# V2PE: Improving Multimodal Long-Context Capability of Vision-Language Models with Variable Visual Position Encoding

Supplementary Material

## A. Experiment on Larger VLMs

In Sec. 4, we observe that our V2PE method exhibits limited improvements on general benchmarks when employing a smaller $\delta$. We hypothesize that this is due to the InternVL2-2B used in our experiments, which underwent pre-training and SFT with conventional position encoding, inherently adapting it to a position increment of $\delta = 1$ for short contexts.

To further validate our hypothesis, we begin with the pre-training of a hybrid InternVL2.5-7B model, whose language backbone is Qwen 2.5-7B[26], distinct from the InternLM-2B backbone used in the main paper. At this stage, the model has not yet adapted to conventional positional encoding for visual tokens (i.e., $\delta = 1$). We conduct pre-training experiments both with and without V2PE, and evaluate the resulting models on a suite of general benchmarks. As shown in Table 2, integrating V2PE yields better performance at smaller $\delta$ values, supporting our hypothesis and highlighting the generalizability of V2PE in larger-scale models.

Furthermore, to assess the architectural generality of V2PE, we apply it to LLaVA-One-Vision-7B[10] and conduct subsequent supervised fine-tuning (SFT). The results, presented in Table 1, consistently demonstrate the effectiveness of V2PE across different model backbones.

| **Model** | **MM-NIAH** | | |
|---|---|---|---|
| | **Image** | **Text** | **Avg** |
| LLaVA-One-Vision-FT32K | 61.9 | 79.5 | 70.7 |
| LLaVA-One-Vision-V2PE32K ($\delta$ = 1/256) | 67.6 | 84.4 | 76.0 |

Table 1. Performance comparison on MM-NIAH after applying V2PE to LLaVA-One-Vision-7B. V2PE significantly improves both image and text understanding.

## B. Optimal $\delta$ Selection

We have conducted a detailed analysis of the context length distribution of training samples, the selection strategy of optimal $\delta$, and corresponding empirical experiments.

### B.1. Training Context Length Distribution.

During training, we uniformly sampled $\delta$ from the set

$$\Delta = \left\{ \frac{1}{1}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{256} \right\} \tag{1}$$

Therefore, the expected context length is given by

$$\begin{aligned} \mathbb{E}[L_{\text{context}}] &= \mathbb{E}[N_{\text{img}} \cdot \delta + N_{\text{text}}] \\ &= \mathbb{E}[N_{\text{img}}] \cdot \frac{1}{4.51} + \mathbb{E}[N_{\text{text}}] \end{aligned} \tag{2}$$

Where $L_{\text{context}}$ is the total context length, $N_{\text{img}}$ is the number of image tokens, and $N_{\text{text}}$ the number of text tokens. We denote this expected value as the optimal context length, denoted $L^*$. Applying the above formula to the *training* datasets for Long-VQA and Long-MR yields $L^*$ values of 3.5k and 11.1k, respectively.

### B.2. Inference $\delta$ Selection strategy.

To align inference with the training distribution, we default to $\delta = \frac{1}{4}$, which is closest to $\frac{1}{4.51}$ in $\log_2$ space.

However, if the actual context length during inference $L_{\text{eval}} = N_{img}/4 + N_{text}$ exceeds the optimal context length $L^*$. In such cases, we aim to find a $\delta'$ such that $L^* = N_{img} \cdot \delta' + N_{text}$. The boundary between the two cases should be continuous. Thus, our delta selection strategy can be formalized as:

$$\delta' = \min\left( \frac{1}{4}, \frac{L^* - N_{\text{text}}}{N_{\text{img}}} \right) \tag{3}$$

We then select the closest available value $\delta$ to $\delta'$ from the discrete set $\Delta$, again using $\log_2$ proximity.

### B.3. Experiments.

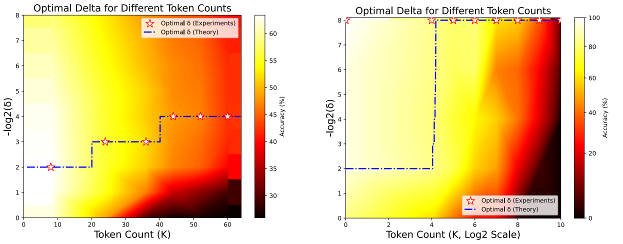
Fig. 1 compares the theoretically predicted and empirically optimal $\delta$ values for Long-VQA and Long-MR tasks. For Long-VQA, the match is strong across the full range. For Long-MR, due to a low image token ratio (39%), there is a sharp change after 20K tokens, with $\delta = \frac{1}{256}$ nearly always optimal, aligning with the prediction. Below 20K tokens, the performance is close across $\delta$. We consider this to be in agreement as well.

