PPE: POSITIONAL PRESERVATION EMBEDDING FOR TOKEN COMPRESSION IN MULTIMODAL LARGE LANGUAGE MODELS

Mouxiao Huang*, **Borui Jiang**✉*, **Dehua Zheng**, **Hailin Hu**✉, **Kai Han**, **Xinghao Chen**✉
Huawei Technologies
✉{jiangborui, hailin.hu, xinghao.chen}@huawei.com

ABSTRACT

Multimodal large language models (MLLMs) have achieved strong performance on vision-language tasks, yet often suffer from inefficiencies due to redundant visual tokens. Existing token merging methods reduce sequence length but frequently disrupt spatial layouts and temporal continuity by disregarding positional relationships. In this work, we propose a novel encoding operator dubbed as **P**ositional **P**reservation **E**mbedding (**PPE**), which has the main hallmark of preservation of spatiotemporal structure during visual token compression. PPE explicitly introduces the disentangled encoding of 3D positions in the token dimension, enabling each compressed token to encapsulate different positions from multiple original tokens. Furthermore, we show that PPE can effectively support cascade clustering — a progressive token compression strategy that leads to better performance retention. PPE is a parameter-free and generic operator that can be seamlessly integrated into existing token merging methods without any adjustments. Applied to state-of-the-art token merging framework, PPE achieves consistent improvements of $2\% \sim 5\%$ across multiple vision-language benchmarks, including MMBench (general vision understanding), TextVQA (layout understanding) and VideoMME (temporal understanding). These results demonstrate that preserving positional cues is critical for efficient and effective MLLM reasoning. Our code is available at https://github.com/MouxiaoHuang/PPE.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently achieved remarkable success across a range of vision-language understanding tasks Bai et al. (2025); Lei et al. (2025); Li et al. (2024a); Yuan et al. (2025); Zhang et al. (2025). A common paradigm involves encoding images or video frames into dense visual tokens, which are then fed into the language model for joint understanding. However, this dense representation is often highly redundant, leading to inefficiencies in computation and inference Song et al. (2024). To address this, recent works Dhouib et al. (2025); Jin et al. (2024); Ma et al. (2023); Zeng et al. (2022); Zhang et al. (2024a) have explored visual token compression, which merges similar tokens to reduce visual sequence length while preserving semantic information, thereby accelerating inference and lowering memory usage.

Despite the efficiency, existing compression methods often disrupt the spatial and temporal structure of visual inputs, limiting their applicability in layout-sensitive tasks such as counting, temporal grounding and sequential understanding. As shown in Figure 1 (a), clustering-based Chat-UniVi Jin et al. (2024) token compression may discard fine-grained spatial or temporal cues. Figure 1 (b) illustrates the recent methods like PACT Dhouib et al. (2025) which have attempted to preserve layouts during compression, but still remain constrained by the insufficient and imprecise positions.

In this work, we introduce *Positional Preservation Embedding* (PPE), a novel positional encoding strategy which explicitly retains the different spatiotemporal layout into one compressed token. Our design is motivated by two key principles. First, we aim to preserve spatial and temporal positions during similar token merging. To this end, PPE firstly assigns each visual token a positional ID (*e.g.*,

*Equal contribution.

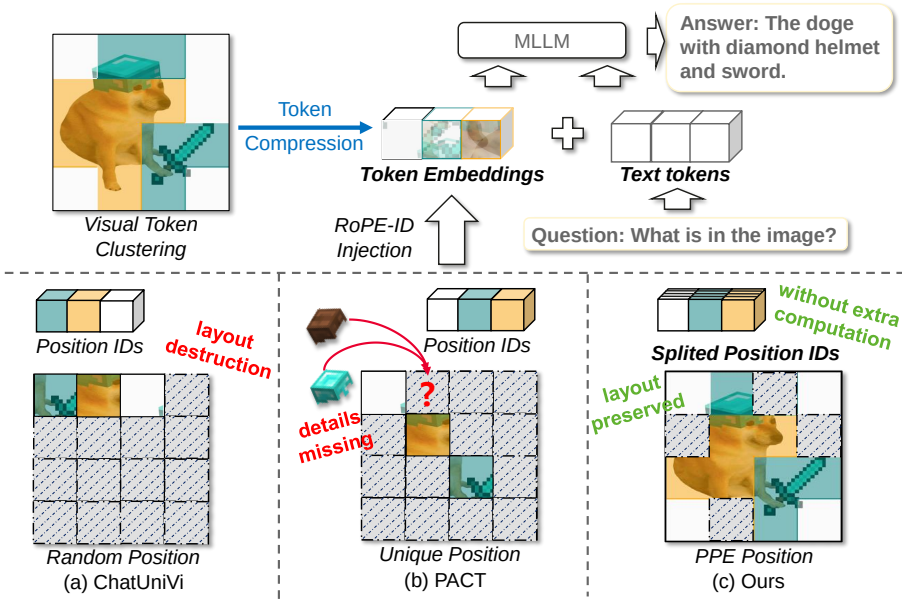


Figure 1: **Comparison between PPE and other token merging methods of processing positional IDs**. To simplify, the components such as the visual encoder are omitted. **(a)** ChatUniVi Jin et al. (2024) mainly assigns randomize ID value to the clustered visual tokens. **(b)** PACT Dhouib et al. (2025) retains the ID of the cluster center for the clustered visual tokens. **(c)** Proposed PPE splits the IDs of compressed token on different dimensions, so that each compressed token could contain **several** original position IDs.

2D spatial for images and 3D spatiotemporal for videos). During compression, each merged token retains several positional IDs of their constituents. This ensures that most of the visual scene layouts are still accessible to MLLM at high compression rates, as shown in Figure 1 (c).

Secondly, we support the widely adopted cascaded compression strategy, which performs token merging progressively across the Transformer Vaswani et al. (2017) layers. This design is inspired by the observation that different layers capture increasingly abstract representations with higher similarity Bolya et al. (2023a); Dhouib et al. (2025); Song et al. (2024); Zhang et al. (2024a). Thus, merging tokens in a multi-stage manner enables higher compression ratios without collapsing the shallow semantics prematurely. Since PPE is decoupled from the merging algorithm and operates solely on position IDs, it naturally extends to multi-stage compression. This allows our method to preserve fine-grained spatiotemporal layouts across multiple compression stages, resulting in higher compression ratio, IDs retention (*Section 4.4.5*) and improved performance (*Section 4.3*).

To our best knowledge, PPE is the first work that explores an effective and lightweight positional preservation solution during visual token compression of MLLM. In contrast to existing positional embedding methods that only preserves one position for single token and lead to the loss of detailed layouts, PPE enables each single token to represent multiple positions, thus preserving the vision layout more completely. Moreover, PPE is parameter-free and can be used in a plug-and-play fashion for easily implanting into existing visual token compression methods without any additional computational costs.

In the experiments, we apply PPE for tackling a variety of vision-language tasks, including general vision-language understanding on MMBench Liu et al. (2024a), text-based visual reasoning on TextVQA Singh et al. (2019), temporal understanding on VideoMME Fu et al. (2024), etc. The reported performances consistently outstrip previous compression methods Dhouib et al. (2025); Jin et al. (2024) by significant margins after fine-tuning. We strongly believe that PPE can make inroads into domains of sparse token representing in MLLM where dense visual representation had previously reigned supreme.

The main contributions of this work are as follows:

- We identify a critical limitation in existing visual token merging methods—namely, the neglect of spatial structure preservation and temporal coherence—which leads to distortion of intra-frame layouts and disruption of inter-frame temporal relations.
- We propose Positional Preservation Embedding (PPE), a novel, plug-and-play approach that explicitly preserves spatiotemporal integrity during token merging, effectively addressing both spatial and temporal challenges.
- We further show that PPE can be applied in cascade compression manner within multiple transformer layers of the MLLM, enabling substantial compression with minimal performance degradation.
- We conduct extensive experiments across serveal image and video benchmarks, demonstrating that PPE maintains accuracy while reducing the visual token count by 90% and outperforms other visual token compression methods at comparable reduction ratios.

