2D format. Subsequently, we adopt a decay pattern for the corresponding position indexes of the image tokens that spread outward from the center following concentric PE (Xing et al., 2024). Given the maximum assignable position index $\mathcal{P}_{max}$, the position as-

Method	Perception	Commonsense QA (Reasoning)	Coarse-grained Perception Tasks				Total Scores
			Existence	Count	Position	Color	
TinyLLaVA-SigLIP-Phi-2							
w/ Raster-scan	1488.30	**120.71**	185.00	143.33	133.33	**180.00**	762.37
w/ Concentric	1465.25	114.29	185.00	**160.00**	131.67	170.00	760.96
w/ All-One	1489.46	117.14	**190.00**	158.33	133.33	175.00	**773.80**
***w/* PyPE (Ours)**	**1500.66**	115.00	**190.00**	150.00	**138.33**	**180.00**	773.33
LLaVA-1.5-7B							
w/ Raster-scan	1510.72	**135.71**	190.00	**158.33**	128.33	175.00	787.37
w/ Concentric	1485.67	120.71	190.00	153.33	133.33	170.00	767.37
w/ All-One	1474.13	131.43	**195.00**	148.33	141.67	175.00	791.43
***w/* PyPE (Ours)**	**1542.19**	130.00	**195.00**	155.00	**146.67**	**180.00**	**806.67**
LLaVA-1.5-13B							
w/ Raster-scan	1581.45	**132.14**	190.00	155.00	135.00	**195.00**	807.14
w/ Concentric	1607.40	128.57	**195.00**	**180.00**	141.67	185.00	830.24
w/ All-One	1608.12	130.00	**195.00**	170.00	146.67	190.00	831.67
***w/* PyPE (Ours)**	**1629.41**	130.71	190.00	**180.00**	**153.33**	180.00	**834.04**

Table 1: Performance evaluation on MME. *Existence*, *Count*, *Position*, and *Color* are coarse-grained subtasks of MME-Perception, while *Commonsense QA* is a subtask of MME-Cognition. *Total Scores* denotes the sum of the results from *Commonsense QA* and *Coarse-grained tasks*. The best results in each setting are in **bold.**

Algorithm 1 Pyramid-descent Visual Position Encoding

```
INPUT: Height $H$, width $W$, descent interval $t$, current layer index $i$, current $\mathcal{P}_{max}$.
OUTPUT: Pyramid-descent position assignment matrix $\mathcal{P}$, causal mask $\mathcal{M}$ and $\mathcal{P}_{max}$ for the next layer.
if $i \mod t == 0$ $and$ $\mathcal{P}_{max} > 1$ then
    $\mathcal{P}_{max} \leftarrow \mathcal{P}_{max} - 1$
end if
Initialize $\mathcal{P}$.
for $p$ $in$ $[1, \mathcal{P}_{max}]$ do
    $\mathcal{P}[p : H - p, p : W - p] \leftarrow p$
end for
Generate $\mathcal{M}$ according to $\mathcal{P}$.
```

signment matrix $\mathcal{P}$ is calculated as follows,

$$\begin{aligned} \mathcal{P}(i, j) = p, \;\; \forall p \in [1, \mathcal{P}_{max}]\,, \\ s.t.\;\; \{(i,j) \mid i \in [p, H-p),\; j \in [p, W-p)\}, \end{aligned} \tag{4}$$

where H and W represent the height and width of the input image, respectively. $\mathcal{P}_{max}$ is initialized to $\lfloor H/2 \rfloor$. This design maintains spatial continuity in the row and column dimensions. It reduces the average distance between significant image tokens and instruction tokens, facilitating cross-attention among the image tokens and cross-attention between the image tokens and instruction tokens.

Subsequently, we propose a gradual expansion of the central receptive field to diminish the influence of anchor tokens and enhance the model's ability to perceive visual elements at varying levels of granularity. Specifically, we reduce $\mathcal{P}_{max}$ every t layers, thereby controlling the granularity of perception through position encoding. When $\mathcal{P}_{max}$ is reduced to 1, the corresponding position encoding transforms into an All-One PE, which perceives more coarse-grained elements. To maintain causal attention, we adjust the attention mask $\mathcal{M}$ based on each assigned position matrix $\mathcal{P}$.

By introducing hierarchical position indices, PyPE facilitates multi-granularity perception of visual elements, allowing the model to dynamically adjust its focus to capture both broad contextual information and fine-grained details within visual data. This innovative approach not only aligns more closely with human cognitive processes but also enhances the model's overall performance in tasks that require both holistic and detailed perception of visual content.

4 Experiment Setup

4.1 Benchmarks

We evaluate PyPE on visual question answering and general multimodal benchmarks, including VQAv2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019), GQA (Hudson and Manning, 2019), VizWizQA (Bigham et al., 2010), TextVQA (Singh

Method	**VQAv2**	**OK-VQA$_{\text{val}}$**	**GQA**	**VizWizQA**	**TextVQA$_{\text{val}}$**	**RealWorldQA**	**ScienceQA^{I}**
TinyLLaVA-SigLIP-Phi-2							
w/ Raster-scan	78.93	56.71	61.07	34.30	50.88	53.99	71.24
w/ Concentric	79.08	57.35	61.15	41.08	50.77	53.59	70.45
w/ All-One	78.89	57.34	61.33	42.50	50.94	53.59	70.55
***w/* PyPE (Ours)**	**79.22**	**57.48**	**61.65**	**44.45**	**51.31**	**54.12**	**71.59**
LLaVA-1.5-7B							
w/ Raster-scan	78.56	54.32	62.12	50.34	46.16	54.80	66.80
w/ Concentric	79.02	52.70	62.28	52.52	45.84	54.77	68.72
w/ All-One	79.02	52.50	62.00	**55.32**	45.98	54.77	68.32
***w/* PyPE (Ours)**	**79.15**	**54.96**	**62.71**	53.11	**46.73**	**55.42**	**69.51**
LLaVA-1.5-13B							
w/ Raster-scan	79.14	**57.38**	63.34	53.75	48.56	55.95	71.15
w/ Concentric	79.90	53.81	63.26	56.38	48.07	55.42	70.00
w/ All-One	**79.95**	51.40	63.34	56.37	48.15	54.64	71.39
***w/* PyPE (Ours)**	**79.95**	55.66	**63.52**	**58.06**	**48.90**	**56.86**	**71.54**

Table 2: Performance evaluation on visual question answering. We utilize *accuracy* as the evaluation metric. OK-VQA$_{\text{val}}$ and TextVQA$_{\text{val}}$ denote the validation set of OK-VQA and TextVQA, respectively. ScienceQA^{I} denote the image subset of ScienceQA. The best results in each setting are in **bold.**

et al., 2019), RealWorldQA (X.AI, 2024), ScienceQA (Lu et al., 2022), MME (Yin et al., 2024), MMBench (Liu et al., 2025), SEED-Bench (Li et al., 2023a), POPE (Li et al., 2023c), AI2D (Kembhavi et al., 2016), MM-Vet (Yu et al., 2023), MMMU (Yue et al., 2024), MMT-Bench (Ying et al., 2024), and MMStar (Chen et al., 2024a). Refer to Appendix A for more details.