## C. Dataset Details

We have introduced two augmented long-context multimodal datasets: Long-VQA and Long-MR, designed to systematically evaluate and analyze the long-context capabilities of Vision-Language Models (VLMs). Representative examples from these datasets are illustrated in Fig.4, Fig.5, Fig.6, and Fig.7. Next, we will provide a detailed description of the dataset construction process.

| Model | $\delta$ | ChartQA | DocVQA | AI2D | InfoVQA | SQA | POPE | MMMU$_{\text{val}}$ | MMBench$_{\text{EN}}$ | SEED$_{\text{I}}$ | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| InternVL2.5-7B | – | 79.4 | 85.4 | 81.1 | 68.4 | 94.4 | 87.9 | 51.6 | 82.0 | 75.9 | 78.5 |
| InternVL2.5-7B + V2PE | 1/256 | 81.2 | 88.5 | 81.0 | 67.7 | 94.4 | 88.3 | 50.7 | 81.4 | 75.9 | 78.8 |
|  | 1/64 | 81.7 | 89.4 | 81.3 | 69.6 | 94.7 | 88.3 | 52.3 | 81.8 | 75.9 | 79.4 |
|  | 1/16 | 81.7 | 90.4 | 81.6 | 70.4 | **95.0** | 88.2 | **53.3** | **81.9** | 76.1 | 79.8 |
|  | 1/4 | **82.4** | **91.0** | **81.8** | 71.7 | 94.9 | 88.1 | 52.6 | **81.9** | 76.1 | **80.1** |
|  | 1/1 | 82.2 | 90.2 | 81.7 | **71.4** | 94.6 | **88.5** | 52.4 | 82.2 | **76.2** | 79.9 |

Table 2. Evaluation results of InternVL2.5-7B and its V2PE variants on multiple multimodal benchmarks. V2PE consistently improves performance across different $\delta$ values, with best average performance at $\delta = 1/4$.



(a) Long VQA Task

(b) Long MR Task

Figure 1. Comparison between the predicted optimal $\delta$ (calculated by Eq. 3) and empirically optimal $\delta$, along with the accuracy heatmap of the experiment results.

### C.1. Long Visual Question Answering (Long-VQA)

The Long-VQA dataset presents a novel challenge to VLMs, necessitating advanced visual perception and sophisticated reasoning capabilities to address tasks involving long context. This dataset is synthesized by combining multiple existing datasets in Tab. 3 to create a set of complex multi-image tasks.

For datasets that primarily consist of document-like images, such as DocVQA [15], we extend the context by merging multiple single-page documents into cohesive multi-page collections. Questions are subsequently sampled from one of the original documents, ensuring that the model's ability to retain and utilize information across an extended multi-page context is rigorously evaluated.

In the case of datasets composed of visual elements like images, charts, and tables, such as those from GQA [7], VizWiz [6], and TabFact [3], we aggregate these components into complex, multi-page documents that emulate naturalistic scenarios. Each visual element, whether an image or a chart, is strategically positioned across different pages and at various locations (e.g., upper-left, center, lower-right). This configuration is designed to evaluate a model's complex reasoning capabilities, as it requires an understanding of the relative positioning of elements throughout the entire document.

By constructing a diverse and challenging dataset, Long-VQA not only evaluates a model's ability to process a wide range of visual inputs but also emphasizes the necessity of navigating through complex, multi-image contexts. This synthesis of data from multiple sources, combined with the deliberate complexity of the spatial layouts, establishes a rigorous benchmark for VLMs. Additionally, we provide the length distribution of the Long-VQA test set in Fig. 2.

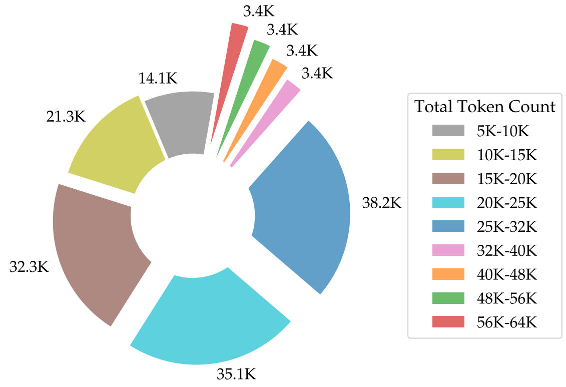


Figure 2. The token length distribution of test set in Long-VQA.