2 Related Work

2.1 Multimodal Large Language Models

MLLMs extend traditional LLMs by incorporating visual inputs, enabling unified vision-language understanding and generation. Recent models such as Flamingo Alayrac et al. (2022), the LLaVA series Li et al. (2024a); Liu et al. (2023); Zhang et al. (2024b), and the Qwen-VL series Bai et al. (2023; 2025); Wang et al. (2024) achieve strong performance across captioning, VQA, and instruction following. These models typically align vision and language via cross-attention, projection modules, or lightweight multimodal transformers. However, most rely on dense visual tokens, posing efficiency and scalability challenges for high-resolution or long-form visual inputs. Extensions such as Video-LLaVA Lin et al. (2024), VideoChat Li et al. (2023b), and Video-LLaMA Zhang et al. (2023) mitigate this via frame sampling or sparse memory. In contrast, our method introduces a Positional Preservation Embedding that enables substantial token reduction while maintaining spatiotemporal coherence, offering a scalable alternative for multimodal reasoning.

2.2 Visual Token Merging

To improve the efficiency of MLLMs, recent works have explored visual token reduction strategies Kong et al. (2025) such as clustering and merging Bolya et al. (2023a); Xu et al. (2022); Ma et al. (2023); Zeng et al. (2022); Dhouib et al. (2025); Song et al. (2024); Jin et al. (2024); Yang et al. (2025). These methods significantly reduce the sequence length of image or video inputs, enabling faster inference and reduced memory consumption. Techniques include token matching Bolya et al. (2023a), hierarchical clustering Ma et al. (2023); Zeng et al. (2022), group-based representation Xu et al. (2022), and pruning-clustering hybrids Dhouib et al. (2025), with extensions to long-video understanding via sparse memory representations Song et al. (2024) and unified token compression across modalities Jin et al. (2024). Despite their effectiveness, these methods typically discard original positional information, which can disrupt spatial layouts and temporal continuity—limiting performance on fine-grained visual reasoning tasks. In contrast, our approach explicitly preserves spatiotemporal position cues throughout compression, maintaining structural fidelity while achieving substantial token reduction.

2.3 Positional Encoding in MLLMs

Positional encoding Vaswani et al. (2017) is essential in MLLMs to maintain spatial and temporal relationships across vision and language tasks. The Rotary Position Embedding (RoPE) Su et al. (2024) is a widely adopted method, which captures relative token positions in sequences using a rotational mechanism. This has been extended to 2D RoPE Heo et al. (2024), which is well-suited for image tasks, enabling models to understand spatial locality in vision transformers (ViTs). For video and temporal tasks, the Qwen series introduced 3D MRoPE Bai et al. (2023); Wang et al. (2024), which integrates rotary encoding across both spatial and temporal dimensions, preserving continuity across frames. However, these methods often face challenges in visual token compression, as reducing token counts can lead to the loss of crucial positional information. In contrast, we propose **PPE**: **P**ositional **P**reservation **E**mbedding, a novel approach that maintains spatiotemporal positional cues

during token reduction, preserving structural integrity while enabling aggressive compression. This method, to the best of our knowledge, is a new exploration in the field.

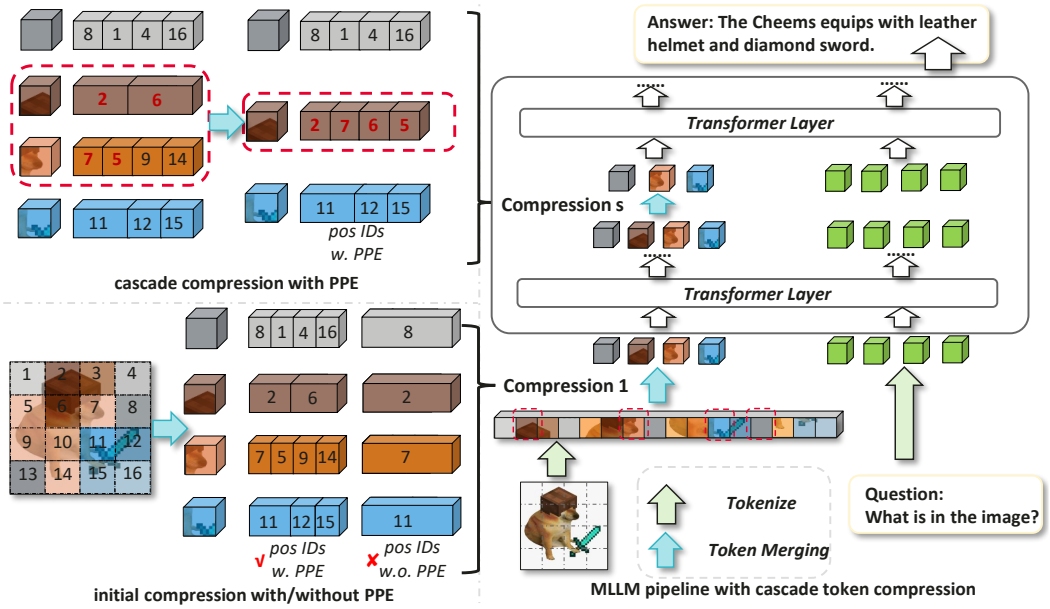


Figure 2: **The overview pipeline of the proposed PPE with cascade compression. Left:** Main idea of Positional Preservation Embedding (PPE) integrated in token compression. For each RoPE ID in compressed token embedding, PPE splits the dimension into chunks to prefill multiple position IDs. The IDs of tokens with high importance scores are reserved preferentially. **Right:** The MLLM pipeline integrating PPE and cascade compression. Token compression is applied in multiple layers, each with PPE. See main text for more explanation.

3 METHODOLOGY

In this section, we first recap the preliminaries of PPE, and then analyze why previous methods fail to hold the positional information. Afterward, we introduce our proposed PPE and cascade compression, as illustrated in Figure 2.

3.1 PRELIMINARY

Rotary Position Embeddings in MLLM. RoPE Su et al. (2024) is a positional encoding technique designed to enhance the Transformer Vaswani et al. (2017) architecture by integrating relative positional information directly into the self-attention mechanism. Unlike traditional absolute positional encodings, RoPE applies a rotation to the query and key vectors in multi-head attention, enabling the model to capture relative positions effectively. Assumed that token vector $\mathbf{z} \in \mathbb{N}^D$ in multi-head attention, the rotation operation can be represented as:

$$\text{RoPE}(z_d, m) = e^{im\theta_d} z_d, \; d = 1...D, \tag{1}$$

where m indicates the position ID of z. For capturing spatiotemporal layouts, Qwen2.5-VL Bai et al. (2025) introduces M-RoPE, a structured embedding scheme that partitions the embedding dimension, each encoding positional information along different visual axis—such as temporal order, image height, and width. Formally, the original RoPE rotation operation is modified to:

$$\text{M-RoPE}(z_d, m_d) = e^{im_d\theta_d} z_d, \; d = 1...D, \tag{2}$$

where $\mathbf{m} \in \mathbb{Z}^D$ is pre-filled 2D or 3D visual position IDs. Assumed that the 3D visual token is at position (t, h, w), the $\{m_d\}$ could be indicated as:

$$m_d = \begin{cases} t, & d = 1 \; ... \; D_1, \\ h, & d = D_1 + 1 \; ... \; D_1 + D_2, \\ w, & d = D_1 + D_2 + 1 \; ... \; D_1 + D_2 + D_3, \end{cases} \tag{3}$$

where D_1, D_2, D_3 are human-crafted integer to control the size of mrope sections Bai et al. (2025) satisfying that $D_1 + D_2 + D_3 = D$.

