MEDSCISSOR: TOWARDS GENERAL AND EFFICIENT MEDICAL IMAGE UNDERSTANDING WITH VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language models have shown substantial promise across various medical tasks. However, the large number of visual tokens generated from medical images introduces significant computational and inference overhead. Existing approaches either rely on internal attention signals, complicating their direct compatibility with efficient attention mechanisms such as FlashAttention, or are primarily designed for 3D medical imaging, limiting their applicability to broader medical imaging scenarios. To address these limitations, we propose **MedScissor**, a general and efficient visual token compression framework for medical vision-language models. Specifically, MedScissor first combines the semantic similarity score and neighborhood consistency score of medical visual tokens to identify redundant relationships among tokens, and then progressively merges related groups until the target compression budget is reached. Finally, MedScissor jointly considers local centrality score and global deviation score to retain one representative token with complementary information from each group. Extensive experiments demonstrate the effectiveness of MedScissor across diverse scenarios. On six representative 2D medical imaging benchmarks, MedScissor reduces the number of visual tokens by **50%** while retaining **98.6%** of the original performance, achieving a **1.4**× prefill speedup and a **49.2%** reduction in KV cache usage. On two representative 3D medical imaging benchmarks, MedScissor reduces the number of visual tokens by **70%** while retaining **99.0%** of the original performance, achieving a **2.1**× prefill speedup and a **65.6%** reduction in KV cache usage. *Our code is available in the supplementary materials, and all data will be released on GitHub.*

1 INTRODUCTION

The rapid development of vision-language models (VLMs) (Liu et al., 2023; Bai et al., 2025) has driven the emergence of medical vision-language models (Medical VLMs) (Tu et al., 2024; Sellergren et al., 2025). By integrating visual and textual information, Medical VLMs can jointly model medical images and clinical texts, demonstrating great potential in medical visual question answering (Li et al., 2023; Zhang et al., 2023), assisted diagnosis (Moor et al., 2023; Zhou et al., 2024), and radiology report generation (Wang et al., 2023; Deperrois et al., 2026). However, as the resolution of medical images continues to increase, Medical VLMs need to process an ever-growing number of visual tokens, as shown in Figure 1(a), posing significant computational and inference challenges.

Although visual token compression methods have achieved significant progress in general image (Xing et al., 2024; Zhang et al., 2026c) and video domains (Fan et al., 2026; Lin et al., 2026), medical images exhibit unique structural characteristics and information distributions. Existing compression methods still suffer from the following ***three key limitations*** when applied to Medical VLMs:

(I) Lack of Medical Adaptation. Critical diagnostic regions in medical images often exhibit low visual saliency and local sparsity, making generic visual token compression methods (Bolya et al., 2022; Chen et al., 2024) prone to removing clinical evidence when transferred to medical scenarios.

(II) Limited Scenario Coverage. Existing medical visual token compression methods are primarily designed for 3D medical imaging (Liu et al., 2026; Yi et al., 2026), while their direct application to broader 2D medical imaging scenarios may not achieve optimal performance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

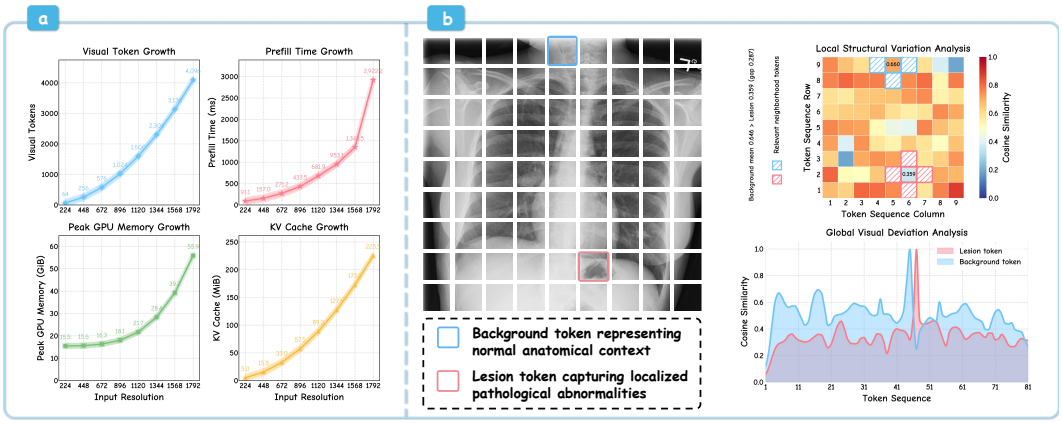

Figure 1: **(a) Significant inference overhead from high-resolution medical images.** As image resolution increases, inference overhead such as *KV Cache usage* and *Prefill Time* increase significantly. **(b) Local structural variation and global visual deviation of medical visual tokens.** In the heatmap, lesion tokens exhibit *lower neighborhood similarity* than background tokens, while their *overall similarity to the image token sequence* is also generally lower than that of background tokens.

(III) Poor Deployment Compatibility. Existing methods often rely on attention signals from the vision encoder (Liu et al., 2026; Yi et al., 2026) or access internal LLM information (Chen et al., 2024; Zhang et al., 2024) for token compression, increasing adaptation complexity and limiting compatibility with inference frameworks such as FlashAttention.

Therefore, developing an architecture-friendly and generalizable visual token compression method for diverse medical scenarios remains an important challenge. To bridge this gap, we analyze visual token distributions in medical images and identify ***two key characteristics:***

(I) Local Structural Variation. As shown in Figure 1(b), lesion tokens exhibit lower neighborhood similarity than background tokens. In this example, the average similarity is only **0.359** for the lesion token, compared with **0.634** for background tokens. This indicates that ***lesion regions disrupt local spatial continuity, exhibiting stronger local structural variation***.

(II) Global Visual Deviation. As shown in Figure 1(b), the background token at **(5,9)** exhibits higher similarity to the entire image token sequence than the lesion token at **(6,2)** across most positions. This indicates that ***lesion tokens tend to deviate more from the global visual pattern dominated by background regions, exhibiting stronger global visual deviation***.

Consequently, our analysis indicates that effective medical visual token compression should consider both local structural variation and global visual deviation. To this end, we propose **MedScissor**, a general and efficient visual token compression framework for Medical VLMs. Specifically, MedScissor leverages the **Semantic-Neighborhood Grouping** module to combine the semantic similarity score and the neighborhood consistency score, thereby identifying redundant relationships among visual tokens. We then apply the **Progressive Group Merging** module to iteratively merge tokens into groups until the target compression budget is achieved. Furthermore, we introduce the **Representative Token Selection** module, which jointly considers the local centrality score and the global deviation score, thereby selecting one original token from each group while preserving representative and complementary visual information. In summary, the main contributions are as follows:

- **Existing Limitation Analysis.** We identify **three key limitations** of existing visual token compression methods for Medical VLMs, including insufficient medical adaptation, limited scenario coverage, and poor deployment compatibility in practical applications.
- **General Compression Framework.** Based on the above analysis, we propose **MedScissor**, a general medical visual token compression framework that leverages local structural variation and global visual deviation to effectively preserve critical visual information.
- **Excellent Empirical Performance.** Across six representative 2D and two representative 3D medical imaging benchmarks, MedScissor retains **98.6%** and **99.0%** of the original performance while reducing visual tokens by **50%** and **70%**, respectively, achieving **1.4**× and **2.1**× prefill speedups with **49.2%** and **65.6%** reductions in KV cache usage.

2 Related Work

2.1 Medical Vision-Language Models

Medical vision-language models (Li et al., 2023; Sellergren et al., 2025) have rapidly advanced by jointly modeling medical images and clinical texts, integrating visual information with medical knowledge. With the development of general vision-language models (Li et al., 2022; Liu et al., 2023), Medical VLMs have demonstrated strong potential in medical visual question answering (Li et al., 2023), assisted diagnosis (Zhou et al., 2024), and radiology report generation (Wang et al., 2023). Early Medical VLMs adapted general vision-language models to the medical domain through medical image-text data and instruction tuning (Li et al., 2023; Sun et al., 2024). Subsequently, medical multimodal foundation models further incorporated medical knowledge and multimodal data, enhancing generalization across complex medical tasks (Deria et al., 2026; Zhang et al., 2026b). However, high-resolution medical images often produce numerous visual tokens, increasing the computational and memory overhead of Medical VLM inference (Chen et al., 2026; Liu et al., 2026; Zhu et al., 2026; Yi et al., 2026). Therefore, reducing visual tokens while preserving critical information has become an important direction for improving Medical VLM efficiency.

2.2 Visual Token Compression

Visual token compression methods (Dorovatas et al., 2025; Huang et al., 2025; Liu et al., 2025; Zhang et al., 2026a) aim to reduce the number of tokens while preserving critical visual information, thereby reducing computational and inference overhead. Based on their compression strategies, existing methods can mainly be categorized into three types: attention-based, similarity-based, and diversity-based approaches. Attention-based methods (Chen et al., 2024; Zhang et al., 2024) leverage internal attention mechanisms to estimate token importance and retain informative tokens. Similarity-based methods (Bolya et al., 2022; Shang et al., 2025) reduce redundancy by merging semantically similar tokens, while diversity-based methods (Wen et al., 2025; Alvar et al., 2025) preserve representative tokens by modeling information differences among tokens. However, these methods are primarily designed for natural image and video scenarios, overlooking the unique characteristics of medical images. Recently, medical visual token compression methods (Chen et al., 2026; Liu et al., 2026; Zhu et al., 2026; Yi et al., 2026) have attracted increasing attention, but existing approaches mainly focus on 3D medical imaging scenarios (Liu et al., 2026; Zhu et al., 2026; Yi et al., 2026), while compression methods for broader 2D medical imaging scenarios remain underexplored.

3 Methodology

3.1 Overview

The overview of MedScissor is illustrated in Figure 2. The framework consists of three modules, enabling efficient compression while preserving representative and complementary information.

3.2 Semantic-Neighborhood Grouping

Given an input medical image I, we first obtain its projected visual token features by sequentially applying the vision encoder $E(\cdot)$ and visual projector $M(\cdot)$:

$$X = M(E(I)) = \{x_i\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^D, \tag{1}$$

where N and D denote the number of visual tokens and the feature dimension, respectively.

For a 2D medical image, the visual tokens form an $H \times W$ grid with $N = H \times W$, where H and W denote the grid height and width. For a 3D medical image, the tokens form an $H \times W \times Z$ grid with $N = H \times W \times Z$, where Z denotes the depth dimension. We then perform L_2 normalization on each token feature, obtaining $\hat{x}_i = x_i / \|x_i\|_2$.