4.2 Implementation Details

To demonstrate the generalizability of our proposed method across models with different parameter sizes, we conduct experiments using three model architectures with 3B, 7B, and 13B parameters. For 3B models, we follow TinyLLaVA (Zhou et al., 2024) to use SigLIP (Zhai et al., 2023) as the visual encoder and Phi-2 (Li et al., 2023d) as the base LLM. For 7B and 13B models, we adopt pre-trained CLIP ViT-L/14 (336^2) (Radford et al., 2021) as visual encoder and Vicuna v1.5 (Zheng et al., 2023) as the base LLM. Following Liu et al. (2024a), we pretrain the models on CC-558K dataset and finetune them on the mix-665K dataset. All experiments are conducted on 8 NVIDIA A100 and 8 NVIDIA H20 GPUs. See Appendix B for more training and implementation details.

5 Empirical Results and Analysis

We evaluate the visual capabilities of the models trained with the PyPE through various visual question answering and general multimodal benchmarks. This novel position encoding demonstrates highly competitive performance at different scales. Our proposed method consistently delivers top-tier performance across most evaluation metrics, frequently surpassing other baselines.

5.1 Results of Visual Question Answering Benchmarks

To rigorously evaluate the capabilities of our models in general visual question answering tasks, we conduct comprehensive assessments across a diverse array of state-of-the-art benchmarks. The results presented in Tables 1 and 2 indicate that the PyPE series demonstrates exceptional performance across all benchmarks, with the three variants consistently achieving or surpassing baseline performance. In the MME benchmark, PyPE exhibits a superior understanding of visual content at various levels of granularity. It retains a coarse-grained perception capability comparable to that of All-One PE while outperforming both Raster-scan and Concentric PE in terms of fine-grained perception. On the RealWorldQA benchmark, which assesses real-world spatial comprehension, PyPE achieves scores of 54.12, 55.42, and 56.86 for the 3B, 7B, and 13B variants, respectively. These results exceed all baseline performances and reflect an enhanced understanding of physical environments. VizWizQA is a dataset comprising images captured by visually impaired individuals using mobile phones, accompanied by recorded spoken questions. The images in this dataset tend to exhibit relatively low clarity, with subjects occupying a significant portion of the

Method	POPE			MMBench		SEED^{I}	AI2D	MM-Vet	MMMU	MMT-Bench	MMStar
	rand	pop	adv	en	cn						
TinyLLaVA-SigLIP-Phi-2											
w/ Raster-scan	88.50	86.93	85.60	67.88	**45.07**	68.54	59.75	33.00	**33.80**	**48.93**	37.37
w/ Concentric	88.63	87.27	85.67	67.83	43.22	68.51	60.98	33.40	33.60	48.86	38.44
w/ All-One	88.53	87.40	86.00	66.48	43.11	68.25	61.20	32.70	**33.80**	48.00	38.06
***w/* PyPE (Ours)**	**89.07**	**87.70**	**85.73**	**68.33**	43.95	**68.55**	**61.53**	**35.00**	33.70	**48.93**	**38.89**
LLaVA-1.5-7B											
w/ Raster-scan	**88.33**	87.13	85.63	64.97	57.90	66.10	55.25	30.80	31.00	47.94	35.25
w/ Concentric	87.83	86.40	85.43	65.41	57.79	66.31	54.83	29.70	31.00	49.02	35.41
w/ All-One	87.30	86.57	85.53	65.47	55.89	66.41	54.73	29.90	30.70	48.99	36.24
***w/* PyPE (Ours)**	88.27	**87.43**	**85.67**	**66.65**	**58.63**	**67.01**	**55.63**	**31.10**	**31.10**	**49.70**	**36.51**
LLaVA-1.5-13B											
w/ Raster-scan	**88.77**	**87.70**	**85.90**	67.74	63.17	67.65	59.49	**37.30**	**33.20**	49.82	36.81
w/ Concentric	87.90	87.13	85.80	68.89	62.67	67.59	58.55	35.90	32.70	48.54	37.33
w/ All-One	87.93	87.13	85.77	67.99	63.06	67.47	58.84	36.00	32.90	49.38	37.32
***w/* PyPE (Ours)**	88.03	86.97	85.47	**69.23**	**63.45**	**68.50**	**59.59**	36.60	**33.20**	**50.40**	**38.71**

Table 3: Evaluation on general multimodal benchmarks. We utilize *accuracy* as the evaluation metric. SEED^{I} denotes the image subset of SEED-Bench. The best results in each setting are in **bold.**

Method	MME^{P}	OK-VQA$_{val}$	TextVQA$_{val}$	MMStar
TinyLLaVA-SigLIP-Phi-2				
PyPE 1x	1479.53	56.99	50.13	37.31
PyPE 2x	**1500.66**	**57.48**	**51.31**	**38.89**
PyPE 3x	1470.45	57.29	50.28	38.32
PyPE 4x	1466.70	55.89	50.59	37.23
LLaVA-1.5-7B				
PyPE 1x	1507.19	52.73	**46.77**	34.82
PyPE 2x	**1542.19**	**54.96**	46.73	**36.51**
PyPE 3x	1503.95	52.87	45.99	36.18
PyPE 4x	1497.18	51.76	46.20	35.79
LLaVA-1.5-13B				
PyPE 1x	1608.01	50.53	48.60	35.89
PyPE 2x	**1629.41**	55.66	**48.90**	**38.71**
PyPE 3x	1583.84	54.90	48.52	36.55
PyPE 4x	1607.63	**57.42**	48.09	37.07

Table 4: Analysis of the descent interval t. PyPE tx denotes using PyPE with interval t. MME^{P} denotes MME-Perception.