### C.2. Long Multimodal Retrieval (Long-MR)

Our proposed Long-MR dataset is constructed upon the MM-NIAH [23] benchmark, designed specifically to evaluate the performance of VLMs in long-context multimodal retrieval tasks. To further assess the generalization capabilities of VLMs within this task, we introduce additional synthetic variations that increase the task complexity.

Unlike the original MM-NIAH, where a single needle is inserted, our Long-MR dataset incorporates multiple needles into the long-context multimodal sequence. Of these needles, only one is considered as the target query, while the remainder serve as negative needles. This configuration introduces significantly more challenging negative instances, compelling the model to accurately distinguish between highly similar yet irrelevant needles in a lengthy contextual sequence.

To enrich the diversity of needles, we leverage advanced large language models (LLMs) to create synthetic needles beyond those included in the official MM-NIAH benchmark. This expansion results in a more heterogeneous dataset that emulates real-world complexity, thereby improving the robustness of the evaluation. Such diversification reduces the risk of the model overfitting to a particular

Table 3. Data statistics of Long-VQA dataset.

| Dataset | Dataset Size | |
|---|---|---|
| | Training | Validation |
| DeepForm [24] | 3.4K | 2.1K |
| DocVQA [15] | 39.4K | 6.0K |
| InfoVQA [16] | 23.9K | 4.1K |
| Kleister [21] | 13.4K | 5.5K |
| SQA [12] | 10.2K | 4.1K |
| VisualMRC [22] | 15.8K | 7.4K |
| ChartQA [14] | 40.1K | 3.3K |
| DVQA [9] | 150.0K | 16.4K |
| TabFact [3] | 91.6K | 13.4K |
| WikitabQS [18] | 14.1K | 5.1K |
| Clevr [8] | 150.0K | 16.4K |
| GQA [7] | 150.0K | 16.4K |
| OcrVQA [17] | 150.0K | 16.4K |
| OKVQA [13] | 9.0K | 5.8K |
| TextCaps [19] | 110.0K | 17.2K |
| TextVQA [20] | 56.5K | 6.5K |
| Vizwiz [6] | 20.5K | 8.8K |
| Total | 1.1M | 155.0K |

needle category, fostering the development of more generalizable retrieval capabilities.

Fig. 3 illustrates the length distribution of our costumed evaluation split, denoted as MM-NIAH$_{1M}$. Notably, the majority of sequences fall within the 512K to 1M token range. For contexts with lengths shorter than 64K, we directly utilize the samples from the original MM-NIAH benchmark.

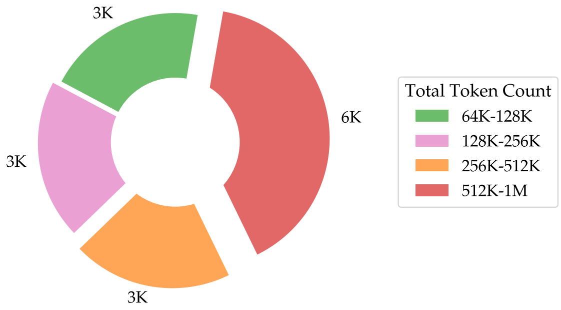


Figure 3. The token length distribution of our MM-NIAH$_{1M}$

## D. Evaluation

To address the out-of-memory challenge encountered during inference on samples exceeding token lengths of 128K in the MM-NIAH$_{1M}$ evaluation dataset, we adopt a perplexity-based approach similar to that employed in LongVA [27]. Specifically, during evaluation, we concatenate the question embedding, which integrates both textual and visual components, with the output answer embedding. Subsequently, a single forward pass is performed using ring attention to predict the logits of the answer. The output is considered correct if the index corresponding to the highest output logit across all tokens within the answer span aligns with the correct answer.

Table 4. Summary of our training hyper-parameters.

| Configuration | V2PE Setting |
|---|---|
| Weight init | InternVL2-2B [4] |
| Loss type | Generative loss |
| Learning rate schedule | Cosine decay |
| Optimizer | AdamW [11] |
| Learning rate | 5e-6 |
| Weight decay | 5e-2 |
| Input image resolution | 448 × 448 |
| Warmup steps | 150 |
| Iterations | 5K |

To facilitate comparison between position encoding extension and our proposed V2PE, we determine the interpolation factor for linear interpolation [2] based on the test sample length and the context window size used during training. Specifically, we interpolate the position indices of the test samples to match the context window range from the training phase. For example, when evaluating InternVL2-FT-32K on the 64K-length Long-VQA task using linear interpolation [2], we utilize an interpolation factor of 2, which effectively maps the position indices of the test samples into the 32K range, consistent with the context length employed during training. Similarly, for evaluations involving 1M-length samples, an interpolation factor of 32 is selected. For the NTK-Aware Scaled RoPE [1], we fix the scaling factor at 5, as our experimental results indicate that it yields consistent performance across tasks of varying lengths.