Based on spatial positions, we construct a local adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For 2D images, token (h, w) is connected to its horizontal and vertical neighbors $(h \pm 1, w)$ and $(h, w \pm 1)$. For 3D images, we additionally connect depth neighbors (h, w, z) and $(h, w, z \pm 1)$.

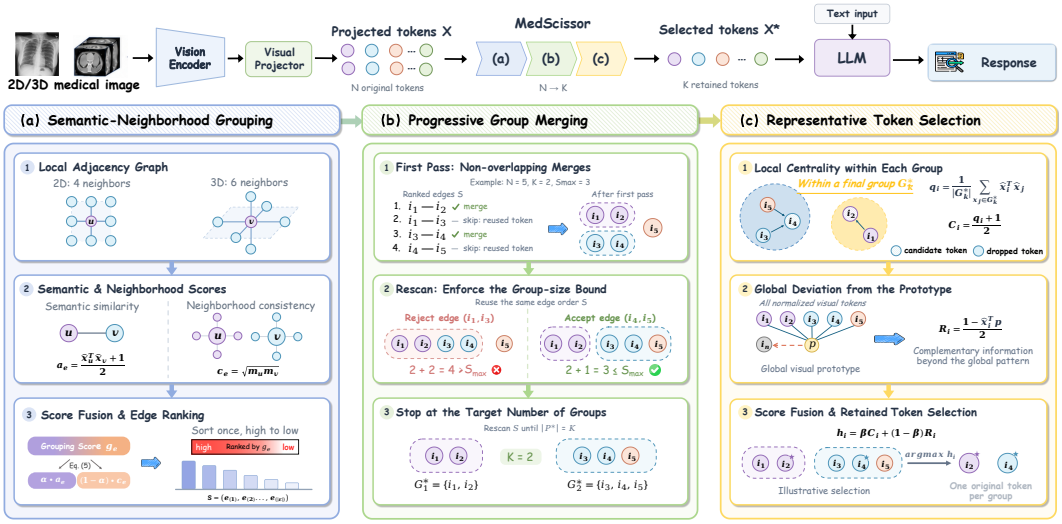

Figure 2: **The overview of MedScissor.** MedScissor exploits semantic and spatial characteristics of medical images to remove redundant visual tokens while retaining critical visual information.

For each local edge $e = (u, v) \in \mathcal{E}$, we define the *semantic similarity score* as:

$$a_e = \frac{\hat{x}_u^{\mathsf{T}} \hat{x}_v + 1}{2}, \tag{2}$$

where $a_e \in [0, 1]$ denotes the normalized cosine similarity between adjacent tokens u and v. A higher a_e indicates that the two tokens exhibit stronger semantic consistency, making them more suitable to be connected and grouped together.

We further compute the average neighborhood support of token i:

$$m_i = \frac{1}{\deg(i)} \sum_{e \ni i} a_e, \tag{3}$$

where $\deg(i)$ denotes the number of neighbors of token i. For an edge $e = (u, v)$, the *neighborhood consistency score* is defined as:

$$c_e = \sqrt{m_u m_v}. \tag{4}$$

A higher c_e indicates that both tokens maintain strong consistency with their local neighborhoods, suggesting that they are more likely to lie in a region with similar and redundant features.

Finally, we combine the two scores to obtain the *semantic-neighborhood score*:

$$g_e = \alpha a_e + (1 - \alpha) c_e, \tag{5}$$

where α is a weighting coefficient that balances the two scores. We sort all local edges according to g_e in descending order and obtain the ordered edge sequence $\mathcal{S} = \big(e_{(1)}, e_{(2)}, \ldots, e_{(|\mathcal{E}|)}\big)$, where $g_{e_{(1)}} \geq g_{e_{(2)}} \geq \cdots \geq g_{e_{(|\mathcal{E}|)}}$. The sequence $\mathcal{S}$ is then used as the input to the next module.

3.3 Progressive Group Merging

Given the ordered edge sequence $\mathcal{S}$, we progressively merge visual tokens into target groups. Given a compression ratio r, the target number of retained tokens is $K = \lfloor N(1 - r) \rfloor$. Initially, each token forms an independent group, denoted as $\mathcal{P} = \{G_i\}_{i=1}^{N}$, where $G_i = \{x_i\}$.

During merging, for two connected groups G_u and G_v, we only perform merging when the resulting group size satisfies:

$$|G_u| + |G_v| \leq S_{\max} = \gamma / (1 - r), \tag{6}$$

where γ controls the maximum size of a single token group. The two groups are then merged into a new group $G_{uv} = G_u \cup G_v$.

First, we scan the ordered edge sequence $\mathcal{S}$ and prioritize forming non-overlapping token groups. For an edge $e_{(i)} = (u, v) \in \mathcal{S}$, the merging operation is performed only when both u and v have

not participated in merging. After this stage, the partition is updated as $\mathcal{P}' = \{G'_k\}$. This strategy restricts each token to participate in at most one merge, preventing a highly similar token from continuously absorbing multiple neighbors and forming oversized local groups.

If the number of groups remains larger than K, we rescan $\mathcal{S}$ and continue merging. At this stage, each edge $e_{(j)} = (s, t) \in \mathcal{S}$ connects the current groups containing tokens s and t. If the merging constraint is satisfied, we merge the two connected groups into a new group as $G'_{st} = G'_s \cup G'_t$.

This process iteratively merges groups along $\mathcal{S}$ until the number of groups is reduced to K. Throughout the process, we directly reuse the edge ordering generated by the semantic-neighborhood score without recomputing inter-group similarity.

Finally, we obtain K visual token groups, denoted as:

$$\mathcal{P}^* = \{G^*_k\}_{k=1}^K. \tag{7}$$

The resulting partition $\mathcal{P}^*$ is subsequently fed into the next module.

3.4 Representative Token Selection

Given the token partition $\mathcal{P}^* = \{G^*_k\}_{k=1}^K$, we select one representative token from each group to preserve major visual information while reducing redundancy. For each candidate token x_i within a group G^*_k, we first define the *local centrality score* as:

$$q_i = \frac{1}{|G^*_k|} \sum_{x_j \in G^*_k} \hat{x}_i^\mathsf{T} \hat{x}_j, \quad C_i = \frac{q_i + 1}{2}, \tag{8}$$

where q_i denotes the average cosine similarity between token x_i and all tokens within the same group, and C_i denotes the normalized local centrality score. A higher C_i indicates that the token exhibits stronger consistency with other tokens in the group, making it more representative of the overall visual information within the group.

However, relying on local centrality may overlook tokens that deviate from the global pattern but contain complementary information. Therefore, we define the *global deviation score* as:

$$R_i = \frac{1 - \hat{x}_i^\mathsf{T} p}{2}, \tag{9}$$

where p denotes the global visual prototype computed from all normalized visual tokens. A higher R_i indicates that the token deviates more from the global visual pattern, potentially containing complementary local visual information.

Finally, we combine the two scores to obtain the *representative selection score*:

$$h_i = \beta C_i + (1 - \beta) R_i, \tag{10}$$

where β is a weighting coefficient that balances the two scores. For each token group G^*_k, we select the token with the highest h_i as the representative token:

$$x^*_k = \arg\max_{x_i \in G^*_k} h_i. \tag{11}$$

Finally, each group retains one original visual token. This process directly preserves original visual tokens without generating new feature representations, reducing token numbers while maintaining representative and complementary visual information.

Algorithm: MedScissor Compression Pipeline

Require: Medical image I; encoder $E(\cdot)$; projector $M(\cdot)$; compression ratio r
Ensure: Compressed visual tokens X^*

$X \leftarrow M(E(I)) = \{x_i\}_{i=1}^N$
// Step 1: Semantic-Neighborhood Grouping
Construct adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
for each $e = (u, v) \in \mathcal{E}$ **do**
  $a_e \leftarrow (\hat{x}_u^\mathsf{T} \hat{x}_v + 1)/2$
  $c_e \leftarrow \sqrt{m_u m_v}$
  $g_e \leftarrow \alpha a_e + (1 - \alpha) c_e$
end for
$\mathcal{S} \leftarrow \mathrm{Sort}(\mathcal{E}, g_e)$
// Step 2: Progressive Group Merging
$K \leftarrow \lfloor N(1 - r) \rfloor$
$\mathcal{P} \leftarrow \{\{x_i\}\}_{i=1}^N$
for each $e_{(j)} = (u, v) \in \mathcal{S}$ **do**
  if $|G_u| + |G_v| \le S_{\max}$ **then**
    $G_{uv} \leftarrow G_u \cup G_v$
  end if
end for
while $|\mathcal{P}| > K$ **do**
  Rescan $\mathcal{S}$ and merge valid groups
end while
$\mathcal{P}^* \leftarrow \{G^*_k\}_{k=1}^K$
// Step 3: Representative Token Selection
for each $G^*_k \in \mathcal{P}^*$ **do**
  $C_i \leftarrow (q_i + 1)/2$
  $R_i \leftarrow (1 - \hat{x}_i^\mathsf{T} p)/2$
  $h_i \leftarrow \beta C_i + (1 - \beta) R_i$
  $x^*_k \leftarrow \arg\max_{x_i \in G^*_k} h_i$
end for
$X^* \leftarrow \{x^*_k\}_{k=1}^K$

Table 1: **Performance comparison of baselines on Lingshu-7B at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. **%** represents the performance retention rate. Best results for each ratio are highlighted in **bold black font**.