Visual Token Compression. ChatUniVi Jin et al. (2024) adopts a lightweight, parameter-free clustering algorithm for token compression to mitigate the computational overhead of long visual token sequences. The key idea is to merge tokens with DPC-KNN Du et al. (2016) based on token similarity. Assumed that N visual token embeddings $\{\mathbf{z}_i \in \mathbb{R}^D\}_{i=1}^N$ are clustered to M groups, $\{\mathbf{z}_i\}_{i \in C_j}$ indicates the token embeddings in group j. The compressed token embeddings $\mathbf{z}'_j$ is defined as:

$$\mathbf{z}'_j = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{z}_i,\ j = 1...M. \tag{4}$$

3.2 Positional Preservation Embedding

While similarity-based compression effectively reduces sequence length, it disrupts fine-grained spatiotemporal layouts. To mitigate this, we propose *Positional Preservation Embedding* (PPE), which retains multiple positional cues per merged token to preserve structural information.

PPE builds on the principle behind RoPE Su et al. (2024), in which the rotation of position embeddings is independent at the dimension. As a result, the m in Equation 1 can be totally different on the dimension D to represent different positions at the same time. For instance, M-RoPE Wang et al. (2024) partitions embedding dimensions into several groups to store spatiotemporal positions, formally in Equation 2. Inspired by this, PPE merges different positions to one merged token by splitting into more groups. Formally, the merged PPE ID $\hat{\mathbf{m}}$ could be indicated as:

$$\hat{m}_d = m_{k,d},\ d = (k-1)\frac{D}{K} + 1\ ...\ k\frac{D}{K}, \tag{5}$$

where K is a fixed hyper-parameter to represent the maximum capacity of PPE, and $\{\mathbf{m}_k\}^K$ is the set of different position IDs in 1D RoPE manner before merging. Note that K is always divisible by dimension D. In M-RoPE manner, K is set to the greatest common divisor of the mrope sections, guaranteed that each dimension is evenly cut into K groups:

$$\hat{m}_d^{3D} = m_{k,d}^{3D},\ d = \begin{cases} (k-1)\frac{D_1}{K} + 1 & ...\ k\frac{D_1}{K}, \\ (k-1)\frac{D_2}{K} + D_1 + 1 & ...\ k\frac{D_2}{K} + D_1, \\ (k-1)\frac{D_3}{K} + D_1 + D_2 + 1 & ...\ k\frac{D_3}{K} + D_1 + D_2. \end{cases} \tag{6}$$

The key insight shared of PPE is that similar token embeddings can share their feature embeddings during token merging. Rather than assigning a single position ID to a merged token, we extend this idea to a multi-position formulation that better reflects the internal diversity of a token cluster - containing more positional information of original input visual tokens. Specifically, consider a cluster group C_j containing token embeddings $\{\mathbf{z}_i\}_{i \in C_j}$ with corresponding position IDs $\{\mathbf{m}_i$. Rather than choosing one representative ID, we select the top-K IDs per cluster which are scored by the distance from the cluster center. Note that the score is higher if the token is closer to the cluster center Jin et al. (2024). If $|C_j| < K$, high-weight tokens are repeated to fill the slots. Denote the merged IDs of PPE as $\hat{\mathbf{m}}$, the PPE rotation of vector $\mathbf{z}$ is simply follow the RoPE which could be formulated as:

$$\mathrm{PPE}(z_d, \hat{m}_d) \ = \ e^{i\,\hat{m}_d \theta_d}\, z_d,\ d = 1...D, \tag{7}$$

3.3 Cascade Compression with PPE

In this section, we further investigate the effectiveness of cascade compression in conjunction with our proposed Positional Preservation Embedding (PPE) strategy. Cascade compression is a widely adopted technique in token compression Bolya et al. (2023a); Dhouib et al. (2025); Song et al. (2024); Zhang et al. (2024a), motivated by the observation that deeper Transformer layers exhibit greater representational redundancy Dong et al. (2021), whereas earlier layers encode critical low-level semantics that are less amenable to aggressive compression.

To this end, we implement a cascaded PPE-based compression pipeline built upon ChatUniVi Jin et al. (2024), as illustrated in Figure 2. In this design, visual token clustering is applied not only prior to feeding tokens into the LLM, but also within selected LLM layers. Leveraging PPE, we preserve

original positional information, enabling efficient computation of new cluster center positions after each merge.

For PPE ID assignment, we follow the standard top-K selection strategy described in *Section 3.2*, choosing token IDs with minimal distance to the cluster center. As shown in Figure 2, during repeated merges, the number of preserved position IDs reduces gradually (e.g., retaining only four IDs when two previously merged tokens are merged again).

Experimental results demonstrate that this cascaded PPE design maintains fine-grained spatial structure across compression stages and achieves higher compression ratio without performance loss. Moreover, this design significantly improves both ID retention (*Section 4.4.5*) and downstream task performance (*Section 4.3*). These findings confirm the strong compatibility of PPE with standard cascaded compression frameworks.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Datasets. To demonstrate the effectiveness of PPE, we construct our training datasets for supervised fine-tuning (SFT) referring to the public projects: LLaVA-Video-178k Zhang et al. (2024b) and LLaVA-OneVision Li et al. (2024a). Limited by the computational costs, we adopted even down-sampling on these datasets. To highlight the improvements of the model on image and video benchmarks, we further set up two different settings: one for images and another for videos. Specifically, we utilized about 120K video samples to fine-tune the models for multimodal video benchmarks, while 300K image samples to the models for the multimodal image benchmarks.

Evaluation Benchmarks. We use VideoMME Fu et al. (2024) as our primary benchmark for multimodal video analysis. To assess generalization, we also report results on NeXT-QA (multi-choice and open-ended) Xiao et al. (2021), SEED-Bench-Video Li et al. (2023a), and MVBench Li et al. (2024b). For image understanding, we evaluate on MMBench (CN/EN) Liu et al. (2024a) and SQA Iyyer et al. (2017) to assess general visual comprehension. To further examine text-rich layout and document understanding, we include TextVQA Singh et al. (2019), DocVQA Mathew et al. (2021), OCRBench Liu et al. (2024b), and ChartQA Masry et al. (2022). All reported metrics follow the evaluation criteria of their respective benchmarks, and unless otherwise noted, higher values indicate better performance.

Implementation Details. We conduct our default experimental settings by using the Qwen2.5-VL-3B-Instruct model Bai et al. (2025). All models are fine-tuned in a fully supervised manner on the down-sampled datasets with all parameters unfrozen. The training setting is mainly referring to the public project Lee (2024), in which the gradient accumulation step is 4, the learning rate is $1\mathrm{e}{-5}$ along with the warm-up ratio of 0.03, while the learning rate of visual encoder and patch merger is $2\mathrm{e}{-6}$ and $1\mathrm{e}{-5}$, respectively. We train for only 1 epoch. To maintain the optimal performance and computational costs for both training and evaluation, we mainly follow the official input size configurations. Specifically, we set `image_min_pixels` = $512 \times 28 \times 28$ and `image_max_pixels` = $1280 \times 28 \times 28$ for image inputs, and `video_min_pixels` = $128 \times 28 \times 28$ and `video_max_pixels` = $768 \times 28 \times 28$ for video inputs. Additionally, the maximum number of video frames is set to 64 due to the limited memory usage. The 3D M-RoPE section is set to $[16, 24, 24]$, while $[32, 32]$ for the 2D manner. The number of preserved token IDs is set to $K = 8$ for 3D and $K = 32$ for 2D, which corresponds to the greatest common divisor (GCD) of M-RoPE sections.

Token compression settings. The proposed PPE can be integrated into most existing compression methods. For our experiments, we adopt the SOTA clustering-based framework Chat-UniVi Jin et al. (2024) and follow its default training strategy, with slight modifications to adapt to the Qwen2.5-VL model. In Chat-UniVi, the number of clustered tokens is controlled by a predefined clustering ratio due to the native resolution technique. Specifically, the spatial clustering ratio is set to 0.45, while the temporal clustering ratio is 0.0625, consistent with the original paper. For fair comparison, token clustering is applied at the interface between the vision encoder and the LLM by default.