## E. Experiment Details

### E.1. Technical details about V2PE

The detailed training configurations are summarized in Table 4. Additionally, for experiments involving the V2PE method, we employ the `Float32` data type when computing positional indices and positional embeddings required for RoPE, to ensure computational precision.

### E.2. Linear interpoation and NTK-Awared Scaled RoPE

When comparing the performance of linear interpolation and V2PE, we select the interpolation coefficient based on the length of the test samples and the context window used during training. Specifically, we interpolate the position indices of the test samples into the range of the training context window. For example, in the case of InternVL2-FT-32K, when evaluating it on the 64K-length Long VQA task, we set the interpolation coefficient to 2 so that the position indices of the test samples are mapped within the 32K range, which corresponds to the context length used during training. Similarly, when testing on 1M-length samples, we set the interpolation coefficient to 32 to ensure that the position indices during testing do not exceed the training context window too much.

When comparing NTK-Aware Scaled RoPE and V2PE, we fix the coefficient of NTK-Aware Scaled RoPE at 5, as our experiments indicate that this value performs well across tasks of different lengths.

### E.3. Visual token compression in ablation study

In the ablation study, we follow InternVL's compression approach (PixelShuffle + MLP) to implement token compression. By training both the token compression and V2PE methods on the same dataset and conducting evaluations, we demonstrate that the effectiveness of V2PE lies in compressing the range of position encoding while preserving the complete token information. To verify the orthogonality of our proposed V2PE with other token compression methods, we also conduct an ablation study for MM-NIAH image retrieval tasks on attention sink [25], which focuses on reducing inference computational cost in long context. The results show that the official InternVL2-2B achieves 26.3 points, which can be enhanced by applying V2PE to achieve 62.8 points. Moreover, InternVL2-2B can also be further improved by applying both V2PE and attention sink to achieve 70.3 points. It should be noted that V2PE and attention sink both target long-context challenges, and V2PE specifically refines visual position encoding.

## F. Attention Matrices Analysis

To investigate the impact of our V2PE on attention mechanism, we follow  [5] to analyze the attention matrices on the Long-VQA evaluation set. Specifically, our analysis focuses on the tail portion of the entire attention matrices, which corresponds to the question segments located at the end of the sequences. This allows us to observe how effectively the model retrieves relevant information when answering questions.

As illustrated in Fig. 8, we observe that as the positional increment parameter $\delta$ decreases, the attention patterns in Layer 1 exhibit an increasingly distinct emphasis on visual tokens. This observation suggests that with smaller values of $\delta$, the model becomes more attentive to visual content, which is crucial for answering questions involving visual inputs. Furthermore, Fig. 9 shows that in deeper layers (e.g., Layer 15), the attention becomes more focused around a specific sequence index, particularly ID=1410, as $\delta$ decreases. Notably, the answer to the corresponding question is located near the 1410-*th* token. This indicates that a smaller $\delta$ not only sharpens the model's focus but also aligns its attention more effectively with the tokens containing the correct answer.

These findings imply that using smaller positional increment $\delta$ allows the model to better align its attention with the critical portions of the input sequence, thereby enhancing its capability to retrieve relevant information, especially in the scenarios of long-context multimodal tasks.

## References

[1] @bloc97. Ntk-aware scaled rope, 2023. 3

[2] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. 3

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 2, 3

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3

[5] Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. The neural data router: Adaptive control flow in transformers improves systematic generalization. *arXiv preprint arXiv:2110.07732*, 2021. 4

[6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2, 3

[7] Drew A Hudson and Christopher D Manning. 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, 2019. 2, 3

[8] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[9] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 3

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1

[11] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[12] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[13] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[14] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 3

[15] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2, 3

[16] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3

[17] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3

[18] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 3

[19] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 3

[20] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[21] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pages 564–579. Springer, 2021. 3