Method	PMC-VQA	PathVQA	SLAKE	VQA-RAD	MMMU-Med	MedXQA	Avg.	%
Lingshu-7B	63.0	57.1	77.0	53.2	54.8	26.5	55.3	100.0%
Visual token compression ratio (↓ 50%)								
FastV [ECCV24]	61.3	55.7	74.1	50.6	49.7	25.7	52.8	95.5%
SparseVLM [ICML25]	61.2	54.3	74.7	52.8	51.3	**26.4**	53.5	96.7%
ToMe [ICLR23]	60.7	56.2	75.7	52.1	51.0	26.2	53.6	97.0%
LLaVA-PruMerge [ICCV25]	56.9	54.6	70.1	52.8	49.0	25.8	51.5	93.2%
DART [EMNLP25]	60.8	**57.0**	76.1	51.2	51.3	**26.4**	53.8	97.3%
DivPrune [CVPR25]	**62.1**	55.9	75.8	52.6	52.3	26.0	54.1	97.8%
MedPruner [MICCAI26]	61.2	55.1	**76.4**	52.8	**52.5**	25.6	53.9	97.5%
CARVE [Arxiv26]	60.5	54.7	75.6	53.0	50.9	26.2	53.5	96.7%
MedScissor	61.8	56.5	76.0	**54.3**	52.4	26.2	**54.5**	**98.6%**
Visual token compression ratio (↓ 70%)								
FastV [ECCV24]	59.4	54.3	69.2	47.2	46.6	25.8	50.4	91.2%
SparseVLM [ICML25]	58.5	53.1	73.3	53.0	50.4	25.3	52.3	94.5%
ToMe [ICLR23]	58.0	55.5	74.4	52.6	49.5	25.5	52.6	95.1%
LLaVA-PruMerge [ICCV25]	55.5	54.0	67.6	**53.7**	47.8	25.1	50.6	91.5%
DART [EMNLP25]	58.2	55.8	75.2	51.7	49.1	**26.1**	52.7	95.2%
DivPrune [CVPR25]	59.8	55.3	73.6	51.7	50.1	25.6	52.7	95.3%
MedPruner [MICCAI26]	59.8	54.2	74.6	51.2	50.4	24.9	52.5	94.9%
CARVE [Arxiv26]	58.7	54.7	73.9	49.2	50.1	25.3	52.0	94.0%
MedScissor	**59.9**	**56.2**	**75.3**	53.4	**50.5**	25.7	**53.5**	**96.7%**
Visual token compression ratio (↓ 90%)								
FastV [ECCV24]	49.8	44.9	55.4	42.6	43.2	24.6	43.4	78.5%
SparseVLM [ICML25]	54.4	51.3	65.6	49.5	**47.5**	25.0	48.9	88.4%
ToMe [ICLR23]	55.6	53.1	**71.0**	50.8	47.4	24.8	50.4	91.2%
LLaVA-PruMerge [ICCV25]	52.1	52.2	63.6	50.8	46.4	25.3	48.4	87.5%
DART [EMNLP25]	52.3	**53.7**	69.3	51.2	47.4	25.4	49.9	90.2%
DivPrune [CVPR25]	54.9	53.3	69.0	49.0	47.1	24.9	49.7	89.8%
MedPruner [MICCAI26]	55.6	52.5	70.1	47.0	46.9	24.3	49.4	89.3%
CARVE [Arxiv26]	53.4	53.1	68.4	47.7	46.7	24.6	49.0	88.5%
MedScissor	**57.2**	52.7	70.8	**52.1**	47.4	**25.6**	**51.0**	**92.2%**

4 Experiments

4.1 Experimental Setup

Baselines. We select FastV (Chen et al., 2024) and SparseVLM (Zhang et al., 2024) as representative attention-based methods, ToMe (Bolya et al., 2022) and LLaVA-PruMerge (Shang et al., 2025) as similarity-based token merging methods, and DART (Wen et al., 2025) and DivPrune (Alvar et al., 2025) as diversity-based pruning methods. Additionally, we compare with medical visual token compression methods, including MedPruner (Liu et al., 2026) and CARVE (Yi et al., 2026). More details on these methods are provided in Appendix A.1.

Datasets. For 2D medical imaging scenarios, we select six medical visual question answering benchmarks, including PMC-VQA (Zhang et al., 2023), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), MMMU-Med (Yue et al., 2024), and MedXpertQA-MM (Zuo et al., 2025), covering multimodal medical image understanding, pathology analysis, radiology question answering, and medical reasoning. For 3D medical imaging scenarios, we select two benchmarks, M3D-VQA (Bai et al., 2024) and 3D-RAD (Gai et al., 2026), covering 3D volume understanding and radiology diagnostic reasoning. More details are provided in Appendix A.2.

Implementation Details. All experiments are conducted on a server with 8 NVIDIA A100 (80GB) GPUs. For 2D medical imaging scenarios, we select Lingshu-7B (Xu et al., 2026) as the base model, while Hulu-Med-7B (Jiang et al., 2025) is used for 3D medical imaging scenarios. The key hyperparameters are set as follows: $\alpha = 0.5$, $\beta = 0.7$, and $S_{\max} = \gamma/(1-r)$. All experimental results are averaged over three runs. More details are provided in Appendix A.3.

Table 2: **Performance comparison of baselines on Hulu-Med-7B at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. **%** represents the performance retention rate. Best results for each ratio are highlighted in **bold black font**.

Method	M3D-VQA	3D-RAD			Avg.	%
	All	Open	Close	All		
Hulu-Med-7B	81.7	25.0	79.0	71.5	76.6	100.0%
Visual token compression ratio (↓ 50%)						
FastV [ECCV24]	80.8	24.0	78.3	70.8	75.8	99.0%
SparseVLM [ICML25]	80.9	24.0	78.4	70.9	75.9	99.1%
ToMe [ICLR23]	80.8	24.1	78.4	70.9	75.9	99.1%
LLaVA-PruMerge [ICCV25]	80.8	24.0	78.4	70.9	75.9	99.1%
DART [EMNLP25]	80.6	23.9	78.4	70.8	75.7	98.8%
DivPrune [CVPR25]	80.7	23.6	78.1	70.5	75.6	98.7%
MedPruner [MICCAI26]	81.0	24.0	78.6	**71.1**	**76.1**	**99.3%**
CARVE [Arxiv26]	81.0	24.0	78.6	71.0	76.0	99.2%
MedScissor	**81.1**	24.1	78.7	**71.1**	**76.1**	**99.3%**
Visual token compression ratio (↓ 70%)						
FastV [ECCV24]	79.7	23.1	78.2	70.5	75.1	98.0%
SparseVLM [ICML25]	79.8	23.1	78.2	70.6	75.2	98.2%
ToMe [ICLR23]	79.8	23.7	77.7	70.3	75.1	98.0%
LLaVA-PruMerge [ICCV25]	80.3	23.7	78.3	70.8	75.6	98.6%
DART [EMNLP25]	80.0	23.1	78.3	70.7	75.4	98.4%
DivPrune [CVPR25]	79.6	23.3	77.3	69.8	74.7	97.5%
MedPruner [MICCAI26]	**80.7**	24.5	78.4	**71.0**	**75.9**	**99.1%**
CARVE [Arxiv26]	80.5	24.4	78.3	70.9	75.7	98.8%
MedScissor	80.6	24.5	78.4	70.9	75.8	99.0%
Visual token compression ratio (↓ 90%)						
FastV [ECCV24]	77.9	20.7	78.3	70.4	74.2	96.9%
SparseVLM [ICML25]	78.0	20.8	78.3	70.3	74.2	96.9%
ToMe [ICLR23]	78.4	23.2	77.7	70.1	74.3	97.0%
LLaVA-PruMerge [ICCV25]	78.2	22.5	78.3	70.6	74.4	97.1%
DART [EMNLP25]	78.0	21.0	78.4	70.4	74.2	96.9%
DivPrune [CVPR25]	78.3	22.3	76.1	68.7	73.5	96.0%
MedPruner [MICCAI26]	78.5	23.6	78.2	70.6	74.6	97.4%
CARVE [Arxiv26]	**78.6**	23.6	78.2	**70.7**	**74.7**	**97.5%**
MedScissor	78.5	23.5	78.1	70.6	74.6	97.4%

4.2 Main Results

Performance Comparison on Lingshu-7B. Table 1 presents the performance comparison of various baselines on Lingshu-7B for 2D medical imaging tasks. The results can be summarized in three points: **(i) Outstanding Performance.** On six 2D medical benchmarks, MedScissor maintains **98.6%** of the original performance with **50%** token compression. Furthermore, MedScissor achieves higher retention rates than other baselines across compression ratios. **(ii) Poor Transferability.** General visual token compression methods often struggle when directly applied to medical imaging. For example, FastV achieves only **78.5%** performance retention under a **90%** token compression ratio, resulting in substantial degradation. **(iii) Cross-Domain Mismatch.** Token compression methods specifically designed for 3D medical imaging may not generalize well to 2D scenarios. For example, at a **90%** token compression ratio, MedPruner and CARVE retain only **89.3%** and **88.5%** of the original performance, respectively, even lower than general methods such as ToMe and DART.

Performance Comparison on Hulu-Med-7B. Table 2 presents the performance comparison of various baselines on Hulu-Med-7B for 3D medical imaging tasks. The experimental results can be summarized in two points: **(i) Limited performance degradation.** In 3D medical imaging tasks, all methods exhibit minor performance drops under different compression ratios, with small gaps among baselines. This indicates substantial token redundancy in 3D medical images, allowing key task information to be preserved even under aggressive compression. **(ii) Cross-scenario competitiveness.** On 3D medical imaging tasks, MedScissor achieves comparable performance to MedPruner and CARVE, which are specifically designed for 3D scenarios. Meanwhile, it significantly outperforms these methods on 2D medical imaging tasks, demonstrating stronger cross-scenario generalization.

Table 3: **Impact of different Semantic-Neighborhood Grouping strategies under different compression ratios across eight representative medical imaging benchmarks.**

Method	PMC-VQA	PathVQA	SLAKE	VQA-RAD	MMMU-Med	MedXQA	M3D-VQA	3D-RAD	Avg.	%
Baseline	63.0	57.1	77.0	53.2	54.8	26.5	81.7	71.5	60.6	100.0%
Visual token compression ratio (↓ 50%)										
w/o a_e	61.6	**56.8**	**76.3**	53.0	**52.4**	**26.3**	80.9	70.9	59.8	98.7%
w/o c_e	60.8	56.5	76.2	53.9	52.2	26.2	81.0	71.0	59.7	98.5%
Ours	**61.8**	56.5	76.0	**54.3**	**52.4**	26.2	**81.1**	**71.1**	**59.9**	**98.8%**
Visual token compression ratio (↓ 70%)										
w/o a_e	59.6	**56.3**	74.8	52.8	50.1	25.9	80.3	70.7	58.8	97.0%
w/o c_e	59.4	56.4	74.9	53.2	**51.5**	**26.0**	80.4	70.8	**59.1**	**97.5%**
Ours	**59.9**	56.2	**75.3**	**53.4**	50.5	25.7	**80.6**	**70.9**	**59.1**	**97.5%**
Visual token compression ratio (↓ 90%)										
w/o a_e	55.2	52.4	69.5	52.3	**48.0**	25.5	78.5	70.2	56.5	93.2%
w/o c_e	56.4	**53.2**	70.6	**52.6**	47.0	**25.8**	**78.6**	70.5	56.8	93.7%
Ours	**57.2**	52.7	**70.8**	52.1	47.4	25.6	78.5	**70.6**	**56.9**	**93.9%**