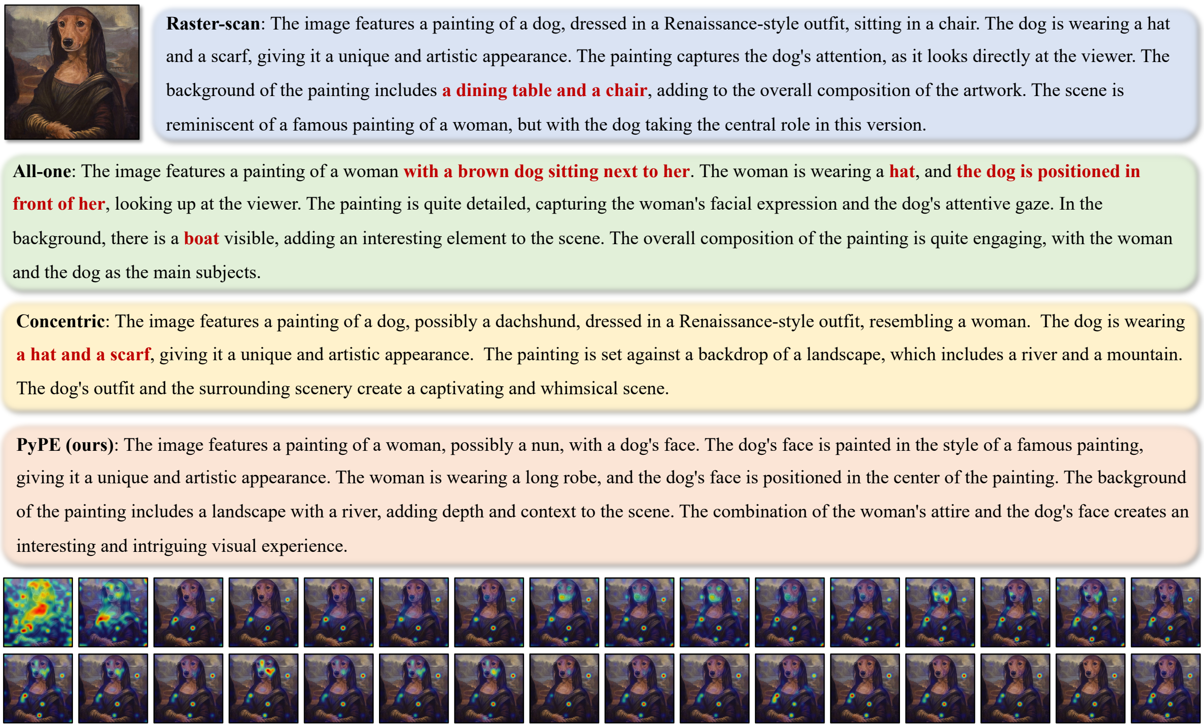
frame. Consequently, as shown in Table 2, All-One PE demonstrates competitive performance on this dataset, while our proposed PyPE exhibits superior zero-shot performance on both VizWizQA and ScienceQA. This improvement can be attributed to the flexible receptive field enabled by PyPE.

5.2 Results of General Multimodal Benchmarks

As illustrated in Table 3, the PyPE series demonstrates exceptional performance on mainstream general multimodal benchmarks. In the MMStar benchmark, which is designed to assess genuine multimodal capabilities using visually indispensable samples, PyPE outperforms all baseline models. On MM-Vet, which evaluates the integration of core vision-language capabilities across 16 complex multimodal tasks, the 3B model of PyPE achieves an impressive score of 35.00, significantly surpassing the scores of 33.00 and 33.40 obtained by Raster-scan and Concentric PE, respectively. In the MMT-Bench evaluation, which assesses advanced reasoning and instruction-following across 32 core meta-tasks and 162 subtasks in multimodal understanding, PyPE markedly exceeds baseline performance, demonstrating its ability to apply expert knowledge and execute deliberate visual recognition, localization, reasoning, and planning. On MMBench, which evaluates fine-grained abilities across 20 dimensions, PyPE exhibits strong performance, matching or leading the state-of-the-art. Additionally, we test the methods on AI2D, a benchmark focusing on multiple-choice questions related to scientific diagrams containing text. The results indicate that PyPE achieves state-of-the-art performance and demonstrates a strong comprehension of textual content within images.

5.3 Analysis of the Descent Interval

As shown in Table 4, we evaluate the performance of different models using PyPE with varying descent intervals on VQA and general multimodal benchmarks. Across all models, a moderate descent interval PyPE 2x generally provides the best or near-best performance, which strikes a bal-



Raster-scan: The image features a painting of a dog, dressed in a Renaissance-style outfit, sitting in a chair. The dog is wearing a hat and a scarf, giving it a unique and artistic appearance. The painting captures the dog's attention, as it looks directly at the viewer. The background of the painting includes **a dining table and a chair**, adding to the overall composition of the artwork. The scene is reminiscent of a famous painting of a woman, but with the dog taking the central role in this version.

All-one: The image features a painting of a woman **with a brown dog sitting next to her**. The woman is wearing a **hat**, and **the dog is positioned in front of her**, looking up at the viewer. The painting is quite detailed, capturing the woman's facial expression and the dog's attentive gaze. In the background, there is a **boat** visible, adding an interesting element to the scene. The overall composition of the painting is quite engaging, with the woman and the dog as the main subjects.

Concentric: The image features a painting of a dog, possibly a dachshund, dressed in a Renaissance-style outfit, resembling a woman. The dog is wearing **a hat and a scarf**, giving it a unique and artistic appearance. The painting is set against a backdrop of a landscape, which includes a river and a mountain. The dog's outfit and the surrounding scenery create a captivating and whimsical scene.

PyPE (ours): The image features a painting of a woman, possibly a nun, with a dog's face. The dog's face is painted in the style of a famous painting, giving it a unique and artistic appearance. The woman is wearing a long robe, and the dog's face is positioned in the center of the painting. The background of the painting includes a landscape with a river, adding depth and context to the scene. The combination of the woman's attire and the dog's face creates an interesting and intriguing visual experience.

Figure 4: Illustration of the multi-granularity perception capability of PyPE with a sample from LLaVA-Bench. The case study is based on LLaVA-1.5-7B and the query is *"Describe this photo in detail"*. The misunderstandings and hallucinations of visual contents are highlighted in **red**. We also provide a corresponding layer-wise attention visualization of PyPE, with the heatmap arranged from the upper left to the lower right, indicating layers 1 to 32.

ance between the model's ability to handle perception (MME), external knowledge integration (OK-VQA), text comprehension (TextVQA), and vision-critical tasks (MMStar). While the 2x interval is generally optimal, there are exceptions, such as the LLaVA-1.5-13B model performing best on OK-VQA with a 4x interval. This indicates that larger models might benefit from longer intervals for specific tasks.