[22] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13878–13888, 2021. 3

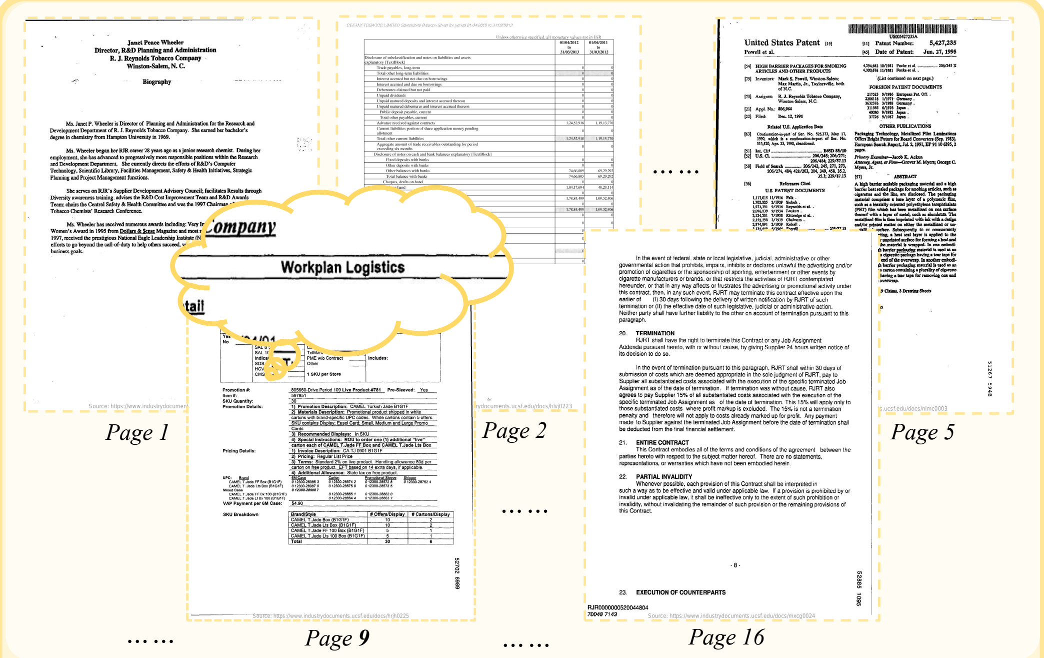
[23] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 2

[24] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. A benchmark for structured extractions from complex documents. *ArXiv, abs/2211.15421*, 2, 2022. 3

[25] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 4

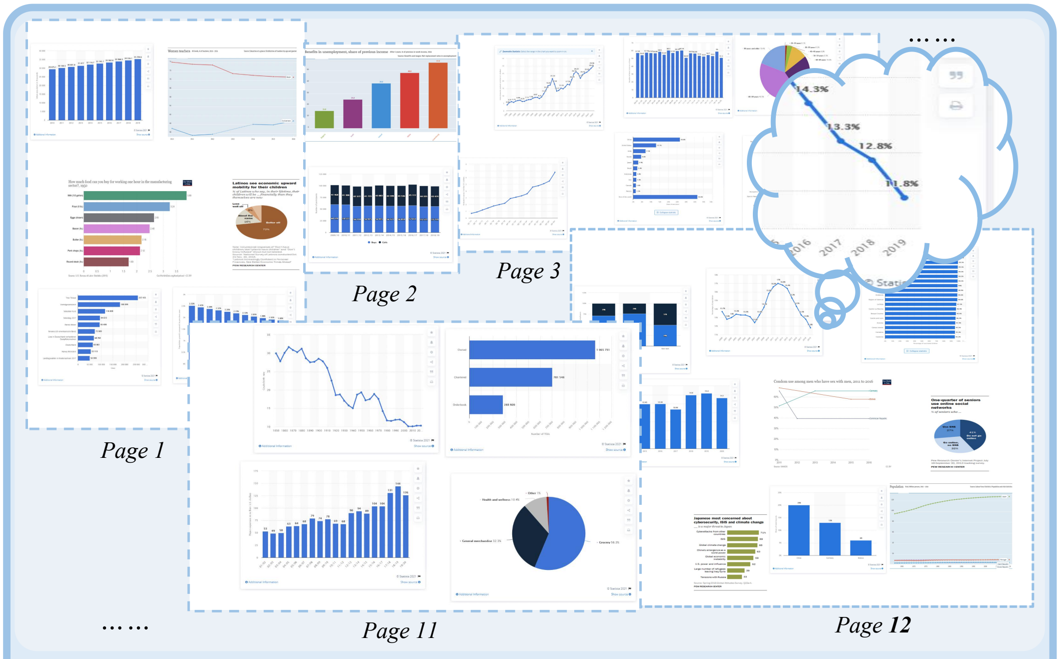
[26] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. 1