Table 1: The overall performance comparison across different benchmarks. The **Dense** model is a simply Qwen2.5-VL-3B-Instruct model fine-tuned on our SFT datasets. The subscript S/ST indicates spatial-only and spatiotemporal compression, respectively. **Bold** denotes the best performance among compressed variants, and underline denotes the best overall including the Dense baseline. The image average score is calculated without OCRBench.

Benchmarks	Dense	+Chat-UniVi$_S$	+PPE$_S$ (Ours)	+Chat-UniVi$_{ST}$	+PPE$_{ST}$ (Ours)
MMBench (EN)	85.89	**84.92**	84.73 (-0.19)	-	-
MMBench (CN)	86.07	83.71	**84.87** (+1.16)	-	-
SQA	76.90	77.30	**77.88** (+0.58)	-	-
TextVQA	79.50	57.66	**77.14** (+19.48)	-	-
DocVQA	89.44	52.48	**76.79** (+24.31)	-	-
OCRBench	761	535	**691** (+156)	-	-
ChartQA	79.96	49.60	**74.52** (+24.92)	-	-
Image Average	*82.96*	*67.61*	***79.32*** (+11.71)	-	-
VideoMME (w/o subs)	57.81	57.22	**58.70** (+1.48)	56.07	**57.41** (+1.34)
VideoMME (w subs)	57.96	57.22	**59.07** (+1.85)	56.15	**57.78** (+1.63)
NeXT-QA (MC)	78.20	77.63	**78.42** (+0.42)	77.59	**77.99** (+0.4)
NeXT-QA (OE)	31.65	25.37	**32.61** (+7.24)	26.55	**31.95** (+5.4)
SEED-Bench-Video	57.60	**56.08**	55.98 (-0.10)	53.47	**54.19** (+0.72)
MVBench	67.90	66.90	**67.38** (+0.48)	64.38	**66.42** (+2.04)
Video Average	*58.52*	*56.74*	***58.69*** (+1.95)	*55.70*	***57.62*** (+1.92)
Reduction Ratio	0%	55%	55%	94%	94%

4.2 Main Results

We evaluate the performance of our proposed method on a range of image and video benchmarks. The results show that PPE consistently outperforms previous approaches, demonstrating its effectiveness in both image and video understanding.

Image Tasks. Table 1 illustrates the strong performance of PPE on image understanding benchmarks. Despite reducing visual tokens by 55%, PPE achieves competitive or superior performance across all tasks. It significantly outperforms Chat-UniVi in overall accuracy (**79.32%** vs. **67.61%**) and on TextVQA (**77.14%** vs. **57.66%**), while maintaining comparable results on other tasks. Compared to the full-token Dense baseline model, PPE shows only a minor acceptable performance drop in overall accuracy (**82.96%** vs. **79.32%**), despite using less than half the visual tokens.

Video Tasks. As shown in Table 1, under a 55% token reduction with spatial-only compression, PPE consistently outperforms Chat-UniVi across most tasks and in overall accuracy (**58.69%** vs. **56.74%**), even exceeding the Dense baseline. With a more aggressive 94% reduction using spatiotemporal compression, PPE again leads in overall score (**57.32%** vs. **55.7%**), demonstrating strong robustness under heavy compression. Notably, on challenging benchmarks like VideoMME and NeXT-QA (OE), PPE shows clear improvements over Chat-UniVi.

Summary. In summary, PPE achieves substantial token reduction—55% for spatial-only compression and 94% for spatiotemporal compression—while preserving, or in some cases even enhancing, downstream task performance. This demonstrates that preserving visual token IDs enables more accurate reconstruction of the spatial layout and temporal order, even under aggressive compression, thereby improving reasoning capabilities, as illustrated in Fig. 1.

Comparison with M-RoPE. As reported in Table 1, Dense model adopts M-RoPE by default, and thus Chat-UniVi directly inherits this design. In contrast, our PPE strategy compresses visual tokens while still retaining the positional information of merged tokens within a single representation. This design leads to consistently stronger results under identical reduction ratios, highlighting that PPE is more effective than M-RoPE in preserving positional cues during token compression.

Comparison of MLLMs with Intact and Compressed Tokens. We compare our PPE with representative dense MLLMs, including LLaVA-OneVision Li et al. (2024a), InternVL2.5 Chen et al. (2024), and Qwen2.5-VL Bai et al. (2025), as well as token compression approaches such as PACT Dhouib et al. (2025) and SparseVLM Zhang et al. (2024a), as shown in Table 2. All baseline

Table 2: Comparison of MLLMs with intact tokens and compressed tokens.

Model	VideoMME (w/o subs)	VideoMME (w subs)	MVBench	MMBench (EN)	MMBench (CN)	TextVQA	Reduction Ratio
LLVA-OneVision-0.5B	44.00	43.50	45.50	52.10	-	-	0%
InternVL2.5-4B	62.30	63.60	71.60	81.10	79.30	76.80	0%
Qwen2.5-VL-3B	61.50	67.60	67.00	79.10	78.10	79.30	0%
PACT-7B	57.60	-	-	80.30	-	75.00	67%
SparseVLM-7B	-	-	-	64.10	-	57.80	66%
PPE-3B (Ours)	**58.70**	**59.07**	**67.38**	**84.78**	**84.85**	**77.08**	55%
PPE*-3B (Ours)	58.48	58.52	67.35	-	-	-	90%

Table 3: Performance of cascade compression, PPE integrated *Before* and/or *Within*-LLM.

Before	*Within*	VideoMME (w/o subs)	VideoMME (w subs)	NeXT-QA (MC)	NeXT-QA (OE)	SEED-Bench -Video	MVBench	*Average*	Reduction Ratio
✗	✗	56.41	56.49	78.07	32.99	55.12	68.25	*57.91*	0%
✓	✗	58.70	59.07	78.42	32.61	55.98	67.38	*58.69*	55%
✓	✗	57.41	57.70	77.98	31.06	55.36	66.65	*57.69*	90%
✗	✓	58.48	58.52	78.20	32.20	56.11	67.35	*58.48*	90%

results are reported by their respective original papers. Note that Chat-UniVi Jin et al. (2024) is excluded due to missing benchmark results. Despite using only 3B parameters, PPE achieves competitive performance compared to both dense and compressed models, with a 55% token reduction ratio. PPE* refers to a cascade compression variant applied *within* the LLM (details in *Section 4.3*), which further improves the reduction ratio to 90% while maintaining comparable performance.

4.3 Cascade Compression

In addition to applying our PPE compression *before* the LLM, we further investigate its integration *within* the LLM to enable multi-layer cascade token compression. Specifically, we conduct a layer-wise insertion study using the Qwen2.5-VL-3B-Instruct model, which contains 36 transformer layers. Specifically, we insert PPE-based clustering modules at layers 11, 23, and 35, while keeping all other configurations unchanged.

As shown in Table 3, applying PPE *within* the LLM using a per-layer clustering ratio of 0.45 achieves a 90% token reduction while maintaining performance comparable to the 55%-reduction case. Under the same 90% reduction budget, the *within*-LLM configuration outperforms applying PPE only *before* the LLM, demonstrating that cascade compression across multiple stages enables higher compression ratios without prematurely collapsing shallow semantics.

4.4 Ablation Study

In this section, we study the insight of PPE, its compatibility with other clustering-based compression methods, inference efficiency, and the effect of retaining different numbers of token ID positions to verify its ability to preserve positional information. Further studies on different reduction ratios *A.1*, aggregation stages *A.2*, model size generalization *A.4*, backbone generalization *A.3*, task generalization *A.5*, etc, are provided in the *Appendix A*.