Table 4: **Impact of different Representative Token Selection strategies under different compression ratios across eight representative medical imaging benchmarks.**

Method	PMC-VQA	PathVQA	SLAKE	VQA-RAD	MMMU-Med	MedXQA	M3D-VQA	3D-RAD	Avg.	%
Baseline	63.0	57.1	77.0	53.2	54.8	26.5	81.7	71.5	60.6	100.0%
Visual token compression ratio (↓ 50%)										
Random	60.7	**56.5**	75.7	50.8	51.0	26.2	81.0	**71.4**	59.2	97.7%
w/o C_i	**62.0**	**56.5**	**76.0**	54.1	52.2	26.1	81.0	71.0	**59.9**	**98.8%**
w/o R_i	60.7	55.8	75.8	53.4	**52.5**	**26.3**	**81.3**	71.3	59.6	98.3%
Ours	61.8	**56.5**	**76.0**	**54.3**	52.4	26.2	81.1	71.1	**59.9**	**98.8%**
Visual token compression ratio (↓ 70%)										
Random	58.1	55.4	74.7	51.7	49.5	25.4	**80.7**	**71.1**	58.3	96.2%
w/o C_i	**59.9**	56.0	74.8	**54.3**	49.6	25.9	80.5	70.4	58.9	97.2%
w/oR_i	58.6	55.6	73.9	52.8	49.4	**26.1**	80.6	**71.1**	58.5	96.6%
Ours	**59.9**	**56.2**	**75.3**	53.4	**50.5**	25.7	80.6	70.9	**59.1**	**97.5%**
Visual token compression ratio (↓ 90%)										
Random	55.1	52.4	69.2	47.0	47.2	24.8	77.2	**70.8**	55.5	91.6%
w/o C_i	54.9	**53.0**	69.7	51.4	47.2	25.2	**79.0**	70.6	56.4	93.1%
w/o R_i	55.7	52.5	68.4	49.5	**48.6**	24.4	78.4	70.7	56.0	92.4%
Ours	**57.2**	52.7	**70.8**	**52.1**	47.4	**25.6**	78.5	70.6	**56.9**	**93.9%**

4.3 Ablation Study

4.3.1 MedScissor Ablation Study

Semantic-Neighborhood Grouping. Table 3 shows that, under three compression ratios, our method consistently outperforms two variants: one removes the *semantic similarity score* a_e, while the other removes the *neighborhood consistency score* c_e. This demonstrates the effectiveness of jointly modeling semantic similarity and neighborhood consistency for *Semantic-Neighborhood Grouping*.

Representative Token Selection. Table 4 shows that, under three compression ratios, our method consistently outperforms three variants: one removes the *local centrality score* C_i, one removes the *global deviation score* R_i, and the other randomly selects representative tokens from each group. These results demonstrate the effectiveness of jointly considering local centrality and global deviation for *Representative Token Selection*, preserving diverse visual information effectively.

4.3.2 Depth Connectivity Analysis

Depth Connectivity Analysis. Table 5 shows that our method consistently outperforms the variant without depth-wise connectivity under different compression ratios. For example, under a **70%** compression ratio, introducing depth-aware connectivity improves the performance from **79.9%** to **80.6%** on M3D-VQA and from **70.7%** to **70.9%** on 3D-RAD. This indicates that extending spatial

Table 5: **Impact of Depth Connectivity on 3D medical imaging benchmarks.**

Method	M3D-VQA			3D-RAD		
Compression Ratio	↓ 50%	↓ 70%	↓ 90%	↓ 50%	↓ 70%	↓ 90%
w/o Depth Connectivity	80.6	79.9	78.1	70.9	70.7	70.1
Ours	**81.1**	**80.6**	**78.5**	**71.1**	**70.9**	**70.6**

connectivity from 2D to 3D by incorporating depth-aware connections is important for modeling 3D medical structures and achieving effective visual token compression.

4.3.3 Inference Efficiency Analysis

Table 6: **Quantitative comparison of inference efficiency metrics: FLOPs (TFLOPs/sample), LLM Prefill Time (ms/sample), Peak KV-Cache (MiB), and Peak GPU-Memory (GiB).**

Method	FLOPs (↓)	LLM Prefill (↓)	Peak KV-Cache (↓)	Peak GPU-Memory (↓)
2D Medical Imaging Scenarios				
Baseline	11.0	76.3	356.6	16.8
MedScissor (↓ 50%)	6.0	54.7	181.2	**16.7**
MedScissor (↓ 70%)	**4.0**	**50.0**	**111.1**	**16.7**
3D Medical Imaging Scenarios				
Baseline	37.8	199.7	151.3	16.0
MedScissor (↓ 50%)	18.8	105.9	80.5	15.5
MedScissor (↓ 70%)	**11.6**	**95.8**	**52.1**	**15.3**

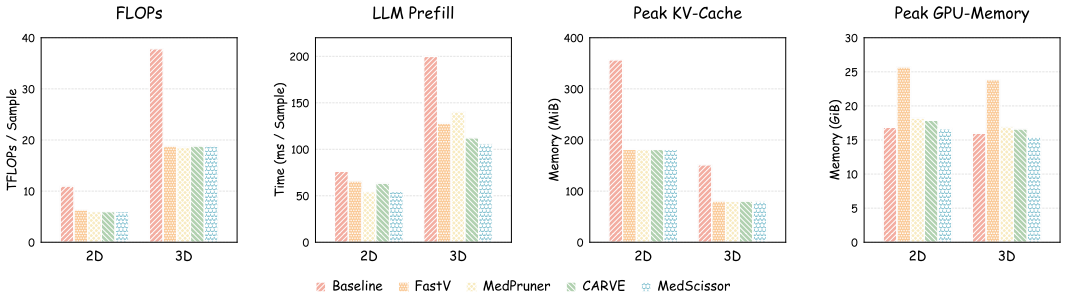

Figure 3: **Inference efficiency comparison among different baselines with 50% compression.**

Table 6 and Figure 3 compare the inference efficiency of MedScissor with different baselines, from which we draw two key observations: **(i) Excellent Efficiency Performance.** MedScissor achieves superior inference efficiency across diverse medical imaging scenarios. It reduces visual tokens by **50%** and **70%**, achieving **1.4**× and **2.1**× prefill speedups with **49.2%** and **65.6%** reductions in KV cache usage, respectively. **(ii) Better Deployment Compatibility.** MedScissor achieves comparable or superior performance across four inference efficiency metrics compared with other baselines. Unlike pruning methods such as FastV that rely on internal LLM attention signals, which are difficult to integrate with FlashAttention and may introduce higher GPU Memory overhead, MedScissor directly compresses visual tokens after the MLP layer, enabling easier deployment.

For more ablation studies, including ***comprehensive inference efficiency analysis***, ***parameter sensitivity analysis***, ***additional accuracy evaluation***, and ***case study***, please refer to Appendices B-E.

5 Conclusion

We propose MedScissor, a general and efficient medical visual token compression framework for medical vision-language models. Extensive experiments demonstrate the effectiveness of MedScissor across diverse medical imaging scenarios. On six representative 2D and two representative 3D medical imaging benchmarks, MedScissor reduces the number of visual tokens by 50% and 70% while retaining 98.6% and 99.0% of the original performance, achieving 1.4× and 2.1× prefill speedups with 49.2% and 65.6% reductions in KV cache usage, respectively.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with grammar polishing during the writing process. We did not use generative AI tools for research design, experimental analysis, interpretation of results, or the generation of the core contributions of this work. All AI-assisted content has been carefully reviewed and verified by the authors. We take full responsibility for the content of this work, including the text, research claims, and any artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the openness and reproducibility of this work. The complete experimental code is provided in the appendix, together with detailed reproduction guidelines to facilitate the verification and replication of our results.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Kaitao Chen, Weiqian Zhao, Jiamin Wu, Qihao Zheng, Shangquan Sun, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Token-sparse medical multimodal reasoning via dual-stream reinforcement learning. *arXiv preprint arXiv:2606.31599*, 2026.

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

Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, et al. Radvlm: A multitask conversational vision-language model for radiology. *Scientific Reports*, 2026.

Ankan Deria, Komal Kumar, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, and Imran Razzak. Medmo: Grounding and understanding multimodal large language model for medical images. *arXiv preprint arXiv:2602.06965*, 2026.

Vaggelis Dorovatas, Soroush Seifi, Gunshi Gupta, and Rahaf Aljundi. Recurrent attention-based token selection for efficient streaming video-llms. *arXiv preprint arXiv:2510.17364*, 2025.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*, 2026.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *Advances in Neural Information Processing Systems*, 38, 2026.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3328–3338, 2025.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36: 28541–28564, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. V-cast: Video curvature-aware spatio-temporal pruning for efficient video large language models. *arXiv preprint arXiv:2603.27650*, 2026.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shengyuan Liu, Zanting Ye, Yunrui Lin, Chen Hu, Wanting Geng, Xu Han, Bulat Ibragimov, Yefeng Zheng, and Yixuan Yuan. Medpruner: Training-free hierarchical token pruning for efficient 3d medical image understanding in vision-language models. *arXiv preprint arXiv:2603.11625*, 2026.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhenyu Yi, Qiang Hu, Zhenhao Li, Jiaxuan Zhao, Yusong Sun, and Lichi Zhang. Carve: Cross-slice anisotropic reallocation of visual evidence for efficient 3d medical volume understanding. *arXiv preprint arXiv:2608.04515*, 2026.

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

Evelyn Zhang, Fufu Yu, Aoqi Wu, Zichen Wen, Ke Yan, Shouhong Ding, Biqing Qi, and Linfeng Zhang. D^2pruner: Debiased importance and structural diversity for mllm token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12412–12420, 2026a.

Kai Zhang, Zhengqing Yuan, Cheng Peng, Songlin Zhao, Mengxian Lyu, Ziyi Chen, Yanfang Ye, Wei Liu, Ying Zhang, Kaleb E Smith, et al. Medgpt-oss: Training a general-purpose vision-language model for biomedicine. *arXiv preprint arXiv:2603.00842*, 2026b.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 2026c.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

Yitao Zhu, Mengjun Liu, Yingji Fu, Haowen Pang, and Anqi Qiu. Medarc: Training-free adaptive redundancy compression of visual tokens for 3d medical vision-language models. *arXiv preprint arXiv:2607.26554*, 2026.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

APPENDIX

A Other Experimental Details **13**

A.1 Baselines 13

A.2 Datasets 14

A.3 Additional Implementation Details 15

B Comprehensive Inference Efficiency Analysis **15**

C Parameter Sensitivity Analysis **16**

D Additional Accuracy Evaluation **17**

E Case study **20**

A OTHER EXPERIMENTAL DETAILS

A.1 BASELINES

FastV. FastV (Chen et al., 2024) is a training-free visual token pruning method based on the observation that visual-token attention becomes highly concentrated after the early layers of a large vision-language model. It processes the complete multimodal sequence before a selected LLM layer, ranks visual tokens according to their average received attention, and removes low-scoring tokens from all subsequent layers. This directly reduces the computational cost of both self-attention and feed-forward operations in the deeper LLM layers.