[27] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3



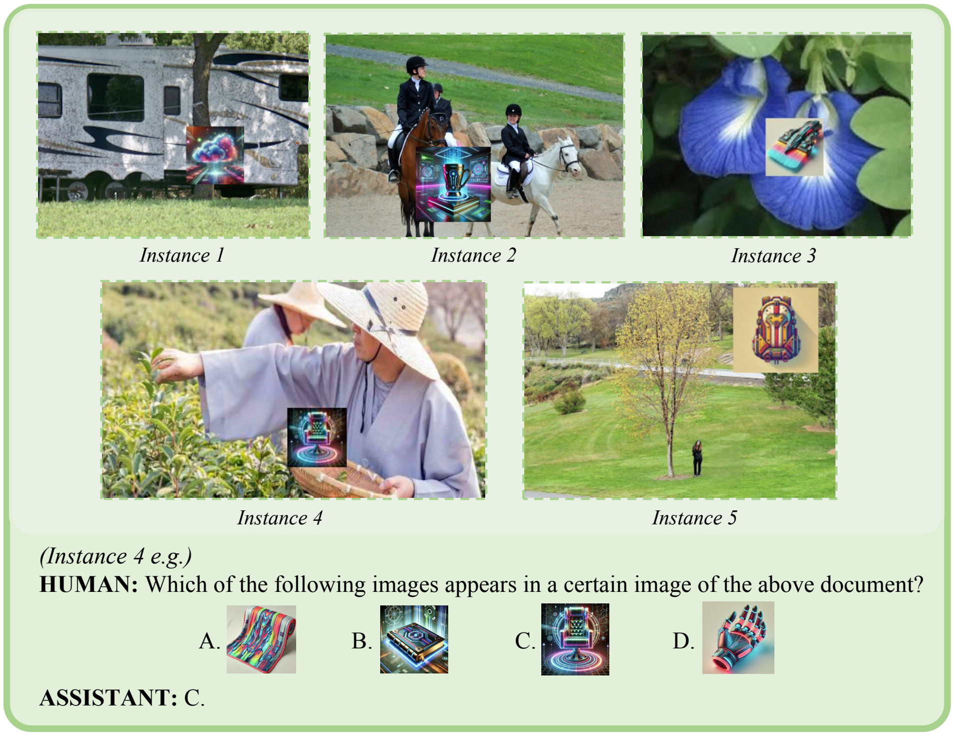
**HUMAN:** Referring to the No.**9** page, What kind of logistics does the title read? Answer the question using a single word or phrase.
**ASSISTANT:** Workplan.

Figure 4. Examples of DocVQA subset from Long-VQA dataset.



**HUMAN:** According to the **top-center** chart on page **twelve**, What is the poverty rate in California in the year 2019? Answer the question using a single word or phrase.
**ASSISTANT:** 11.8.

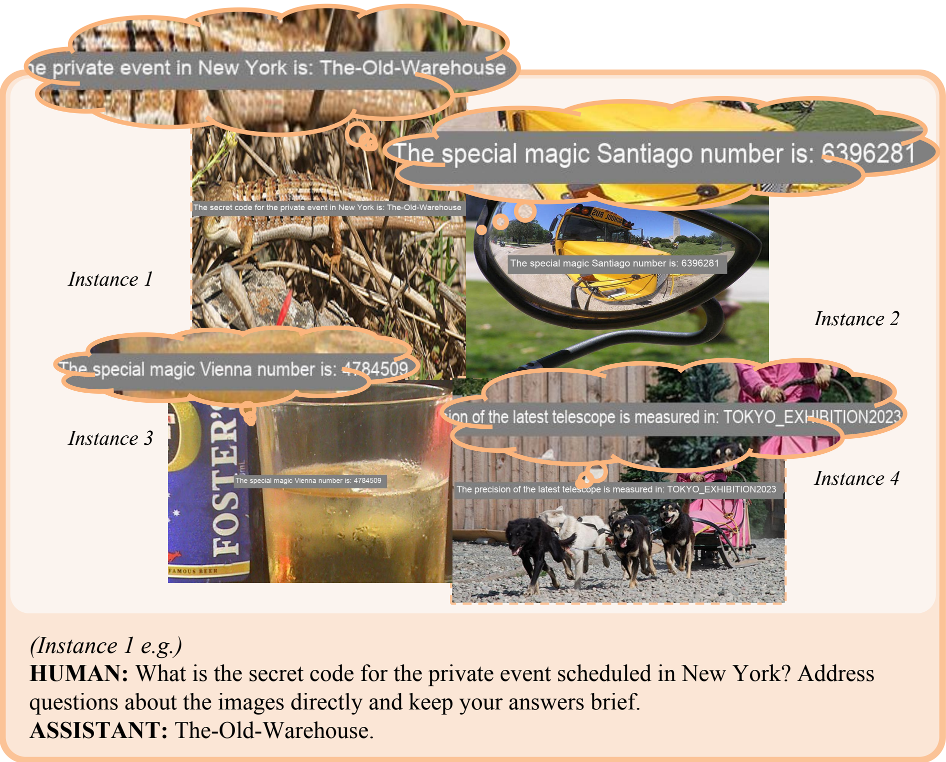
Figure 5. Examples of ChartVQA subset from Long-VQA dataset.