5.4 Qualitative Results on LLaVA-Bench

Figure 4 demonstrates a case study on how, given identical prompts and images, other baselines misperceive or inadequately process visual information, resulting in the generation of hallucinatory content. For instance, in the displayed example, the baseline methods exhibit object hallucinations, identifying nonexistent items such as "*dining table*", "*hat*", "*scarf*", and "*boat*". In contrast, the implementation of PyPE notably mitigates these hallucination issues while simultaneously maintaining the coherence and informativeness of the output text. This can be attributed to the multi-scale visual modeling capability afforded by the dynamic local receptive fields of PyPE, in conjunction with the stable global receptive fields. Furthermore, the visualization results of layer-wise attention indicate that our proposed method effectively alleviates the phenomenon of "aggregation pattern", thereby creating a synergistic effect with the former. Refer to Appendix C for a more in-depth analysis of anchor tokens and Appendix E for more case studies.

6 Conclusion

In this work, we conduct an in-depth analysis of how visual position encoding affects visual perception in VLMs (Vision-lnaguage Models), particularly from the aspect of long-term decay and the "aggregation pattern" (also known as "anchor tokens"). Our findings indicate that conventional visual position encoding methods are constrained by the "aggregation pattern" derived from LLMs (Large Language Models) and lack multi-scale perceptual capabilities. To address these limitations, we introduce Pyramid-descent Visual Position Encoding (PyPE), a novel approach designed to enhance the perception of visual tokens within VLMs. Extensive experiments across multiple benchmarks and VLM families demonstrate the efficacy of PyPE in addressing these challenges and ensuring a thorough understanding of visual content.

Limitations

Although PyPE demonstrates exceptional performance in enhancing the overall capabilities of Vision-language Models (VLMs), it is currently limited to single-frame images and has not yet been extended to video and other modalities. Future research will focus on effectively integrating the temporal dimension for unified position encoding and extending PyPE to a broader range of VLMs.

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*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024b. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. 2024. Visionllama: A unified llama interface for vision tasks. *arXiv preprint arXiv:2403.00522*.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. 2021. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*.

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

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*.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Md Amirul Islam, Sen Jia, and Neil DB Bruce. 2020. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

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, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining 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. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023d. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

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.

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. 2024. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. 2025. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

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*.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578.

X.AI. 2024. Grok-1.5 vision preview. `https://x.ai/blog/grok-1.5v`.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. 2024. Gpa: Global and prototype alignment for audio-text retrieval. In *Interspeech 2024*, pages 5078–5082.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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.

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

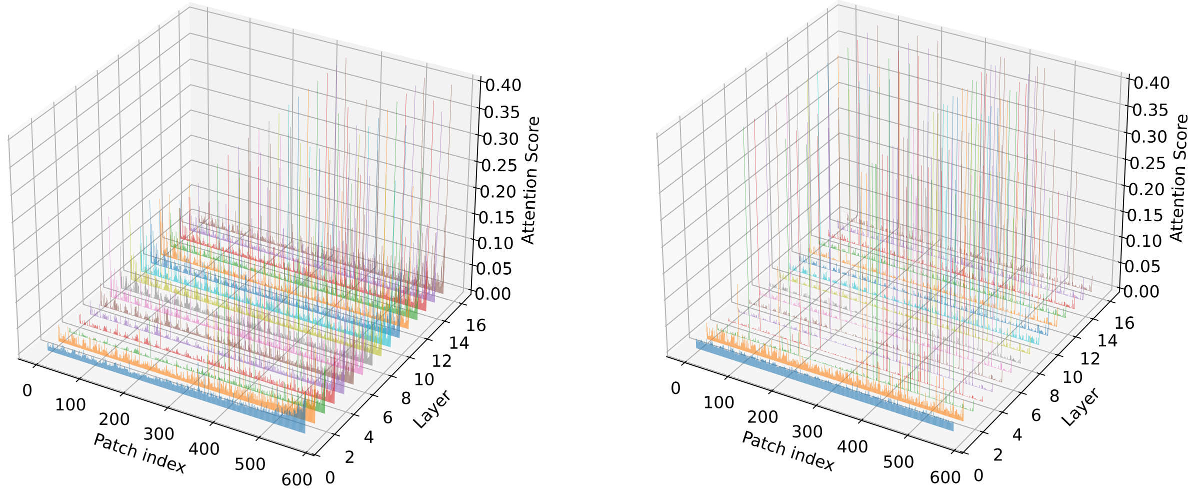
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

A Benchmarks

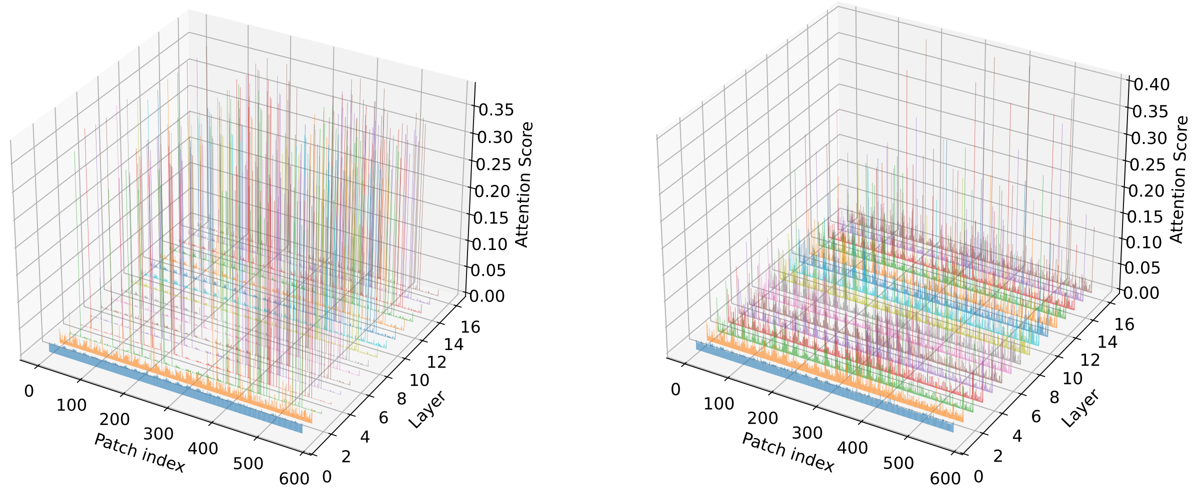
Visual Question Answering The VQAv2 dataset is currently the largest available dataset for visual question answering. OK-VQA includes questions that necessitate external knowledge beyond the multimodal inputs provided. GQA is specifically designed to assess the reasoning capabilities of the model. VizWizQA is composed of question-answer pairs derived from visually impaired users. TextVQA places a greater emphasis on evaluating the model's ability to comprehend text within natural scenes. RealWorldQA is a benchmark specifically designed to evaluate the spatial understanding capabilities of multimodal AI models in real-world contexts. ScienceQA comprises multimodal multiple-choice questions across a diverse range of science topics. These datasets are strategically selected to comprehensively evaluate our method's capacity to understand and reason across diverse visual contexts and knowledge domains.