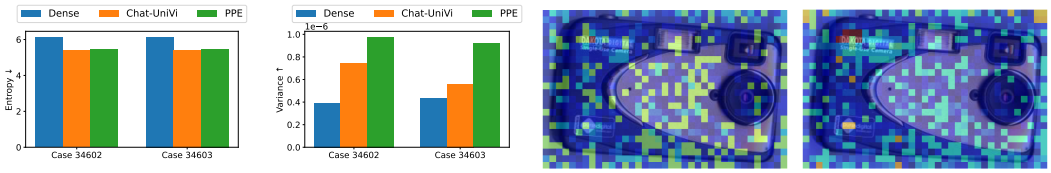
4.4.1 Insight into why PPE better preserves token semantics

The key insight of PPE lies in its ability to simultaneously encode multiple positions within a single token. Specifically, unlike traditional RoPE where multiple frequencies are rigidly bound to a single positional index, PPE introduces flexibility by allowing different frequencies to bind to different positional indices—as long as the corresponding tokens are considered similar in the clustering. As a result, during attention computation, each token can attend to multiple combinations of frequencies and positional indices. This enables each token pair to perceive a richer set of relative positional relationships, thereby preserving the global positional layout despite compression.

We randomly selected two samples and evaluated text–visual attention at the final LLM layer under a 55% token compression ratio. As shown in Figure 3(a–b), both Chat-UniVi and PPE reduce entropy

Table 4: Ablation of PPE integration with clustering-based compression and inference efficiency.

Method	MMBench (EN)	MMBench (CN)	TextVQA	LLM Generation (s) Time (s) ↓	Peak Memory Usage (GB) blue↓	Reduction Ratio
PACT	74.14	74.17	73.73	0.08	15.82	89%
PACT + PPE	**74.48**	**75.00**	**73.87**	0.09	15.82	89%
ToMe	74.31	73.63	74.94	0.90	15.81	57%
ToMe + PPE	**74.57**	**74.74**	**76.16**	0.91	15.81	57%



(a) Entropy comparison ↓ (b) Variance comparison ↑ (c) Chat-UniVi: "d" ✗ (d) PPE:"dakota digital" ✓

Figure 3: **Attention statistics and visualizations** of samples from TextVQA. (a–b) Quantitative comparison of entropy and variance. (c–d) Qualitative attention score visualizations of case 34602.

compared to the dense baseline, while PPE achieves similar entropy but consistently higher variance, indicating sharper and more confident grounding under compression. This is further illustrated in Figure 3(c–d): for case 34602 (question: *"What is the brand of this camera?"*), both models attend to the correct region, but Chat-UniVi's focus is narrowly confined due to positional information loss and answers incorrectly, whereas PPE preserves coverage across the text, recovering the full brand name. More samples and analyses are provided in *Appendix A.7*.

4.4.2 Integration with Other Clustering-Based Methods

PPE is compatible with other clustering-based methods. When integrated into the training-free PACT Dhouib et al. (2025) and ToMe Bolya et al. (2023b), experiments on Qwen2-VL-7B-Instruct Wang et al. (2024) show consistent improvements over the baselines (Table 4). This compatibility arises because PPE preserves the correspondence between tokens and their RoPE components: cluster averaging keeps embeddings close, and associating merged tokens with all original RoPE indices maintains positional information. When training is allowed, PPE further improves RoPE allocation, yielding more robust joint token-position representations and enhancing performance (e.g., Chat-UniVi + PPE, Table 1).

To further demonstrate PPE's generality across different compression frameworks, we conduct extensive experiments on VisionZip Yang et al. (2025), which combines pruning and merging and supports both training-free and SFT settings. We simply integrate PPE into VisionZip on Qwen2.5-VL-3B-Instruct following the official instructions ($dominant\_ratio$ and $contextual\_ratio$ are set 0.40 and 0.05 respectively to achieve 55% reduction ratio), randomly sampling 1/10 of our curated dataset for the SFT setting and only fine-tuning the projector for one epoch. The results are shown in Table 5. In both settings, PPE delivers improvements in most cases, especially on layout- and OCR-sensitive benchmarks, demonstrating that PPE retains more spatial information under the same reduction ratio. Moreover, when training is allowed, PPE achieves larger gains. Our method is decoupled from model size., extensive experiments are provided in *Appendix A.4*.

4.4.3 SFT or Training-Free

PPE is a plug-and-play, parameter-free method. Whether SFT is required depends on the underlying compression method, not PPE itself: 1) Chat-UniVi is a merging-based method that requires SFT, and performs bad in training-free setting especially on layout-sensitive benchmarks. 2) PACT and ToMe are fully training-free, and PPE is directly applied to the official pre-trained models without any additional training. 3) VisionZip supports both training-free and SFT usage, and PPE improves performance in both settings without modifying their token selection or aggregation mechanisms. This supports our key claim that PPE can enhance existing token compression pipelines in a plug-and-play manner and can further improve RoPE allocation when training is allowed.

Table 5: Extensive comparisons under both training-free and SFT using Qwen2.5-VL-3B-Instruct.

Setting	Benchmarks	Dense	Chat-UniVi	Chat-UniVi+PPE	VisionZip	VisionZip+PPE
Training-Free	**MMBench** (EN)	79.10	81.50	**82.28** (+0.78)	**83.48**	83.18 (-0.30)
	MMBench (CN)	78.10	80.06	**81.43** (+1.37)	**81.75**	81.64 (-0.11)
	TextVQA	79.30	37.60	**73.78** (+36.18)	79.00	**79.62** (+0.62)
	DocVQA	93.90	19.58	**66.16** (+46.58)	83.98	**85.63** (+1.65)
	OCRBench	797	307	**598** (+291)	713	**725** (+12)
	ChartQA	84.00	18.72	**67.08** (+48.52)	79.72	**80.72** (+1.00)
SFT	**MMBench** (EN)	85.89	**84.92**	84.73 (-0.19)	**83.99**	83.78 (-0.21)
	MMBench (CN)	86.07	83.71	**84.87** (+1.16)	82.12	**82.63** (+0.15)
	TextVQA	79.50	57.66	**77.14** (+19.48)	80.02	**82.06** (+2.04)
	DocVQA	89.44	52.48	**76.79** (+24.31)	84.84	**90.52** (+5.68)
	OCRBench	761	535	**691** (+156)	711	**780** (+69)
	ChartQA	79.96	49.60	**74.52** (+24.92)	80.20	**82.36**(+2.16)
	Reduction Ratio	0%	55%	55%	55%	55%

Table 6: Ablation study on performance and retained ID ratio with varying K.

K	VideoMME (w/o subs)	VideoMME (w subs)	NeXT-QA (MC)	NeXT-QA (OE)	SEED-Bench -Video	MVBench	*Average*	Reduction Ratio	IDs Retained
1	57.74	58.04	77.88	28.77	55.07	**68.08**	*57.97*	55%	45%
8	**58.70**	**59.07**	**78.42**	**32.61**	55.98	67.38	***58.69***	55%	77%
24	58.19	58.56	78.02	31.64	55.52	67.73	*58.28*	55%	84%

4.4.4 Analysis of Inference Efficiency

PPE introduces no additional parameters and incurs negligible computational overhead. Since it operates by redistributing existing RoPE dimensions, the computational and memory savings primarily result from token reduction itself. To ensure fairness and reproducibility, we reused the official PACT code to report runtime and memory usage, as presented in Table 4. With same reduction ratios, PPE does not need extra parameter and introduce very few inference time demonstrating its efficiency.