SparseVLM. SparseVLM (Zhang et al., 2024) is a training-free, text-aware visual token sparsification method. It first identifies text tokens that are relevant to the visual input and then uses the self-attention matrix to estimate the importance of each visual token. Visual tokens are progressively reduced across LLM layers, while a rank-based strategy dynamically determines the sparsification ratio at each layer. A token-recycling mechanism further compresses information from discarded tokens into compact representations to alleviate information loss.

ToMe. ToMe (Bolya et al., 2022) reduces computation by merging similar tokens instead of directly discarding them. At each selected transformer layer, tokens are divided into two sets and matched using a lightweight bipartite soft-matching procedure based on feature similarity. The most similar token pairs are merged through size-aware weighted averaging, and this process is applied progressively across layers. Consequently, redundant information is consolidated while the representative content of the original token sequence is retained.

LLaVA-PruMerge. LLaVA-PruMerge (Shang et al., 2025) performs adaptive visual token reduction through a combination of pruning and merging. It uses the sparse attention between the visual encoder's class token and spatial tokens, together with an interquartile-range outlier criterion, to adaptively identify informative visual tokens. The remaining tokens are assigned to nearby selected tokens according to feature similarity and merged using weighted averaging.

DART. DART (Wen et al., 2025) is a training-free duplication-aware token pruning method that argues token redundancy is a more reliable pruning criterion than conventional importance scores. It selects a small subset of visual tokens as pivots and measures the duplication of every remaining token with respect to these pivots. Tokens that are highly similar to the pivots are considered redundant, whereas tokens with low duplication are preferentially retained. This preserves diverse visual evidence while supporting regular dense attention computation after pruning.

DivPrune. DivPrune (Alvar et al., 2025) formulates visual token pruning as a Max-Min Diversity Problem. Given a target token budget, it selects a subset of visual tokens that maximizes the minimum pairwise distance among the retained tokens. The resulting subset therefore provides broad coverage

of the original visual representation while avoiding repeated or highly similar features. DivPrune is training-free and does not require calibration data or additional model parameters.

MedPruner. MedPruner (Liu et al., 2026) is a training-free hierarchical pruning framework designed for volumetric medical image understanding. It first applies inter-slice anchor-based filtering to remove anatomically redundant slices according to their relationship with representative anchor slices. It then performs Dynamic Information Nucleus Selection within the retained slices, using cumulative attention mass to adaptively determine how many visual tokens should be preserved. This hierarchical design addresses both slice-level redundancy and heterogeneous token-level information density.

CARVE. CARVE (Yi et al., 2026) is a training-free, budget-constrained visual token reduction framework for slice-based 3D medical volumes. It partitions the depth axis into coherent windows and allocates the global token budget non-uniformly according to normalized cross-slice evidence. Within each window, CARVE constructs spatial anchors from representative slices, retrieves locally varying evidence from the volume, and merges the remaining eligible tokens into nearby anchors.

A.2 DATASETS

PMC-VQA. PMC-VQA (Zhang et al., 2023) is a large-scale medical visual question answering benchmark constructed from the PubMed Central Open Access subset. Its scalable generation and filtering pipeline converts biomedical image–caption pairs into question–answer pairs covering diverse diseases, anatomical regions, and medical imaging modalities, with radiological images accounting for a large proportion. It also provides a manually verified test set for evaluating medical visual understanding and knowledge-intensive reasoning.

PathVQA. PathVQA (He et al., 2020) is a visual question answering benchmark developed specifically for pathology, with images collected from pathology textbooks and digital libraries. It is constructed through a semi-automatic pipeline that extracts images and captions, generates questions from textual information, and manually verifies the resulting questions. PathVQA evaluates pathological structure recognition and the integration of visual evidence with pathology knowledge.

SLAKE. SLAKE (Liu et al., 2021) is a semantically labeled and knowledge-enhanced medical visual question answering benchmark containing bilingual English–Chinese question–answer pairs. Its images span multiple radiological modalities, anatomical regions, and diseases, while its questions include both open-ended and closed-ended forms. It further provides physician-annotated semantic labels and a structured medical knowledge base to support visual recognition, medical knowledge reasoning, and cross-modal semantic understanding.

VQA-RAD. VQA-RAD (Lau et al., 2018) is a manually constructed radiology visual question answering benchmark with images sampled from MedPix, covering the head, chest, and abdomen across modalities such as X-ray, CT, and MRI. Its questions are naturally formulated and validated by clinicians and include both open-ended and closed-ended forms. It evaluates clinically relevant recognition, localization, abnormality identification, and radiological reasoning.

MMMU-Med. MMMU-Med (Yue et al., 2024) denotes the Health and Medicine portion of the MMMU benchmark for college-level multimodal understanding and deliberate reasoning. It covers Basic Medical Science, Clinical Medicine, Diagnostics and Laboratory Medicine, Pharmacy, and Public Health, with heterogeneous visual materials including radiological scans, clinical photographs, diagrams, charts, and tables. The benchmark requires models to combine fine-grained visual perception with domain-specific medical knowledge and multi-step reasoning.

MedXpertQA-MM. MedXpertQA-MM (Zuo et al., 2025) is the multimodal subset of MedXpertQA, designed to evaluate expert-level medical understanding and reasoning across medical specialties and body systems. Each question may combine medical images with patient histories, examination findings, and clinical evidence. Compared with caption-derived medical VQA benchmarks, it emphasizes realistic diagnostic integration, difficult medical decisions, and multi-step clinical reasoning.

M3D-VQA. M3D-VQA (Bai et al., 2024) is the visual question answering component of M3D-Data, a multimodal benchmark for 3D medical image analysis. Its questions are derived from 3D medical volumes and diagnostic reports and cover imaging plane, acquisition phase, organ, abnormality, and anatomical location. It supports both open-ended and closed-ended evaluation, requiring models to interpret volumetric anatomy and pathology rather than relying on individual 2D slices.

3D-RAD. 3D-RAD (Gai et al., 2026) is a large-scale radiology visual question answering benchmark built from 3D CT examinations. It defines six complementary tasks, including anomaly detection, image observation, medical computation, existence detection, static temporal diagnosis, and longitudinal temporal diagnosis. The benchmark contains both open-ended and closed-ended questions and evaluates spatial, computational, and multi-temporal clinical reasoning.

A.3 Additional Implementation Details

For Lingshu-7B and the six 2D medical imaging benchmarks, we follow the evaluation settings of the corresponding original papers (Xu et al., 2026). For Hulu-Med-7B and the two 3D medical imaging benchmarks, we likewise follow the original evaluation protocols (Jiang et al., 2025) and uniformly sample 32 slices along the depth dimension from each 3D medical image for testing.

B Comprehensive Inference Efficiency Analysis

Table 7: **Inference efficiency analysis with 50% compression: FLOPs (TFLOPs/sample), LLM Prefill Time (ms/sample), Peak KV-Cache (MiB), and Peak GPU-Memory (GiB).**

Method	FLOPs (↓)	LLM Prefill (↓)	Peak KV-Cache (↓)	Peak GPU-Memory (↓)
Baseline	17.7	107.1	305.3	16.6
FastV [ECCV24]	9.4	81.4	156.3	25.2
SparseVLM [ICML25]	11.0	85.3	156.0	28.3
ToMe [ICLR23]	9.2	71.9	156.1	**16.4**
LLaVA-PruMerge [ICCV25]	9.6	72.5	156.1	17.2
DART [EMNLP25]	9.7	75.6	156.0	16.5
DivPrune [CVPR25]	9.2	74.8	156.1	**16.4**
MedPruner [MICCAI26]	**9.1**	75.7	**155.8**	17.9
CARVE [Arxiv26]	9.2	75.7	156.1	17.5
MedScissor	9.2	**67.5**	156.0	**16.4**

Table 8: **Inference efficiency analysis with 70% compression: FLOPs (TFLOPs/sample), LLM Prefill Time (ms/sample), Peak KV-Cache (MiB), and Peak GPU-Memory (GiB).**

Method	FLOPs (↓)	LLM Prefill (↓)	Peak KV-Cache (↓)	Peak GPU-Memory (↓)
Baseline	17.7	107.1	305.3	16.6
FastV [ECCV24]	6.3	74.6	96.4	24.8
SparseVLM [ICML25]	7.1	74.3	96.4	27.8
ToMe [ICLR23]	5.9	66.0	96.3	**16.3**
LLaVA-PruMerge [ICCV25]	6.2	61.2	96.3	17.2
DART [EMNLP25]	6.7	67.9	96.5	16.4
DivPrune [CVPR25]	5.9	67.2	96.3	16.4
MedPruner [MICCAI26]	5.9	**58.8**	96.2	17.8
CARVE [Arxiv26]	**5.7**	68.4	**95.4**	17.5
MedScissor	5.9	61.4	96.3	**16.3**

Tables 7–9 further provide a comprehensive comparison of inference efficiency between MedScissor and all eight baselines under different compression ratios, covering four metrics: FLOPs, LLM Prefill Time, Peak KV-Cache, and Peak GPU-Memory. Overall, MedScissor achieves the best or near-best inference efficiency across different medical imaging scenarios and compression ratios, effectively reducing visual tokens while lowering computational and memory overhead. More importantly, MedScissor offers better deployment compatibility. FastV and SparseVLM rely on internal LLM attention signals, introducing additional memory overhead and limiting compatibility with efficient attention mechanisms such as FlashAttention. MedPruner and CARVE require internal signals from the vision encoder, while ToMe requires modifications to the vision encoder. In contrast, MedScissor compresses projected visual tokens directly after the MLP layer without accessing internal LLM or vision-encoder signals or modifying the original vision encoder. This enables efficient compression with low additional overhead and easier integration into existing Medical VLMs.