*(Instance 4 e.g.)*
**HUMAN:** Which of the following images appears in a certain image of the above document?

**ASSISTANT:** C.

Figure 6. Examples of *Retrieval-Image-Needle* in our proposed Long-MR dataset.


he private event in New York is: The-Old-Warehouse
The secret code for the private event in New York is: The-Old-Warehouse
Instance 1
The special magic Santiago number is: 6396281
The special magic Santiago number is: 6396281
Instance 2
The special magic Vienna number is: 4784509
The special magic Vienna number is: 4784509
Instance 3
ion of the latest telescope is measured in: TOKYO_EXHIBITION2023
The precision of the latest telescope is measured in: TOKYO_EXHIBITION2023
Instance 4
FOSTER'S

*(Instance 1 e.g.)*
**HUMAN:** What is the secret code for the private event scheduled in New York? Address questions about the images directly and keep your answers brief.
**ASSISTANT:** The-Old-Warehouse.

Figure 7. Examples of *Image-Needle-In-A-Haystack* with complex needles in our proposed Long-MR dataset. These needles vary in answer format, font-size and style.

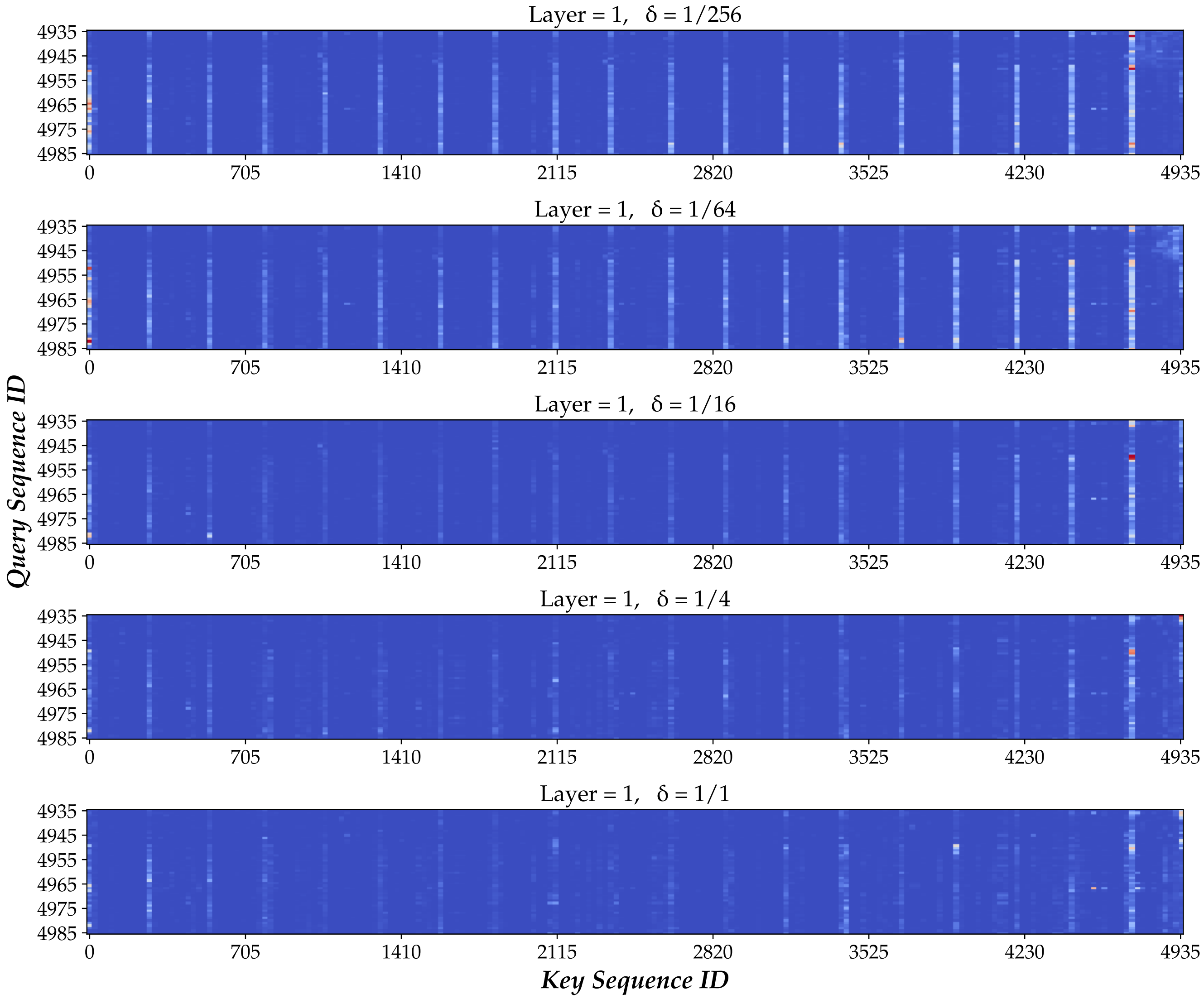


Figure 8. Attention map visualization in layer 1 (Maximum over 16 heads).