4.4.5 Effect of K on Performance and Retained ID Ratio

The hyperparameter K determines how many token ID positions are retained after merging, thus affecting spatiotemporal preservation. As shown in Table 6 and Figure 1(b), $K = 1$ causes severe degradation due to insufficient positional information, while $K = 24$ introduces redundancy and slightly underperforms $K = 8$. The choice of $K = 8$ achieves the best trade-off, aligning with the GCD of the 3D M-RoPE dimensions [16, 24, 24] to preserve essential signals without over-retention, and is adopted as our default. Moreover, Table 6 shows that larger K values yield higher valid ID retention which refers to the ratio of distinct positional embeddings preserved after compression; for $K > 1$, the retention rate exceeds the visual token retention ratio (45%), confirming that PPE captures positional information beyond token counts. Further visualization analyses for different choices of K are presented in *Appendix A.10*.

5 Conclusion

In this work, we propose **P**ositional **P**reservation **E**mbedding (PPE), a simple yet effective operator for retaining spatiotemporal positional information during visual token compression in MLLMs. PPE encodes fine-grained spatial and temporal cues into each compressed token and is parameter-free, plug-and-play, and compatible with existing pipelines without architectural changes. Experiments on MMBench, TextVQA, and VideoMME show consistent $2\% \sim 5\%$ gains, validating its effectiveness and generality. PPE also supports cascade clustering for progressive compression, offering a flexible trade-off between efficiency and performance. Our results underscore the importance of positional information in improving MLLM efficiency.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Annual Conference on Neural Information Processing, NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations, ICLR*, 2023a.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024. doi: 10.48550/ARXIV.2412.05271. URL https://doi.org/10.48550/arXiv.2412.05271.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2025.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021.

Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. Based Syst.*, 99:135–145, 2016. doi: 10.1016/J.KNOSYS.2016.02.001. URL https://doi.org/10.1016/j.knosys.2016.02.001.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024. doi: 10.48550/ARXIV.2405.21075. URL https://doi.org/10.48550/arXiv.2405.21075.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference of Computer Vision,ECCV*, 2024.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision-language models. *CoRR*, abs/2510.23095, 2025. doi: 10.48550/ARXIV.2510.23095. URL https://doi.org/10.48550/arXiv.2510.23095.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. Search-based neural structured learning for sequential question answering. In *Annual Meeting of the Association for Computational Linguistics, ACL*, 2017.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models - from vision, language to multimodality. *CoRR*, abs/2505.18227, 2025. doi: 10.48550/ARXIV.2505.18227. URL https://doi.org/10.48550/arXiv.2505.18227.

Yuwon Lee. Qwen2-vl-finetune, 2024. URL https://github.com/2U1/Qwen2-VL-Finetune.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *CoRR*, abs/2504.10462, 2025. URL https://arxiv.org/abs/2504.10462.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023a. doi: 10.48550/ARXIV.2307.16125. URL https://doi.org/10.48550/arXiv.2307.16125.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023b. doi: 10.48550/ARXIV.2305.06355. URL https://doi.org/10.48550/arXiv.2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2024b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Annual Conference on Neural Information Processing, NeurIPS*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference of Computer Vision,ECCV*, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of OCR in large multimodal models. *Sci. China Inf. Sci.*, 67(12), 2024b. doi: 10.1007/S11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *International Conference on Learning Representations, ICLR*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL https://doi.org/10.18653/v1/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 2199–2208. IEEE, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://doi.org/10.1109/WACV48630.2021.00225.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2024.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Annual Conference on Neural Information Processing, NeurIPS*, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19792–19802. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01843. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_VisionZip_Longer_is_Better_but_Not_Necessary_in_Vision_Language_CVPR_2025_paper.html.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. doi: 10.1162/TACL_A_00166. URL https://doi.org/10.1162/tacl_a_00166.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying SAM2 with llava for dense grounded understanding of images and videos. *CoRR*, abs/2501.04001, 2025. doi: 10.48550/ARXIV.2501.04001. URL https://doi.org/10.48550/arXiv.2501.04001.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2023.

Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. *CoRR*, abs/2504.10465, 2025. URL https://arxiv.org/abs/2504.10465.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *CoRR*, abs/2410.04417, 2024a. doi: 10.48550/ARXIV.2410.04417. URL https://doi.org/10.48550/arXiv.2410.04417.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *CoRR*, abs/2410.02713, 2024b.

A APPENDIX

A.1 ANALYSIS OF DIFFERENT REDUCTION RATIOS

We conduct experiments under different reduction ratios, as shown in Table 7. Overall, we observe that moderate compression tends to preserve the most useful positional information while eliminating redundancy. In particular, a reduction ratio of 55% consistently yields the best results across most benchmarks. When the reduction is too aggressive (e.g., 90%), the performance drops significantly due to the loss of critical visual content. On the other hand, smaller reductions do not fully exploit the potential for compression and efficiency. Based on these observations, we adopt a default clustering ratio of 0.45 (i.e., 55% reduction) in all our experiments, balancing efficiency and effectiveness.

Table 7: Ablation study comparing different reduction ratios.

Reduction Ratio	VideoMME (w/o subs)	VideoMME (w subs)	NeXT-QA (MC)	NeXT-QA (OE)	SEED-Bench -Video	MVBench	*Average*
25%	58.41	58.22	78.53	29.37	**56.77**	**68.03**	*58.22*
40%	57.30	57.52	78.15	30.99	56.59	67.40	*57.99*
55%	**58.70**	**59.07**	78.42	**32.61**	55.98	67.38	***58.69***
70%	57.74	57.78	**78.61**	32.17	55.90	67.38	*58.26*
90%	57.41	57.70	77.98	31.06	55.36	66.65	*57.69*

A.2 ANALYSIS OF AGGREGATION STAGES

We conduct experiments to compare two aggregation strategies: a three-stage setting (with spatial clustering ratios 0.25/0.5/0.5) and a single-stage setting (with a clustering ratio of 0.45). As shown in Table 8, the single-stage approach consistently outperforms the three-stage counterpart under comparable reduction ratios. Based on this observation, we adopt single-stage spatial clustering with a ratio of 0.45 as the default configuration throughout our experiments.

Table 8: Ablation study comparing different aggregation stages.

Aggregation Stages	VideoMME (w/o subs)	VideoMME (w subs)	NeXT-QA (MC)	NeXT-QA (OE)	SEED-Bench -Video	MVBench	*Average*	Reduction Ratio
Three-Stage	57.30	57.04	78.14	32.33	**56.06**	66.95	*57.97*	57%
Single-Stage	**58.70**	**59.07**	**78.42**	**32.61**	55.98	**67.38**	***58.69***	55%

A.3 ANALYSIS OF BACKBONE GENERALIZATION

To evaluate the generalizability of PPE across backbones, we additionally test it on the LLaVA-OV-0.5B model. As shown in Table 9, the Dense baseline reaches 44.8%, while PPE achieves 44.74% with 96% token reduction, showing strong generalization. PPE also outperforms Chat-UniVi (43.07%), especially in preserving spatiotemporal information.

Table 9: Ablation study on backbone generalization, conducted on the VideoMME (w/o subs) benchmark using the LLaVA-OV-0.5B (SI) model.

Method	Short	Medium	Long	*Average*	Reduction Ratio
Dense	54.70	<u>43.20</u>	<u>36.20</u>	*<u>44.80</u>*	0%
Chat-UniVi	53.00	40.40	**35.80**	*43.07*	96%
PPE (Ours)	**<u>56.40</u>** (+3.40)	**42.10** (+7.70)	35.70 (-0.10)	***44.74*** (+1.67)	96%

A.4 ANALYSIS OF MODEL SIZE

Our method is decoupled from model size. To further demonstrate that PPE scales with model size, we additionally evaluate PPE on Qwen2.5-VL-7B-Instruct, results are shown in Table 10, showing

Table 10: Ablation study on model size using Qwen2.5-VL-7B-Instruct.