Table 9: **Inference efficiency analysis with 90% compression: FLOPs (TFLOPs/sample), LLM Prefill Time (ms/sample), Peak KV-Cache (MiB), and Peak GPU-Memory (GiB).**

Method	FLOPs (↓)	LLM Prefill (↓)	Peak KV-Cache (↓)	Peak GPU-Memory (↓)
Baseline	17.7	107.1	305.3	16.6
FastV [ECCV24]	3.2	53.2	40.3	25.1
SparseVLM [ICML25]	3.3	55.3	40.1	28.1
ToMe [ICLR23]	2.7	48.9	40.1	16.3
LLaVA-PruMerge [ICCV25]	2.9	50.7	40.1	17.2
DART [EMNLP25]	3.7	49.5	40.4	16.4
DivPrune [CVPR25]	2.7	49.6	40.1	16.3
MedPruner [MICCAI26]	2.7	45.9	**40.0**	17.8
CARVE [Arxiv26]	2.7	**44.1**	40.1	17.5
MedScissor	**2.6**	47.4	40.1	**16.2**

C Parameter Sensitivity Analysis

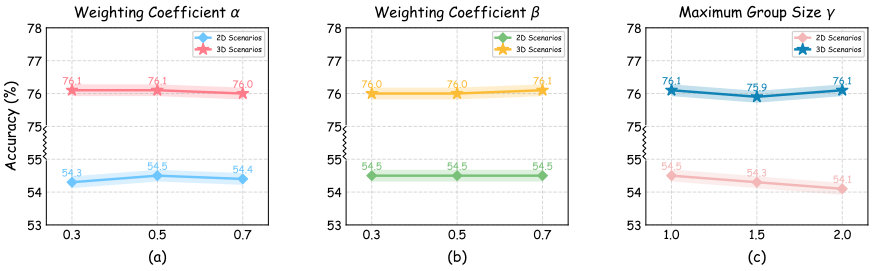

Figure 4: **Impact of different parameter values with 50% compression.**

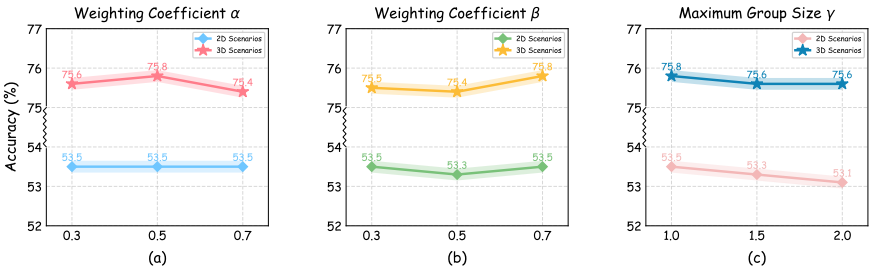

Figure 5: **Impact of different parameter values with 70% compression.**

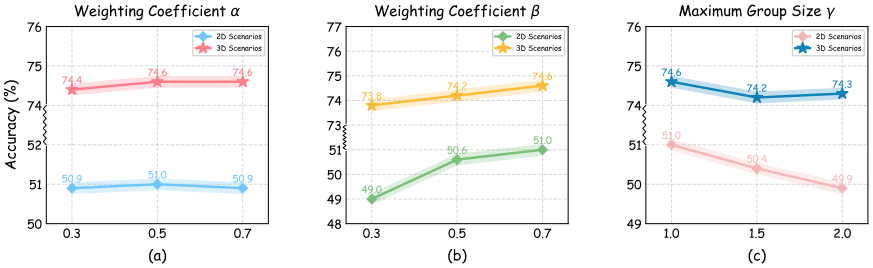

Figure 6: **Impact of different parameter values with 90% compression.**

Parameter sensitivity analyses under different compression ratios are presented in Figures 4–6, where the 2D and 3D results are averaged over six 2D benchmarks and two 3D benchmarks, respectively. Overall, the performance varies only slightly across different parameter settings and compression ratios, indicating that MedScissor is robust to hyperparameter choices and does not require extensive dataset-specific tuning. Based on the overall results across both 2D and 3D scenarios, we finally set $\alpha = \mathbf{0.5}$, $\beta = \mathbf{0.7}$, and $\gamma = \mathbf{1.0}$. Specifically, α controls the balance between semantic similarity and

neighborhood consistency. A smaller α weakens the contribution of semantic similarity, while a larger α reduces the effect of neighborhood consistency, making $\alpha = \mathbf{0.5}$ a balanced choice. The parameter β balances local centrality and global deviation during representative token selection. A smaller β overemphasizes global deviation while weakening within-group representativeness, leading to performance degradation, especially under high compression ratios. Finally, γ controls the maximum group size. A larger γ allows more tokens to concentrate in a few larger groups, which may reduce grouping balance and cause the loss of fine-grained visual information. Therefore, $\gamma = \mathbf{1.0}$ provides a more balanced grouping strategy across different medical imaging scenarios.

D Additional Accuracy Evaluation

Table 10: **Additional accuracy evaluation metrics of baselines on PathVQA at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. Best results for each ratio are highlighted in **bold black font**.

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
Lingshu-7B	25.9	88.2	30.2	33.4	17.5	13.2	11.2	30.2	17.5	57.1
Visual token compression ratio (↓ 50%)										
FastV [ECCV24]	24.7	86.8	29.0	32.2	16.7	12.7	10.8	29.0	16.6	55.7
SparseVLM [ICML25]	21.9	86.7	26.2	29.6	16.0	12.2	10.4	26.1	15.1	54.3
ToMe [ICLR23]	24.8	87.6	29.0	32.3	17.0	12.9	10.9	29.0	16.7	56.2
LLaVA-PruMerge [ICCV25]	22.4	86.8	26.6	29.8	15.8	12.0	10.3	26.6	15.2	54.6
DART [EMNLP25]	**26.4**	87.6	**30.5**	**33.6**	**17.2**	**13.0**	**10.9**	**30.4**	**17.4**	**57.0**
DivPrune [CVPR25]	24.5	87.2	28.8	32.0	16.9	12.8	10.8	28.8	16.6	55.9
MedPruner [MICCAI26]	22.6	87.6	26.7	30.2	16.4	12.5	10.6	26.6	15.5	55.1
CARVE [Arxiv26]	22.0	87.3	26.1	29.7	16.2	12.4	10.6	26.1	15.2	54.7
MedScissor	25.3	**87.7**	29.5	32.7	16.9	12.9	10.9	29.4	16.9	56.5
Visual token compression ratio (↓ 70%)										
FastV [ECCV24]	22.9	85.7	27.1	30.5	16.1	12.3	10.6	27.1	15.5	54.3
SparseVLM [ICML25]	20.2	85.9	24.2	27.7	15.1	11.7	10.1	24.1	13.9	53.1
ToMe [ICLR23]	23.7	**87.2**	27.9	31.2	16.7	12.7	10.8	27.9	16.2	55.5
LLaVA-PruMerge [ICCV25]	21.4	86.5	25.5	28.8	15.5	11.8	10.1	25.4	14.6	54.0
DART [EMNLP25]	24.8	86.7	29.0	32.2	16.7	12.6	10.7	29.0	16.6	55.8
DivPrune [CVPR25]	23.7	86.8	28.0	31.2	16.6	12.6	10.7	27.9	16.1	55.3
MedPruner [MICCAI26]	21.1	87.2	25.1	28.8	16.0	12.3	10.5	25.1	14.6	54.2
CARVE [Arxiv26]	22.6	86.7	26.5	30.1	16.2	12.3	10.5	26.5	15.4	54.7
MedScissor	**25.1**	87.2	**29.2**	**32.4**	**17.0**	**12.8**	**10.9**	**29.1**	**16.8**	**56.2**
Visual token compression ratio (↓ 90%)										
FastV [ECCV24]	9.0	80.7	12.5	17.6	12.0	10.1	9.1	12.5	7.6	44.9
SparseVLM [ICML25]	17.4	85.1	21.1	25.0	14.1	11.0	9.7	21.0	12.1	51.3
ToMe [ICLR23]	19.6	**86.5**	23.5	27.3	15.3	11.8	10.2	23.5	13.7	53.1
LLaVA-PruMerge [ICCV25]	19.0	85.5	22.8	26.5	14.6	11.3	9.8	22.8	13.1	52.2
DART [EMNLP25]	**21.8**	85.6	**25.5**	**29.0**	**15.4**	11.9	10.2	**25.4**	**14.5**	**53.7**
DivPrune [CVPR25]	21.1	85.5	24.7	28.4	15.4	**11.9**	10.2	24.6	14.2	53.3
MedPruner [MICCAI26]	18.6	86.3	22.5	26.3	14.8	11.6	10.1	22.5	13.0	52.5
CARVE [Arxiv26]	20.7	85.4	24.4	28.2	15.4	11.9	10.2	24.4	14.1	53.1
MedScissor	19.6	85.8	23.5	27.2	15.3	11.8	**10.2**	23.5	13.7	52.7

Table 11: **Additional accuracy evaluation metrics of baselines on SLAKE at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. Best results for each ratio are highlighted in **bold black font**.

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
Lingshu-7B	73.5	82.3	81.6	80.9	52.6	33.7	25.7	81.3	58.7	77.0
Visual token compression ratio (↓ 50%)										
FastV [ECCV24]	70.6	79.3	79.7	79.0	50.4	32.2	24.6	79.3	56.7	74.1
SparseVLM [ICML25]	70.1	81.8	78.6	77.8	50.1	32.1	24.4	78.3	56.2	74.7
ToMe [ICLR23]	72.1	81.2	80.3	79.7	51.4	33.1	**25.4**	80.0	57.5	75.7
LLaVA-PruMerge [ICCV25]	66.5	75.5	75.2	74.5	47.4	30.3	23.1	74.9	53.5	70.1