General Multimodal Benchmarks MME measures both perception and cognition abilities on a total of 14 subtasks. MMBench comprehensively evaluates a model's multimodal capabilities in both Chinese and English contexts. SEED-Bench focuses on assessing generative comprehension in Vision-language Models. POPE evaluates the extent of multimodal hallucinations present in a model. AI2D assesses a model's ability to interpret scientific diagram inputs. MM-Vet evaluates the multimodal conversational abilities of a model using GPT-4 as a benchmark. MMMU is designed to assess multimodal models on extensive



(a) Raster-scan PE.

(b) All-One PE.



(c) Concentric PE.

(d) PyPE.

Figure 5: Visualization of anchor tokens in baselines and PyPE.

multi-disciplinary tasks that require college-level subject knowledge and deliberate reasoning. MMT-Bench is a comprehensive benchmark developed to evaluate VLMs across a wide range of multimodal tasks that necessitate expert knowledge and deliberate visual recognition, localization, reasoning, and planning. MMstar is a premier, vision-critical multimodal benchmark comprising 1,500 challenge samples meticulously curated by human experts.

B Hyperparameters and More Implementation Details

We show the training hyperparameters for both first-stage vision-language alignment pretraining and the second-stage visual instruction tuning in Table 5. We use LMMs-Eval (Zhang et al., 2024) to conduct experiments on VQA and general multimodal benchmarks.

Hyperparameter	Pretrain	Finetune
batch size	256	128
lr	1e-3	2e-5
lr schedule	cosine decay	
lr warmup ratio	0.03	
weight decay	0	
epoch	1	
optimizer	AdamW	
DeepSpeed stage	2	3

Table 5: **Hyperparameters** of TinyLLaVA-SigLIP-Phi-2 and LLaVA-1.5-7B/13B.

C Visualization of Anchor Tokens

To further analyze the aggregating attention pattern, we visualize the attention score of each patch in the first 16 layers. As illustrated in Figure 5, both the All-One PE and the Concentric PE ex-

Method	**RefCOCO**			**RefCOCO+**			**RefCOCOg**	
	val	test-A	test-B	val	test-A	test-B	val	test
TinyLLaVA-SigLIP-Phi-2								
w/ Raster-scan	**31.85**	15.77	44.29	31.28	18.65	40.83	56.86	56.98
w/ Concentric	30.89	15.97	42.22	31.16	19.31	38.77	59.45	58.74
w/ All-One	29.13	15.24	39.11	30.64	18.38	38.34	54.81	55.28
***w/* PyPE (Ours)**	31.33	**16.02**	**45.13**	**31.86**	**19.40**	**42.25**	**59.72**	**59.79**
LLaVA-1.5-7B								
w/ Raster-scan	34.19	**18.07**	46.89	**34.30**	**21.63**	**43.53**	61.21	59.40
w/ Concentric	32.23	16.51	42.49	32.66	20.00	40.41	59.72	58.47
w/ All-One	32.99	16.46	41.26	33.28	20.73	39.83	63.07	61.90
***w/* PyPE (Ours)**	**35.16**	16.51	**48.04**	34.17	21.22	41.46	**64.62**	**64.13**
LLaVA-1.5-13B								
w/ Raster-scan	36.86	19.29	50.01	36.12	22.37	43.59	**63.66**	60.96
w/ Concentric	35.87	18.54	48.17	36.07	21.94	42.65	61.66	61.07
w/ All-One	36.84	19.06	49.16	37.10	22.71	41.72	61.58	59.75
***w/* PyPE (Ours)**	**37.81**	**21.82**	**51.88**	**37.14**	**25.74**	**44.73**	63.16	**62.59**

Table 6: Performance comparison on referring expression comprehension tasks. We use CIDEr (Vedantam et al., 2015) to evaluate the quality of the descriptions. The highest results in each setting are indicated in **bold**, while the second-best results are underlined.

hibit a relatively uniform distribution of attention in the initial two layers. However, a significant phenomenon of attention aggregation emerges in the subsequent layers, where non-anchor patches demonstrate a suppression of attention, particularly pronounced in Concentric PE. Though Raster-scan PE shows slight improvement, the attention in each layer tends to be preferentially allocated to patches that are closer to the instruction token, resulting in a discontinuous and fragmented attention pattern. This indicates a limitation of the Raster-scan PE in effectively modeling patches with similar semantics. In contrast, PyPE not only reduces the number of anchor tokens but also yields significantly lower attention scores for these tokens compared to the baselines, thereby facilitating the model's exploration of image details more effectively. Furthermore, in each layer, the attention distribution of the PyPE is more continuous, highlighting the superiority of our proposed method in modeling semantically similar information.

D Performance on Referring Expression Comprehension

In the context of the visual localization task, we evaluate PyPE using the RefCOCO, RefCOCO+, and RefCOCOg datasets (Kazemzadeh et al., 2014; Mao et al., 2016). The results, presented in Table 6, indicate that PyPE achieves top-tier performance among baselines. Its superior structural design enables PyPE to effectively perceive intricate details within images, resulting in significant improvements over baseline models. The performance of PyPE underscores its potential to advance the field of visual localization and its applicability in real-world scenarios that require precise visual understanding.

E More Case Studies

We provide more examples of visual description in Table 7. As illustrated in the table, our proposed PyPE exhibits a reduced incidence of generating visual hallucinations or misunderstandings. More importantly, compared to other baseline methods, PyPE demonstrates a finer granularity in perceiving visual elements, thereby uncovering additional information, such as "blueberries" in the first example and "My joke website (funny joke push to reveal punchline)" in the second example. To further analyze the model's attention distribution across each decoder layer, we visualize the corresponding attention values for these examples. The results in Figure 6, 7, and 8 indicate that while other baselines remain hindered by anchor tokens, PyPE consistently mitigates this issue, facilitating a more rational allocation of attention.