Setting	Benchmarks	Dense	Chat-UniVi	Chat-UniVi+PPE	VisionZip	VisionZip+PPE
Training-Free	**MMBench** (EN)	83.50	83.23	**83.58** (+0.35)	**85.24**	85.19 (-0.05)
	MMBench (CN)	83.40	80.18	**82.35** (+2.17)	83.62	**83.92** (+0.30)
	TextVQA	84.90	35.82	**63.44** (+27.42)	84.64	**85.02** (+0.40)
	DocVQA	95.70	27.06	**66.42** (+39.36)	87.30	**88.52** (+1.22)
	OCRBench	864	479	**577** (+98)	780	**794** (+14)
	ChartQA	**87.30**	19.92	**46.72** (+29.90)	**60.56**	59.88 (-0.68)
SFT	**MMBench** (EN)	**86.90**	86.23	**86.26** (+0.03)	86.14	**87.24** (+0.16)
	MMBench (CN)	**85.35**	84.25	**84.85** (+0.60)	85.12	**85.15** (0.03)
	TextVQA	87.20	54.92	**82.46** (+27.54)	87.08	**87.24** (+0.16)
	DocVQA	92.97	50.01	**85.84** (+35.83)	93.16	**93.57** (+0.41)
	OCRBench	826	584	**764 (+180)**	**828**	**828** (+0)
	ChartQA	**86.32**	43.96	**78.88** (+34.92)	85.40	**86.08** (+0.68)
	Reduction Ratio	0%	55%	55%	55%	55%

that PPE provides improvements regardless of model capacity. For the SFT setting, we trained it on only 1/5 of the data used for the 3B model due to time and resource constraints.

We also note that in the training-free setting of VisionZip, PPE shows a slight performance drop on ChartQA for the 7B model, while improving the 3B model. This may be because the stronger 7B baseline already leverages fine-grained local patterns for such task, and without SFT, the additional positional details introduced by PPE can slightly perturb the attention distribution. After SFT, PPE consistently brings positive gains. For instance, the slight drop on VisionZip+PPE for ChartQA becomes an improvement, indicating that PPE benefits more when SFT is applied.

A.5 Analysis of Task Generalization

PPE is inherently task-agnostic and can be seamlessly integrated into any clustering-based token merging method, regardless of downstream task. While our main experiments focus on MLLM QA tasks—largely due to their popularity and the convenience of evaluation—we emphasize that PPE is not limited to QA scenarios.

To further demonstrate its generalization capability, we conducted additional experiments on the image captioning task using Flickr30k Young et al. (2014) benchmark. We reused the same models from Table 1 and directly evaluated their zero-shot captioning performance without any further fine-tuning. The results are in Table 11.

	Bleu_1	Bleu_2	Bleu_3	Bleu_4	METEOR	ROUGE_L	SPICE	CIDEr
Dense	0.311	0.188	0.120	0.079	0.147	0.322	0.219	0.722
+ Chat-UniVi	0.280	0.167	0.104	0.068	0.137	0.311	**0.214**	0.649
+ PPE	**0.313** (+0.033)	**0.185** (+0.018)	**0.117** (+0.013)	**0.076** (+0.008)	**0.144** (+0.007)	**0.312** (+0.001)	0.211 (-0.003)	**0.690** (+0.041)

Table 11: Ablation study on task generalization beyond QA.

From the table above, as expected, Dense achieves the best overall performance. Chat-UniVi, without PPE, suffers a noticeable performance drop. For instance, CIDEr decreases from 0.722 to 0.649, reflecting reduced content relevance. BLEU scores also decline, suggesting degraded coherence due to spatial misalignment caused by compression. SPICE shows a slight decrease (0.219 $\rightarrow$ 0.214), indicating minor semantic loss. With PPE, performance is largely restored across most metrics: BLEU-1/4 improve from 0.280/0.068 to 0.313/0.076, surpassing Dense in BLEU-1, and CIDEr rises from 0.649 to 0.690, narrowing the gap with Dense. Additionally, qualitative cases further illustrate PPE's effectiveness, as presented in Appendix A.6.

Both quantitative and qualitative results confirm PPE's ability to recover structural and positional cues often lost during naive token merging. While it does not fully close the gap with Dense, PPE consistently mitigates performance degradation.

A.6 Qualitative Cases on Image Captioning

We randomly select two representative cases from Flickr30k, as shown in Figure 4.

(a) Case 1: 10287332, roof scene

(b) Case 2: 900144365, marathon scene

Figure 4: **Qualitative examples for the image captioning task**, from Flickr30k benchmark.

Case 1 (10287332.jpg). The task prompt is: "Provide a one-sentence caption for the provided image. Do not provide any explanation." The ground truth caption is: "Two men sitting on the roof of a house while another one stands on a ladder." The Dense model predicts: "Three men work on a roof with a blue sky in the background." Under compression, Chat-UniVi simplifies it to: "Three men work on a roof," omitting background and clothing cues. In contrast, PPE preserves richer details: "Three men working on a roof with one wearing an orange shirt," showing stronger spatial recall under compression.

Case 2 (900144365.jpg). The ground truth caption is: "Marathon runners are racing on a city street, with other people standing around." The Dense model predicts: "A woman with number 1397 is running in a race." Chat-UniVi predicts: "A woman with number 1947 on her shirt is running in a race." PPE outputs: "A woman with the number 247 on her shirt is running in a race." All models focus primarily on the runners' numbers. However, there are multiple runners, including men and women, making the predictions less accurate. In the foreground, two women runners are the main focus of the image. The number of the front runner is only partially visible as "27", while the runner behind wears "1597". Consequently, the Dense model appears to focus on the runner behind and all models misrecognize the number.

A.7 Attention Visualization Analysis

Using the same models from Table 1, we extract the attention maps of the final LLM layer (layer_id=35) between all text and visual tokens and project them back onto the original images as heatmaps, revealing each question's focus across image regions. Both Chat-UniVi and PPE are under a reduction ratio of 55%. We randomly selected some cases to demonstrate our PPE can preserve valid positional information and focus more on the key region of images.

Case 1. MMBench (EN), (a–c) in Figure 5. Official QA index: `1505`. Question: *"How many types of fruits are there in the image? Please give the answer directly. A. 3 B. 2 C. 5 D. 4 E. None."* The image contains five fruit types: a central banana, surrounded by an apple, avocado, pear, and a partially visible orange in the background. The correct answer is *C*. PPE predicts correctly, while Chat-UniVi outputs *D*.

Attention heatmaps show that Chat-UniVi partly focuses on the background and fails to capture the orange, likely leading to the incorrect prediction. In contrast, PPE successfully attends to all fruit regions, including the occluded orange, demonstrating better spatial alignment under token compression and improved grounding for object counting in complex scenes.

Case 2. TextVQA, (d–f) in Figure A.7. Official image name: `af67237a4a504a29.jpg`, question ID: 37890. Question: *"What number is visible on the shorts in front?"* The image depicts three football players on a grassy field. The player in the foreground, jumping mid-air, wears shorts with

(a) Original (b) Chat-UniVi (c) PPE

(d) Original (e) Chat-UniVi (f) PPE

(g) Original (h) Chat-UniVi (i) PPE

Figure 5: **Qualitative comparison of attention visualizations.** Each row corresponds to a different sample: (a, d, g) original input, (b, e, h) Chat-UniVi outputs, and (c, f, i) PPE outputs.

the number ”2”. Behind him, another player shows a partially visible ”14”, of which only the ”1” is clear. The correct answer is *2*. Both Chat-UniVi and PPE incorrectly output *1*.

Heatmaps indicate that models attend to the ”2” on the foreground player, the partial ”1” from the second player, and even the text ”gen” on an advertising board. PPE places more emphasis on the ”2”, while Chat-UniVi is more distracted by background text. Although both fail in this case, the richer attention of PPE may sometimes introduce ambiguity, but also reflects broader contextual awareness.

Case 3. TextVQA, (g–i) in Figure A.7. Official image name: `23183bf2db16d88c.jpg`, question ID: 37376. Question: *”What kind of drink is this?”* The image shows an orange beverage bottle labeled with the brand *maaza*. The ground-truth answer is *maaza*. Both the Dense model and Chat-UniVi predict correctly, while PPE outputs *”orange juice”*.