Continued on next page

Table 11 – continued from previous page

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
DART [EMNLP25]	**72.5**	81.4	**81.1**	**80.5**	51.7	33.1	25.3	**80.7**	**58.1**	76.1
DivPrune [CVPR25]	71.7	81.9	79.9	79.3	51.3	33.0	25.3	79.6	57.3	75.8
MedPruner [MICCAI26]	72.5	**82.4**	81.1	80.4	51.7	32.8	25.0	80.6	58.0	**76.4**
CARVE [Arxiv26]	72.1	80.8	80.6	80.1	51.1	32.7	25.0	80.1	57.2	75.6
MedScissor	72.2	81.8	80.6	80.0	**51.8**	**33.1**	25.3	80.1	57.8	76.0
Visual token compression ratio (↓ 70%)										
FastV [ECCV24]	66.5	73.2	75.8	75.4	47.7	30.8	23.6	75.4	53.5	69.2
SparseVLM [ICML25]	68.9	79.9	77.8	76.9	49.3	31.6	24.1	77.4	55.4	73.3
ToMe [ICLR23]	70.7	79.9	79.1	78.6	50.5	32.3	24.9	78.7	56.5	74.4
LLaVA-PruMerge [ICCV25]	63.9	73.1	72.9	72.3	45.5	28.9	22.1	72.6	51.3	67.6
DART [EMNLP25]	71.8	80.2	**80.5**	**79.9**	**51.4**	**32.9**	**25.3**	**80.0**	57.5	75.2
DivPrune [CVPR25]	69.7	79.4	78.5	77.8	49.8	31.8	24.4	78.2	55.8	73.6
MedPruner [MICCAI26]	70.3	**80.9**	79.5	78.8	50.5	32.0	24.5	79.0	56.7	74.6
CARVE [Arxiv26]	70.8	78.5	79.5	79.0	50.6	32.3	24.8	79.0	56.5	73.9
MedScissor	**71.9**	80.5	80.1	79.5	51.3	32.8	25.0	79.8	**57.5**	**75.3**
Visual token compression ratio (↓ 90%)										
FastV [ECCV24]	53.2	58.7	63.2	64.1	40.0	26.6	21.0	62.6	43.7	55.4
SparseVLM [ICML25]	60.4	73.4	70.2	69.6	45.1	29.0	22.2	69.9	50.5	65.6
ToMe [ICLR23]	**67.8**	75.8	**76.7**	**76.0**	**48.3**	**30.9**	23.6	**76.4**	**54.3**	**71.0**
LLaVA-PruMerge [ICCV25]	59.2	70.3	68.4	68.2	43.1	27.7	21.2	68.1	48.3	63.6
DART [EMNLP25]	66.2	74.0	74.9	74.8	47.8	30.8	**23.7**	74.4	53.0	69.3
DivPrune [CVPR25]	64.6	75.6	74.0	73.6	46.8	30.2	23.3	73.7	52.2	69.0
MedPruner [MICCAI26]	66.0	**76.4**	75.8	75.0	47.3	30.0	22.9	75.3	53.4	70.1
CARVE [Arxiv26]	65.2	73.3	74.4	74.1	47.1	30.1	23.2	73.8	52.5	68.4
MedScissor	67.2	76.2	76.0	75.4	47.9	30.6	23.4	75.7	53.8	70.8

Table 12: **Additional accuracy evaluation metrics of baselines on VQA-RAD at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. Best results for each ratio are highlighted in **bold black font**.

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
Lingshu-7B	26.0	74.9	39.2	39.0	21.5	17.1	13.6	39.0	24.3	53.2
Visual token compression ratio (↓ 50%)										
FastV [ECCV24]	24.3	71.5	36.3	37.1	19.6	15.2	12.5	36.0	21.8	50.6
SparseVLM [ICML25]	25.8	74.3	39.4	38.4	**21.0**	**15.8**	12.7	**39.4**	**23.9**	52.8
ToMe [ICLR23]	25.3	73.5	38.8	38.6	20.8	16.2	13.0	38.6	23.6	52.1
LLaVA-PruMerge [ICCV25]	24.8	75.1	36.5	36.1	17.1	13.5	11.1	36.5	19.8	52.8
DART [EMNLP25]	22.3	74.3	36.2	36.0	19.0	14.6	12.1	36.0	21.6	51.2
DivPrune [CVPR25]	23.8	75.5	37.5	37.0	18.8	14.7	12.1	37.3	21.8	52.6
MedPruner [MICCAI26]	**26.8**	73.5	39.3	38.5	20.2	15.5	12.5	39.3	23.1	52.8
CARVE [Arxiv26]	25.8	74.7	**39.5**	**38.8**	20.2	15.3	12.5	39.3	23.6	53.0
MedScissor	26.0	**76.9**	38.5	37.7	20.4	15.8	**12.8**	38.5	23.3	**54.3**
Visual token compression ratio (↓ 70%)										
FastV [ECCV24]	19.3	69.5	32.8	33.1	17.8	14.0	11.8	32.5	19.9	47.2
SparseVLM [ICML25]	25.3	75.1	36.7	36.7	20.3	15.7	12.6	36.7	22.0	53.0
ToMe [ICLR23]	25.8	73.9	38.9	38.1	20.4	**15.9**	**12.8**	38.9	23.3	52.6
LLaVA-PruMerge [ICCV25]	**26.3**	75.5	37.4	37.3	18.7	14.2	11.5	37.4	21.1	**53.7**
DART [EMNLP25]	20.8	**76.3**	32.3	33.1	17.1	13.7	11.6	32.2	18.6	51.7
DivPrune [CVPR25]	21.8	75.5	35.1	34.8	18.8	14.5	12.0	35.1	21.1	51.7
MedPruner [MICCAI26]	24.8	72.3	**39.2**	**38.6**	**20.6**	15.5	12.7	**39.2**	23.6	51.2
CARVE [Arxiv26]	23.3	69.9	37.4	37.1	19.3	14.7	12.2	37.2	21.8	49.2
MedScissor	25.0	76.1	**39.9**	38.6	20.3	15.5	12.7	**39.9**	**23.9**	53.4
Visual token compression ratio (↓ 90%)										
FastV [ECCV24]	15.3	64.3	26.7	28.3	15.6	12.9	10.9	26.7	15.4	42.6
SparseVLM [ICML25]	19.3	73.5	30.6	31.0	16.7	13.3	10.9	30.6	17.2	49.5
ToMe [ICLR23]	24.3	71.9	35.0	36.1	19.3	14.9	12.3	35.0	20.5	50.8
LLaVA-PruMerge [ICCV25]	24.3	71.9	34.4	34.7	18.3	14.4	11.5	34.4	19.8	50.8
DART [EMNLP25]	22.3	74.3	31.8	32.8	17.3	13.6	11.4	31.8	18.2	51.2
DivPrune [CVPR25]	21.8	70.7	31.0	32.6	18.3	14.5	11.9	31.0	18.6	49.0

Continued on next page

Table 12 – continued from previous page

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
MedPruner [MICCAI26]	22.3	66.7	33.7	34.2	17.8	14.3	11.9	33.7	19.3	47.0
CARVE [Arxiv26]	21.3	68.7	31.8	33.2	16.9	13.9	11.7	31.7	18.0	47.7
MedScissor	**26.0**	72.9	**38.0**	**38.4**	**20.0**	**15.1**	**12.5**	**37.8**	**22.0**	**52.1**

Table 13: **Additional accuracy evaluation metrics of baselines on MMMU-Med at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. Best results for each ratio are highlighted in **bold black font**.

Method	Open	Close	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	All
Lingshu-7B	42.5	55.1	30.8	35.8	15.6	12.5	11.4	30.8	16.0	54.8
Visual token compression ratio (↓ 50%)										
FastV [ECCV24]	34.9	50.0	23.3	28.7	13.6	11.4	10.5	23.3	12.2	49.7
SparseVLM [ICML25]	27.4	51.9	22.4	28.9	13.6	11.3	10.4	22.4	11.2	51.3
ToMe [ICLR23]	34.9	51.3	25.8	30.9	14.2	11.6	10.7	25.8	13.5	51.0
LLaVA-PruMerge [ICCV25]	27.4	49.5	19.1	25.2	12.5	10.6	9.9	19.1	9.8	49.0
DART [EMNLP25]	**39.9**	51.6	**29.1**	**34.7**	**15.2**	**12.3**	**11.2**	**29.1**	**14.8**	51.3
DivPrune [CVPR25]	37.4	52.7	26.6	32.4	14.7	12.0	11.0	26.6	13.6	52.3
MedPruner [MICCAI26]	37.4	52.8	26.6	32.1	14.4	11.7	10.7	26.6	13.5	**52.5**
CARVE [Arxiv26]	29.9	51.4	21.6	27.6	13.3	11.1	10.3	21.6	11.0	50.9
MedScissor	32.5	**52.9**	21.7	27.8	13.4	11.2	10.4	21.7	11.1	52.4
Visual token compression ratio (↓ 70%)										
FastV [ECCV24]	24.9	47.1	15.8	21.9	12.0	10.5	9.9	15.8	8.5	46.6
SparseVLM [ICML25]	24.9	**51.0**	19.9	26.4	12.8	10.8	10.0	19.9	9.9	50.4
ToMe [ICLR23]	**34.9**	49.8	20.8	26.4	13.1	11.1	10.3	20.8	11.0	49.5
LLaVA-PruMerge [ICCV25]	19.9	48.5	17.4	24.2	12.2	10.5	9.8	17.4	8.7	47.8
DART [EMNLP25]	29.9	49.6	21.6	27.9	13.6	**11.4**	**10.6**	21.6	11.1	49.1
DivPrune [CVPR25]	32.4	50.5	19.1	25.4	12.8	10.9	10.2	19.1	9.8	50.1
MedPruner [MICCAI26]	32.4	50.8	21.6	27.6	13.3	11.1	10.3	21.6	11.0	50.4
CARVE [Arxiv26]	29.9	50.5	**24.1**	**29.9**	**13.8**	11.4	10.5	**24.1**	**12.3**	50.1
MedScissor	30.0	50.9	19.2	25.8	13.1	11.2	10.5	19.2	9.9	**50.5**
Visual token compression ratio (↓ 90%)										
FastV [ECCV24]	24.9	43.6	19.1	**25.2**	12.5	10.6	9.9	19.1	9.8	43.2
SparseVLM [ICML25]	22.4	**48.1**	17.4	24.4	12.5	10.7	10.1	17.4	8.7	**47.5**
ToMe [ICLR23]	27.4	47.9	16.6	22.7	11.7	10.1	9.5	16.6	8.6	47.4
LLaVA-PruMerge [ICCV25]	12.4	47.2	9.9	17.4	10.6	9.6	9.2	9.9	4.9	46.4
DART [EMNLP25]	19.9	48.1	14.1	21.2	12.0	10.6	**10.0**	14.1	7.3	47.4
DivPrune [CVPR25]	19.9	47.7	12.4	19.7	11.1	9.9	9.4	12.4	6.2	47.1
MedPruner [MICCAI26]	22.4	47.4	17.4	24.4	12.4	10.7	10.0	17.4	8.6	46.9
CARVE [Arxiv26]	14.9	47.4	9.9	17.4	10.6	9.5	9.2	9.9	4.9	46.7
MedScissor	**27.5**	47.9	**19.1**	25.0	**12.7**	**10.9**	**10.2**	**19.1**	**10.5**	47.4

Table 14: **Additional accuracy evaluation metrics of baselines on 3D-RAD at different compression ratios.** Blue denotes general compression methods, Red denotes medical compression methods. Best results for each ratio are highlighted in **bold black font**.