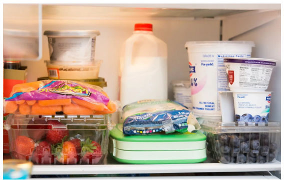
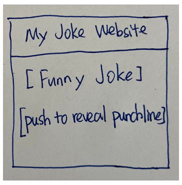
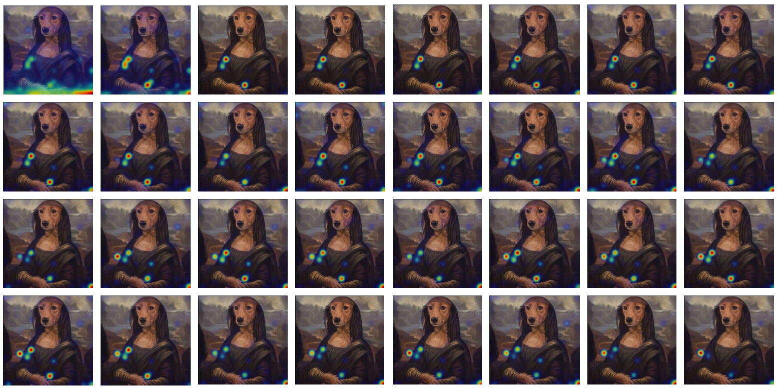
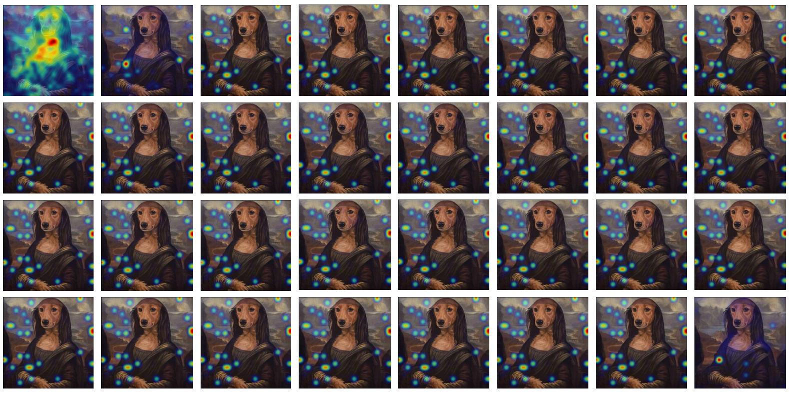
Visual input example, Visual Description 1	
User	 Describe this photo in detail.
Raster-scan	The image displays a refrigerator filled with various food items, including milk, juice, and fruits. There are **multiple bottles of milk** placed throughout the refrigerator, with some located **on the top shelves and others on the bottom shelves**. In addition to the milk, there are **several bottles of juice, some of which are placed on the top shelves and others on the bottom shelves**. A few carrots can be seen in the refrigerator, with some placed on the top shelves and others on the bottom shelves. The refrigerator is well-stocked with a variety of food items, making it a convenient place to store and access groceries.
All-One	The image displays a refrigerator filled with various food items. Inside the refrigerator, there are several bottles, including **one on the left side, two in the middle, and one on the right side**. There are also multiple carrots, with some placed in the middle and others on the left side of the refrigerator. In addition to the carrots, there are several strawberries, with some located on the left side and others **on the right side of the refrigerator.** A bowl can be seen on the left side of the refrigerator, possibly containing more food items. The refrigerator is well-stocked with a variety of fruits and vegetables, making it a healthy and diverse selection.
Concentric	The image displays a refrigerator filled with various food items, including a variety of fruits and vegetables. There are several carrots placed throughout the refrigerator, with some in the top left, middle, and bottom sections. A few strawberries can be seen in the bottom left corner of the refrigerator. In addition to the fresh produce, there are **multiple bottles of milk and juice** stored in the refrigerator. Some of these bottles are located in the **top left, middle, and bottom** sections of the refrigerator. The refrigerator is well-stocked with a diverse selection of food items, making it a convenient place to store and access groceries.
PyPE (ours)	The image displays a refrigerator filled with various food items, including several containers of berries, milk, and other foods. There are **multiple bottles of milk** placed throughout the refrigerator, with some located on the top shelves and others on the bottom shelves. In addition to the milk, there are several containers of berries, such as strawberries and **blueberries**, spread across the refrigerator. The refrigerator is well-stocked with a diverse range of food items, making it a convenient place to store and access groceries.
Visual Description 2	
User	 My Joke Website [Funny Joke] [push to reveal punchline] Describe this photo in detail.
Raster-scan	The image features a handwritten sign with a joke written on it. **The sign is placed on a table**, and it appears to be a website or a piece of paper with a humorous message. The joke is written in a **foreign language**, which adds a unique touch to the sign. The sign is placed in the center of the image, drawing attention to its content.
All-One	The image features a handwritten sign with a joke website written on it. The sign is placed on a white surface, possibly a piece of paper or a wall. The joke website is written in blue ink, and **the text is accompanied by a question mark, indicating that the joke is a riddle or a puzzle**. The sign is placed in the center of the image, drawing attention to the joke website.
Concentric	The image features a handwritten sign with a joke on it, likely meant to be displayed on a website. The joke is written in a **foreign language, and the sign is accompanied by a question mark, indicating that it is meant to be a humorous or amusing statement**. The sign is placed on a white background, which contrasts with the dark blue writing.
PyPE (ours)	The image features a handwritten sign with a joke on it, likely meant to be displayed on a website. The joke is written in blue ink, and it reads, **"My joke website (funny joke push to reveal punchline)."** The sign is placed on a white background, which contrasts with the blue ink used for the joke. The handwriting style suggests a casual and humorous tone, likely meant to entertain or amuse the viewers.