Visualizations reveal that PPE’s attention covers both the brand logo and the broader bottle region. Although its answer deviates from the ground truth, it remains semantically reasonable, suggesting that PPE’s broader positional retention can capture additional contextual cues beyond the exact label.

A.8 Limitations and Failure Cases

(a) Case 1

(b) Case 2

Figure 6: **Representative failure cases of PPE**, from TextVQA benchmark.

As shown in Tables 1 and 7, PPE underperforms Dense in certain cases, highlighting the trade-off between compression and fidelity. We visualize representative failure cases to illustrate PPE’s limitations.

Using the same models from Table 1 (Qwen2.5-VL-3B-Instruct), we extract attention maps from the final LLM layer (layer_id=35) between text and visual tokens and visualize them as heatmaps. Under identical compression ratios (55% and 90%), we analyze two failure examples from the TextVQA benchmark:

Case 1 (55% compression). Official image: 23183bf2db16d88c.jpg; Question: “What kind of drink is this?” (question ID: 37376). The image shows an orange beverage labeled “maaza”. Ground truth: “maaza”. Dense and Chat-UniVi predict correctly, while PPE outputs “orange juice”. Heatmaps indicate that PPE attends to both the logo and the bottle, producing a semantically relevant but less precise answer. This suggests that PPE’s positional retention captures broader context, though at the cost of exactness.

Case 2 (90% compression). Official image: 0b9a494c78985373.jpg; Question: “What word is written on the collar of the jacket?” (question ID: 36877). The image contains multiple words: “kutxa” (collar), “menabi” (chest), “gipuzkoa”, and partial “deba”. Ground truth: “kutxa”. Dense predicts correctly; Chat-UniVi outputs “menabi”; PPE outputs “deb?”. Heatmaps show that both

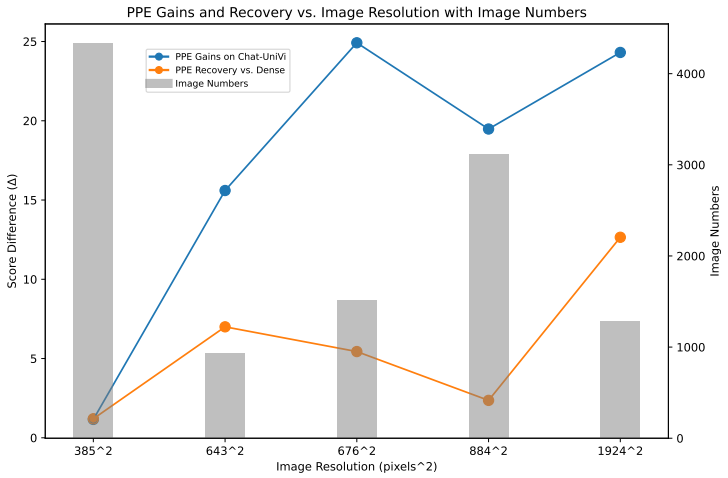


Figure 7: PPE obtains consistent improvements across different resolutions.

Chat-UniVi and PPE attend to larger central regions, missing the smaller collar text. This demonstrates a limitation of PPE under aggressive compression, where maintaining spatial precision becomes challenging.

A.9 PPE Gains across Different Image Resolutions

To evaluate the effect of PPE under varying image resolutions, we calculated the average resolution of TextVQA, MMBench, DocVQA, OCRBench, and ChartQA, based on the resolution statistics of their samples. As illustrated in Figure 7, the horizontal axis denotes the average image resolution for each dataset, while the left vertical axis presents the performance difference (Δ). This metric reflects both the improvements introduced by PPE over Chat-UniVi and its ability to recover performance relative to the dense model. For OCRBench, whose raw score range is not from 0 to 100, scores are normalized by a factor of 10 to ensure consistent visualization. The right vertical axis represents the number of image samples in each dataset.

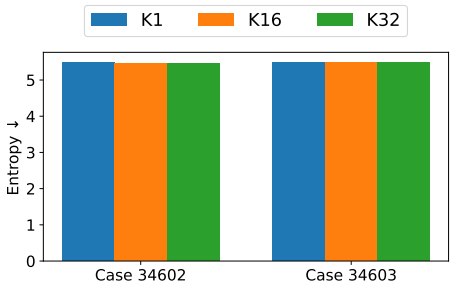
From the figure, we observe that PPE consistently provides notable performance gains across a wide spectrum of image resolutions: from 385^2 to 1924^2. These results demonstrate that PPE imposes no limitations on input resolution and remains robust and effective across images of different scales.

A.10 Visualization of the Effect of K

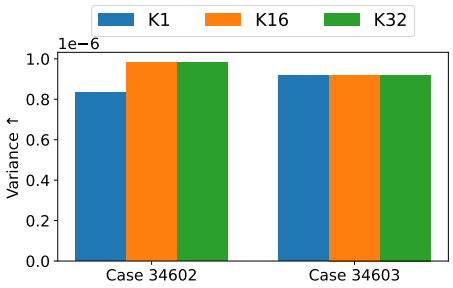
To better understand how different values of K influence attention behavior, we conduct a visual analysis and support it with quantitative indicators such as entropy and variance. Following the analytical framework introduced in Section 4.4.1 and Figure 3, we compare different choices of K directly on the images.

As shown in Figure 8, when $K = 1$, entropy is slightly higher and variance is lower, indicating that the model's attention becomes more dispersed compared with $K = 16$ or $K = 32$. The visualizations further reveal that at $K = 1$, the model mainly focuses on the first half of the brand text (with brighter activations in red). Meanwhile, tokens on the camera backplate also become more activated, showing that attention is distracted by background regions and fails to capture the full brand string, unlike the cases with $K = 16$ or $K = 32$. For images, we set $K = 32$ based on the GCD of M-RoPE configuration $[32, 32]$, which maximizes the preservation of positional information per merged token. The same rationale applies to video settings, $K = 8$ of $[16, 24, 24]$.

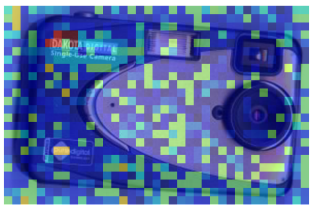
M-RoPE splits embedding dimensions into multiple periodic components (temporal, vertical, horizontal), assigning each axis contiguous frequency bands. Existing analysis (e.g., Figure 3 and 4 in Reference Huang et al. (2025)) shows that uneven allocation can lead to some axes occupying only high- or low-frequency channels, impairing long-range and spatial modeling. Choosing K based on



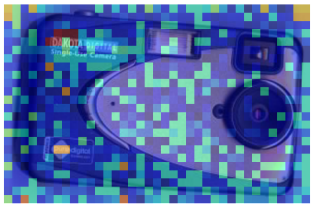
(a) Entropy comparison ↓



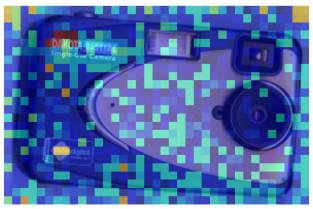
(b) Variance comparison ↑



(c) PPE (K=1)



(d) PPE (K=16)



(e) PPE (K=32)

Figure 8: **Attention statistics and visualizations** of samples from TextVQA. (a–b) Quantitative comparison of entropy and variance. (c–e) Qualitative attention score visualizations of case 34602.

the GCD ensures that each merged token preserves multiple positional anchors evenly across the frequency spectrum, preventing collapse into only high- or low-frequency bands. In short, K is grounded in M-RoPE's frequency structure, providing theoretical justification for PPE's ability to preserve positional information after token merging.

In short, refer to the analysis of positional encoding in VLMs Huang et al. (2025), choosing an appropriate K ensures that positional indices are distributed evenly across frequency bands; otherwise, some indices may collapse into only high- or low-frequency ranges, degrading positional fidelity.