Method	Open							Close				All
	T1 B-1	T1 R-1	T2 B-1	T2 R-1	T3 B-1	T3 R-1	All	T4	T5	T6	All	
Hulu-Med-7B	32.1	45.9	25.5	38.6	16.6	19.9	25.0	81.0	62.5	79.6	79.0	71.5
Visual token compression ratio (↓ 50%)												
FastV [ECCV24]	31.4	46.4	24.6	38.2	14.9	18.5	24.0	80.4	60.4	78.8	78.3	70.8
SparseVLM [ICML25]	31.4	**46.5**	24.6	38.2	14.9	18.5	24.0	80.5	60.5	**78.9**	78.4	70.9
ToMe [ICLR23]	**31.6**	**46.5**	25.1	38.4	13.6	17.0	**24.1**	80.5	60.8	**78.9**	78.4	70.9
LLaVA-PruMerge [ICCV25]	31.1	46.4	24.6	38.2	**15.4**	**19.0**	24.0	80.5	**61.2**	**78.9**	78.4	70.9
DART [EMNLP25]	30.7	46.1	24.7	38.2	15.0	18.7	23.9	80.5	60.4	78.7	78.4	70.8
DivPrune [CVPR25]	30.8	44.5	24.8	37.9	13.2	16.8	23.6	80.1	60.8	**78.9**	78.1	70.5
MedPruner [MICCAI26]	31.2	45.1	**25.3**	38.4	13.3	16.8	24.0	80.8	60.9	78.7	78.6	**71.1**
CARVE [Arxiv26]	31.2	45.1	25.2	38.4	13.3	16.8	24.0	80.8	60.9	78.6	78.6	71.0

Continued on next page

Table 14 – continued from previous page

Method	Open							Close				All
	T1 B-1	T1 R-1	T2 B-1	T2 R-1	T3 B-1	T3 R-1	All	T4	T5	T6	All	
MedScissor	31.3	45.2	**25.3**	**38.5**	13.4	16.9	**24.1**	**80.9**	61.0	78.8	**78.7**	**71.1**
Visual token compression ratio (↓ 70%)												
FastV [ECCV24]	27.9	47.2	24.3	38.2	14.7	18.5	23.1	80.4	58.8	79.1	78.2	70.5
SparseVLM [ICML25]	27.9	47.2	24.4	38.2	14.7	18.5	23.1	80.5	58.9	79.1	78.2	70.6
ToMe [ICLR23]	30.2	45.9	25.5	38.9	12.5	16.0	23.7	80.1	58.1	78.3	77.7	70.3
LLaVA-PruMerge [ICCV25]	28.9	46.1	25.2	38.9	14.4	18.3	23.7	80.6	**59.4**	**79.2**	78.3	70.8
DART [EMNLP25]	29.0	46.7	23.9	38.0	**15.1**	**18.6**	23.1	80.6	59.1	79.0	78.3	70.7
DivPrune [CVPR25]	30.5	44.2	24.5	37.5	12.7	16.3	23.3	79.5	58.1	78.4	77.3	69.8
MedPruner [MICCAI26]	**31.9**	**47.6**	**25.8**	**39.1**	13.5	16.8	**24.5**	**80.8**	58.9	78.6	**78.4**	**71.0**
CARVE [Arxiv26]	**31.9**	47.5	**25.8**	39.0	13.4	16.7	24.4	80.7	58.8	78.5	78.3	70.9
MedScissor	**31.9**	47.5	**25.8**	**39.1**	13.4	16.8	**24.5**	80.7	58.9	78.5	**78.4**	70.9
Visual token compression ratio (↓ 90%)												
FastV [ECCV24]	22.5	42.1	23.0	37.8	13.0	16.8	20.7	**81.0**	55.5	79.0	78.3	70.3
SparseVLM [ICML25]	22.5	42.1	23.0	37.8	13.0	16.8	20.8	**81.0**	55.6	79.1	78.3	70.4
ToMe [ICLR23]	28.4	46.4	**25.2**	38.5	12.4	15.9	23.2	80.4	54.2	79.0	77.7	70.1
LLaVA-PruMerge [ICCV25]	26.5	44.2	24.4	38.3	**13.4**	**17.3**	22.5	80.9	**56.0**	**79.5**	78.3	70.6
DART [EMNLP25]	24.3	43.1	23.1	37.8	12.2	16.1	21.0	**81.0**	**56.0**	79.4	**78.4**	70.4
DivPrune [CVPR25]	27.2	45.9	24.5	38.0	11.4	15.4	22.3	78.7	53.3	78.3	76.1	68.7
MedPruner [MICCAI26]	**30.8**	**48.0**	25.0	**38.9**	12.6	16.4	**23.6**	80.9	55.0	78.9	78.2	70.6
CARVE [Arxiv26]	**30.8**	**48.0**	25.0	**38.9**	12.6	16.5	**23.6**	**81.0**	55.0	79.0	78.2	**70.7**
MedScissor	30.7	**48.0**	24.9	38.8	12.5	16.4	23.5	80.9	54.9	78.9	78.1	70.6

Tables 10–14 provide a more detailed additional accuracy evaluation to further assess model performance under different question types. For the four 2D medical imaging benchmarks containing open-ended questions, we further evaluate open-ended performance. Since these questions require free-form textual responses, we follow the official evaluation protocols of each benchmark for the primary open-ended metric and additionally report F1, BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR to provide a more comprehensive evaluation of response quality. For the 3D medical imaging benchmarks, only 3D-RAD contains open-ended questions, where both open-ended and closed-ended questions are further divided into three tasks. For open-ended questions, we use BLEU-1 as the primary score and additionally report BLEU-1 and ROUGE-1 for each task, while the accuracies of the three tasks are reported separately for closed-ended questions. All accuracy and evaluation metrics are computed following the official settings (Xu et al., 2026; Jiang et al., 2025).

E Case Study

To further analyze the token retention behavior of different methods in medical visual question answering, we conduct a qualitative comparison on a representative case. As shown in Figure 7, the question is *"What do the arrows indicate in the confocal microscope image?*, where the correct answer is *D (attachment of H. pylori to cell membrane)*, while the other options lead to incorrect predictions. FastV, SparseVLM, and MedPruner tend to remove critical visual tokens related to the lesion regions and question-relevant areas under high compression ratios, resulting in incorrect answers. DART and DivPrune preserve more visual information but still retain some redundant regions. In contrast, MedScissor consistently preserves informative regions related to the question across different compression ratios while reducing irrelevant background tokens, enabling correct prediction. This case further demonstrates the effectiveness of MedScissor in leveraging local structural variation and global visual deviation for precise visual token compression.

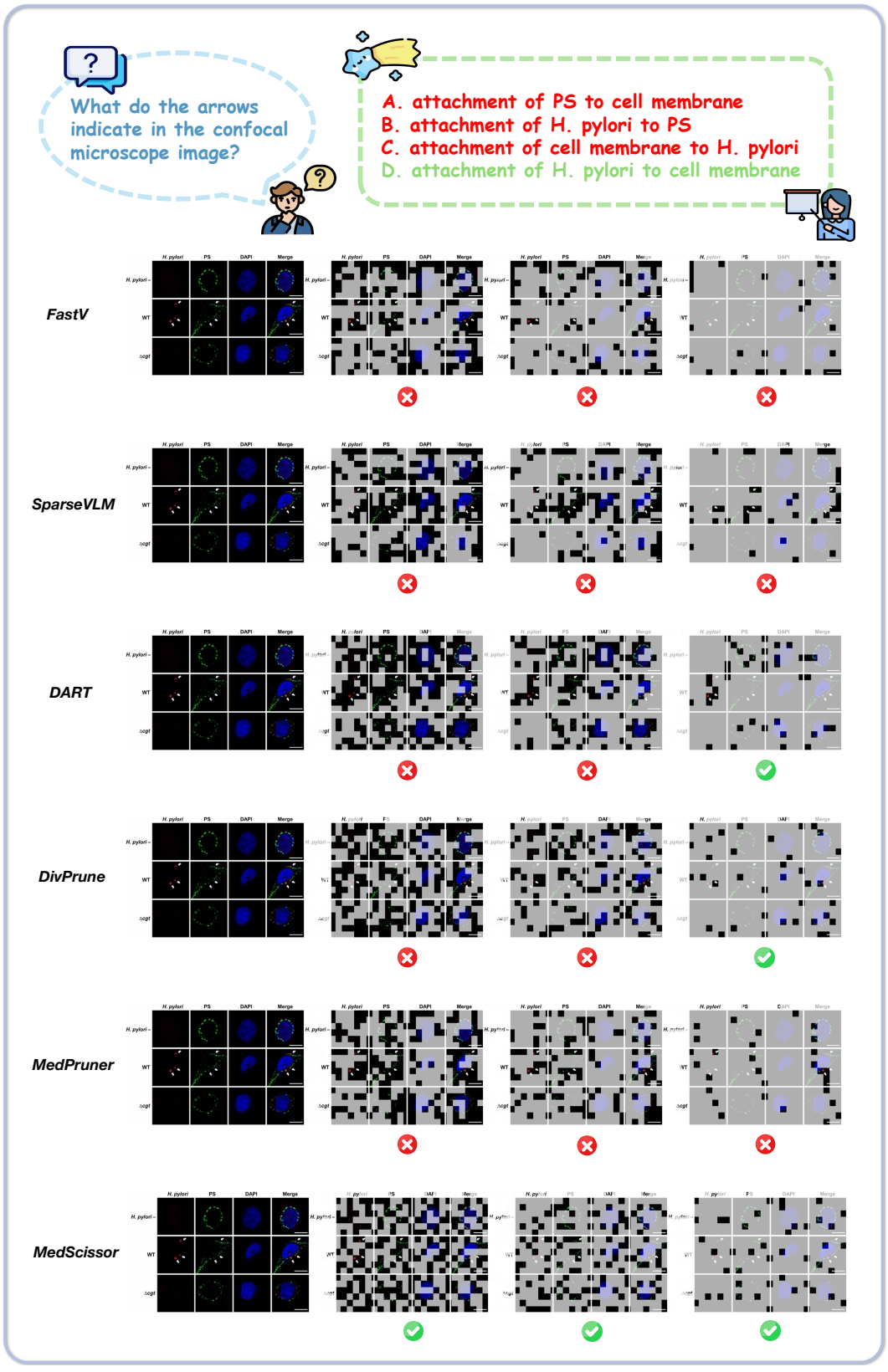

Figure 7: **Token retention visualization of different baselines under various compression ratios.**