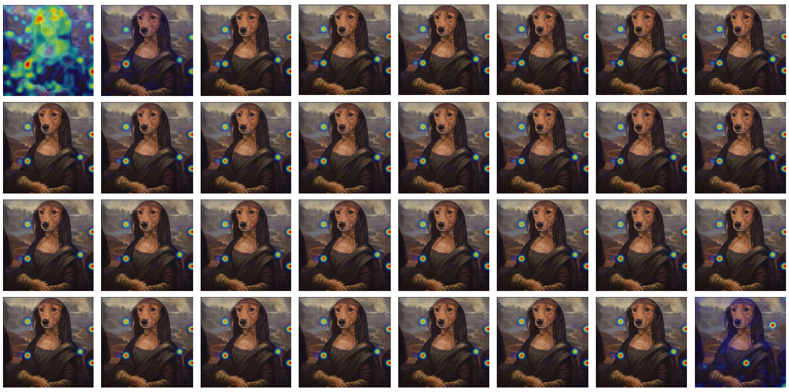
Table 7: More examples from LLaVA-Bench. The misunderstandings and hallucinations of visual contents are highlighted in **red**. The descriptions that are not mentioned in baselines but are accurately represented by PyPE are highlighted in **green**.



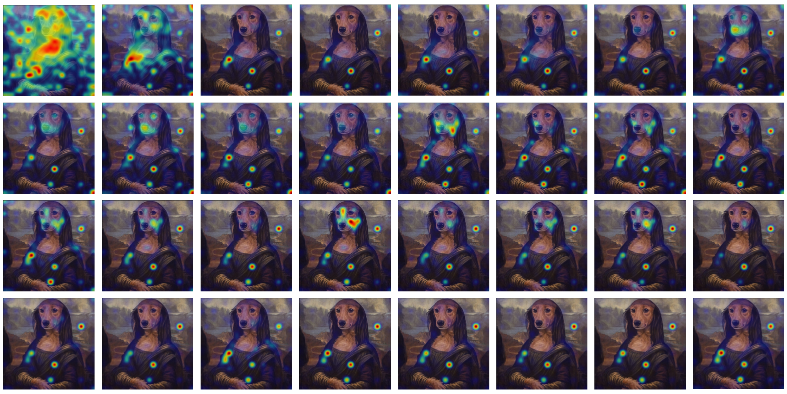
(a) Raster-scan



(b) Concentric

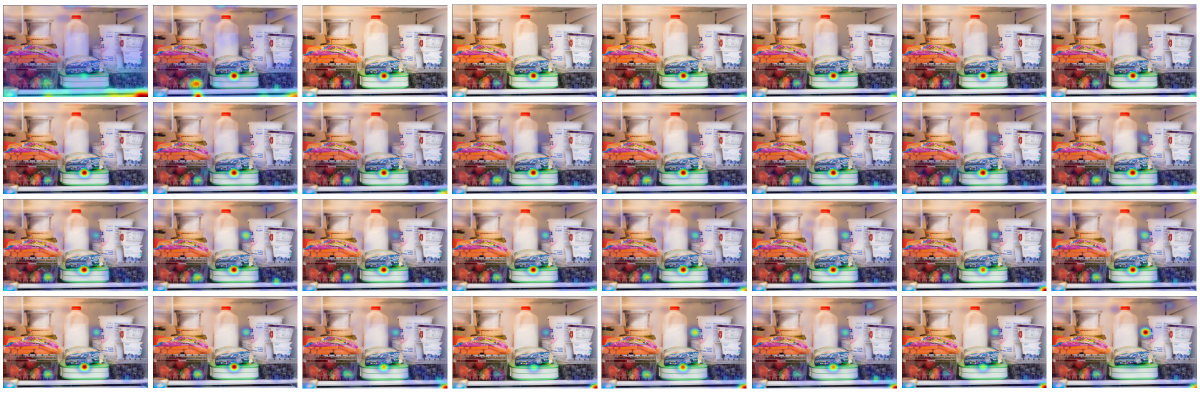


(c) All-One

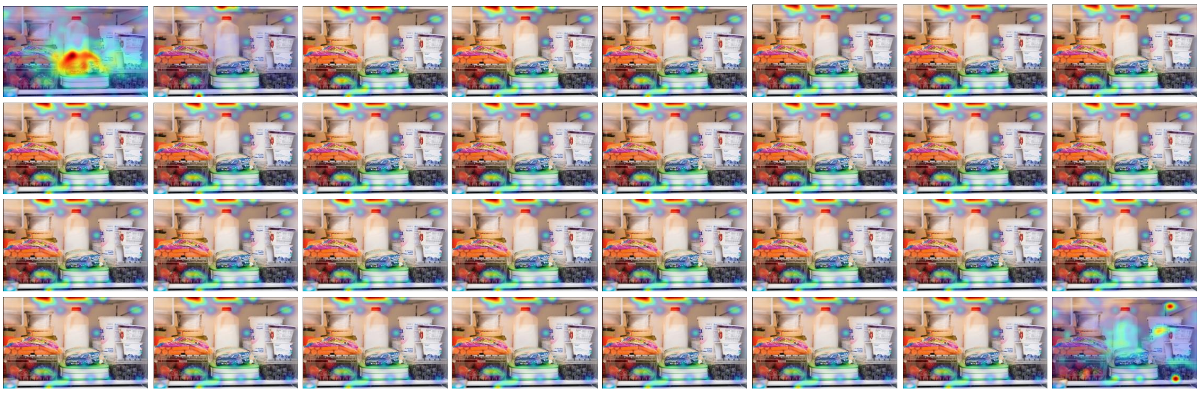


(d) PyPE

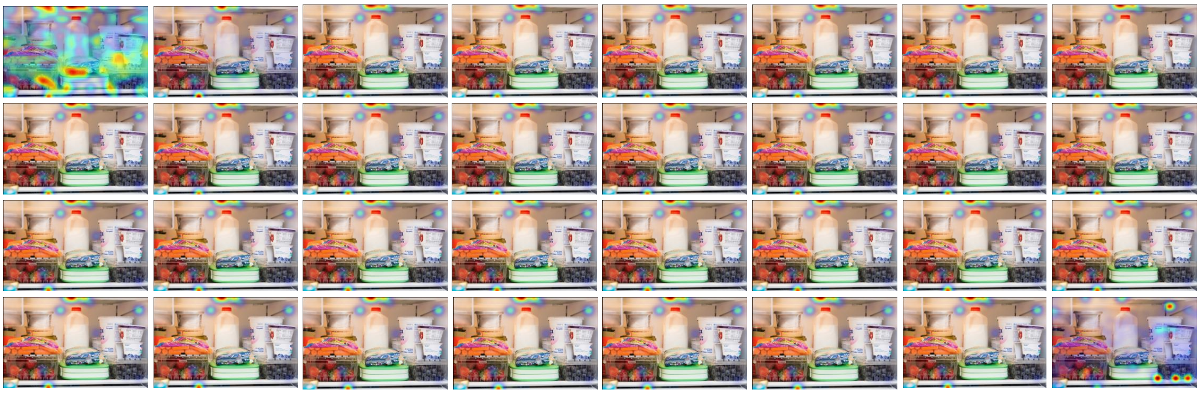
Figure 6: Layer-wise attention visualization (left to right, up to down) of the example from Figure 4.