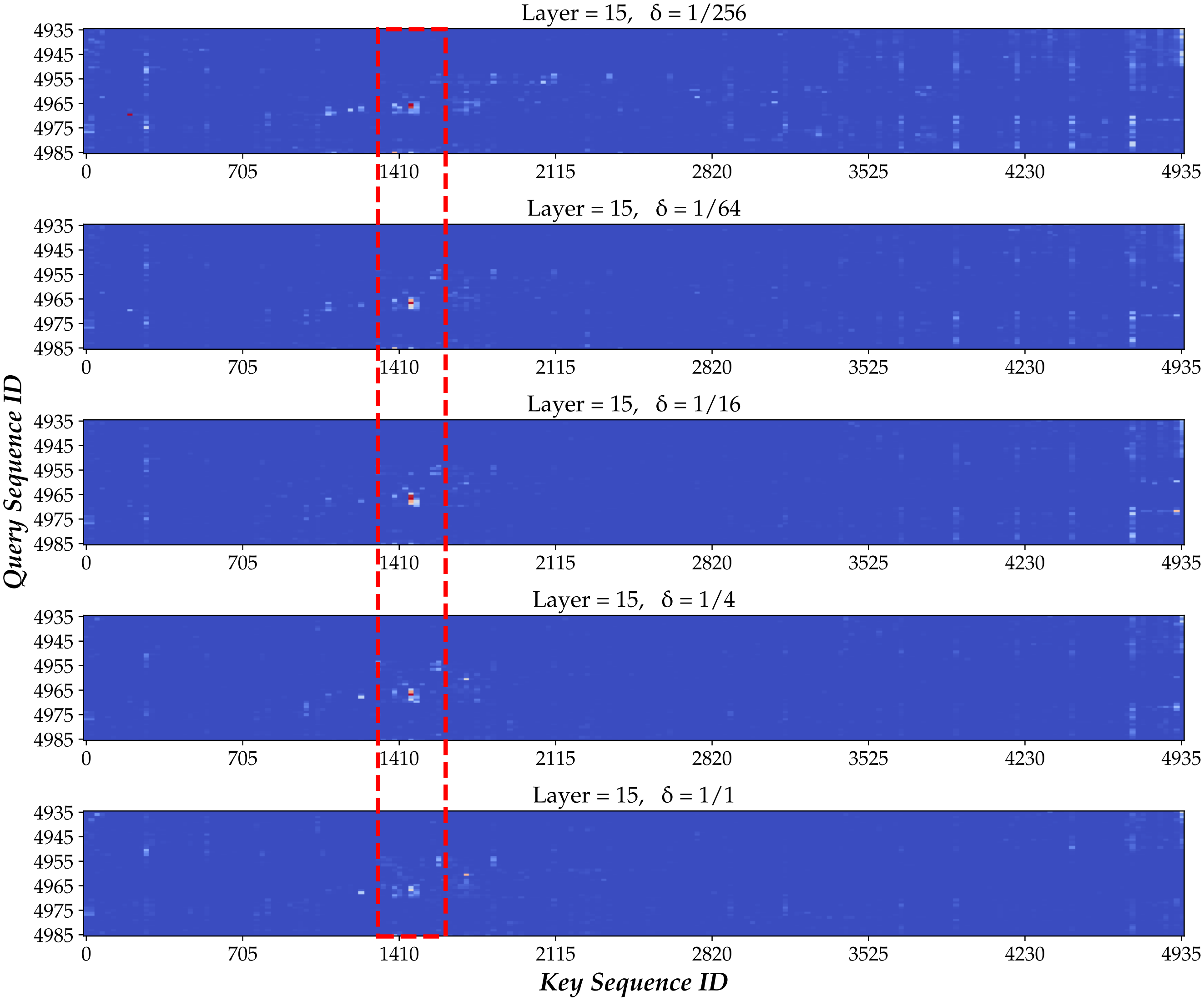


Figure 9. Attention map visualization in layer 15 (Maximum over 16 heads).