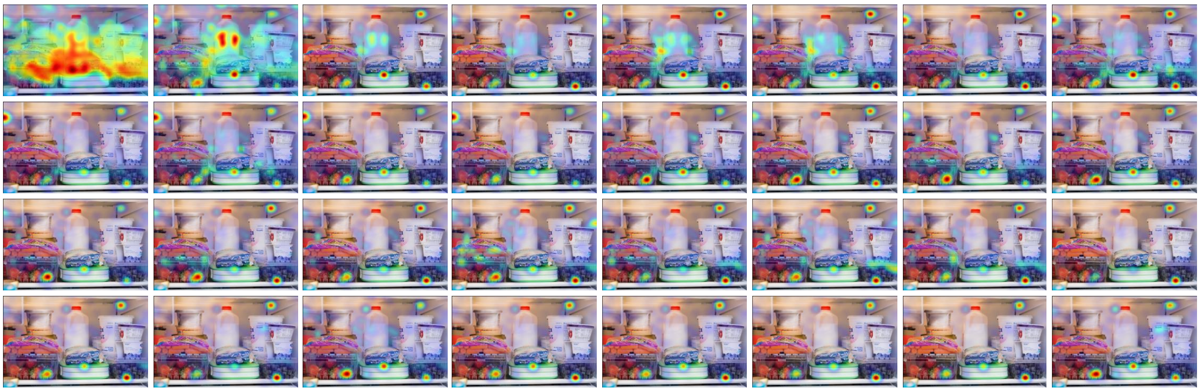
(a) Raster-scan



(b) Concentric

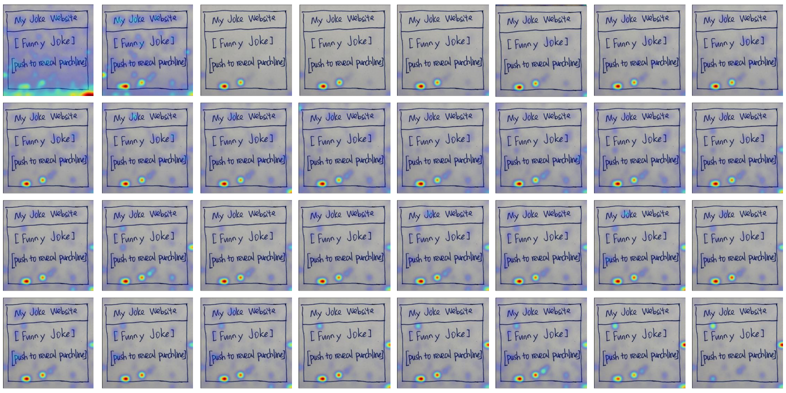


(c) All-One

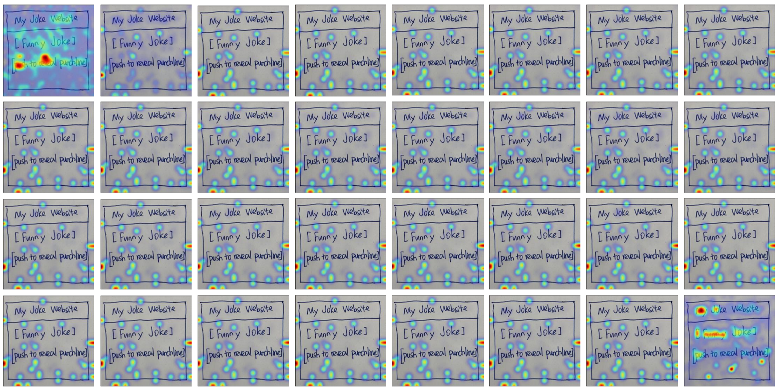


(d) PyPE

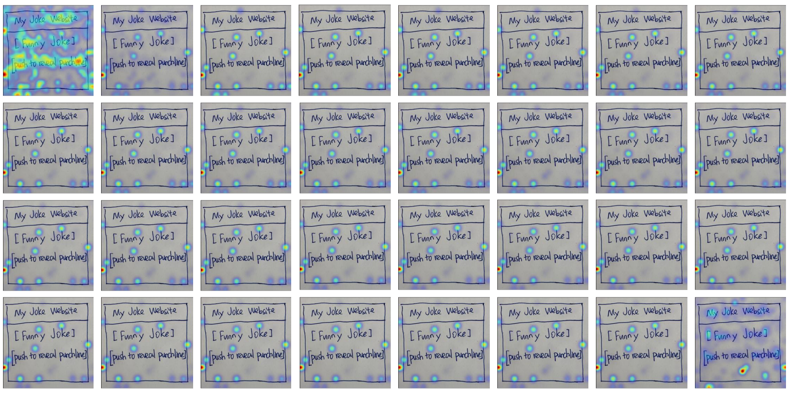
Figure 7: Layer-wise attention visualization (left to right, up to down) of the first example from Table 7.



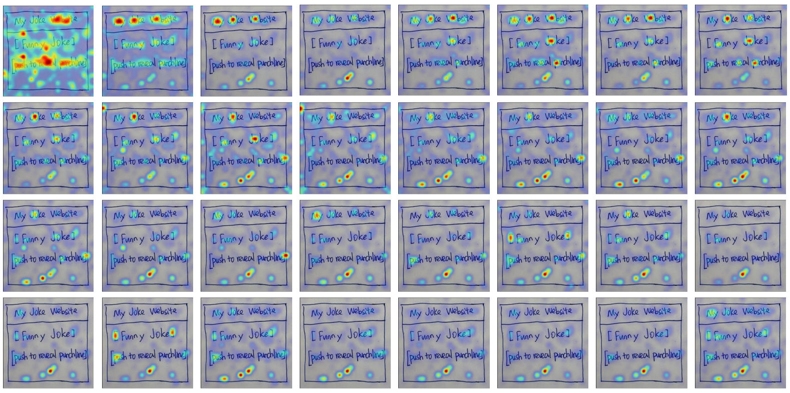
(a) Raster-scan



(b) Concentric



(c) All-One



(d) PyPE

Figure 8: Layer-wise attention visualization (left to right, up to down) of the second example from Table 7.