ToSA: Token Merging with Spatial Awareness

Abstract— **Token merging has emerged as an effective strategy to accelerate Vision Transformers (ViT) by reducing computational costs. However, existing methods primarily rely on the visual token's feature similarity for token merging, overlooking the potential of integrating spatial information, which can serve as a reliable criterion for token merging in the early layers of ViT, where the visual tokens only possess weak visual information. In this paper, we propose ToSA, a novel token merging method that combines both semantic and spatial awareness to guide the token merging process. ToSA leverages the depth image as input to generate pseudo spatial tokens, which serve as auxiliary spatial information for the visual token merging process. With the introduced spatial awareness, ToSA achieves a more informed merging strategy that better preserves critical scene structure. Experimental results demonstrate that ToSA outperforms previous token merging methods across multiple benchmarks on image question answering and spatial understanding while largely reducing the runtime of the ViT, making it an efficient solution for ViT acceleration.**

I. Introduction

Most of the recent vision foundation models [1], [2], [3] adopt Vision Transformer (ViT) [4] as the backbone, achieving advanced performance on various perception tasks such as classification, detection, and segmentation. On the other hand, these vision foundation models also play an important role in the field of generative AI, especially serving as the visual encoder of Vision Language Model (VLM) [5], [6]. Although achieving great success, the attention mechanism of the ViT introduces heavy computational overhead and limits its further applications in real-world scenarios such as robotics, and autonomous driving, where high throughput and low computational cost are preferred.

Many previous works [7], [8], [9], [10] have explored more efficient ViT architecture to accelerate the runtime of ViT training and inference. These new architectures introduced extra learnable parameters or pooling layers that can reduce the number of visual tokens, therefore lowering the computational cost and runtime. Despite all these efforts, these methods require extra training due to their newly introduced model parameters, which limited their practicability compared with recent plug-and-play token reduction methods.

Recently, training-free, plug-and-play token reduction methods have been introduced to improve the efficiency of ViT. Token pooling [11] demonstrates effectiveness in reducing computational cost and the number of visual tokens in the ViT by using clustering method. Recent work such as ToMe [12] serves as a popular training-free method that largely reduces the computational cost of ViT, by introducing Bipartite Soft Matching (BSM), ToMe merges the visual token progressively in each layer of ViT based on the visual token features' similarity, achieving superior performance on

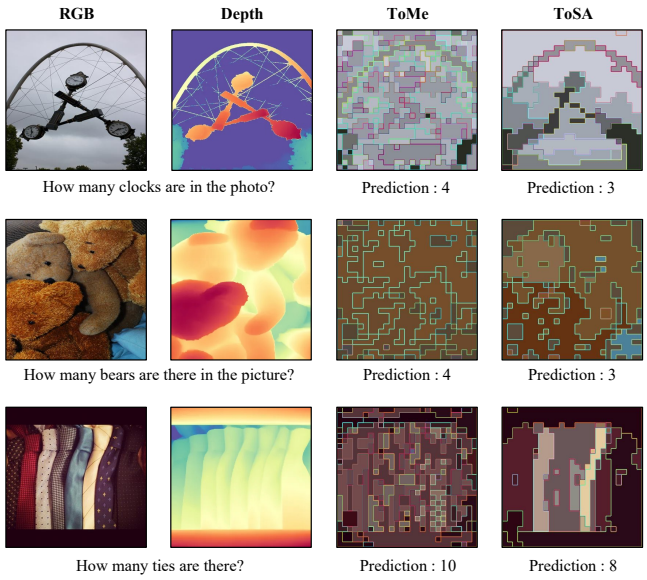

Fig. 1. A merging comparison between ToMe and ToSA. By leveraging depth input, ToSA utilizes spatial awareness in the token merging process and leads to more spatially coherent merging results, helping the models to answer the question correctly. Both merging results retain 16 visual tokens for better visual comparison. Merged token is denoted by the same patch and inner edge color.

image classification tasks compared with previous efficient ViT designs that require further training. Although these methods serve as an effective way to balance the trade-off between ViT efficiency and performance, they heavily rely on the visual features for the token merging and clustering process. Previous exploration work demonstrates [13] that the visual features in the early layer mostly capture low-level information such as edges and textures. For this reason, we argue the low-level features in the early layers are sub-optimal criteria for token merging, as the visual tokens from different objects can also demonstrate similarity in low-level visual features. In tasks such as embodied question answering or spatial question answering, preserving local information is also important, especially when facing questions related to counting, object existence...etc. For this reason, a token merging method that can preserve important local information for more fine-grained question answering task, and also enhance the ViT efficiency in VLM is needed.

In this work, we present **To**ken merging with **S**patial **A**wareness (ToSA), aim to explore the potential of using the depth map from the RGB-D input as an auxiliary spatial awareness for training-free token merging. We compared ToSA with semantic-based token merging method ToMe [12]

on various image and video question answering benchmarks including SpatialBench [14], VQAv2-Counting [15], GQA [16], and OpenEQA [17]. Experiment results show that state-of-the-art VLM incorporated with ToSA can demonstrate better performance compared with the existing token merging method. Furthermore, by leveraging the auxiliary spatial information, ToSA can generate more spatial coherent token merging results compared with previous work, as shown in Fig. 1, which further improves the question answering accuracy. The contributions of this work are summarized as follows:

- We present **ToSA**, a training-free token merging method that conducted token merging based on semantic similarity and spatial awareness. We conduct extensive experiments on multiple images and video question answering benchmarks, demonstrating ToSA outperforms the previous semantic-based method ToMe with minimal additional computational cost, showing its potential for real-world applications.
- We present detailed qualitative results and merging comparison between previous work ToMe and ToSA, demonstrating ToSA can generate more spatial coherent merging results.

II. Related Work

A. Efficient Vision Transformer

To enhance the efficiency of ViT, most of the previous works adopt different ViT architectures that can downsample the number of visual tokens or reduce the computation of attention operations in the model. EfficientViT [8] introduced a cascaded group attention module and a parameter reallocation strategy to reduce the computational cost. DynamicViT [9] and AdaViT [10] proposed additional learnable modules to reduce the number of visual tokens. SP-ViT [18] proposed spatial prior-enhanced attention to reduce the computational cost of self-attention blocks. However, these methods possess several limitations. Firstly, most of these methods lead to a dynamic number of visual tokens, which can not be applied to batch training and inference. Moreover, these models are optimized for image classification tasks, and their practicability for visual question answering task are not fully justified yet. Lastly, these methods' additional training parameters make them challenging to be directly adopted by recent VLMs [5], [6], especially when most of the available model checkpoints are only trained on classification tasks. These reasons lead to the current VLMs still lean to leverage CLIP [2] or SigLIP [3] as visual encoder despite their adoption of original ViT architecture.

B. Training-free Token Reduction

In contrast to previous training-required methods, to enhance the practicability, several recent works do not require extra training and serve as a plug-and-play module to reduce the number of visual tokens and can enhance ViT and VLM efficiency. Work such as Evo-ViT [19] or EViT [20] utilized attention score to conduct token merging on less attentive tokens. ToMe [12] proposed bipartite soft matching and merged the visual tokens that share similar semantic visual features between each layer of ViT. These methods rely heavily on attention score or visual feature similarity as a criterion for token merging. However, as indicated by previous studies [13], [21], early layers of ViT features tend to demonstrate low-level information as well as similar attention scores across visual tokens within the same layer. This nature of ViT can potentially hinder the early-stage token merging process of the current methods, given the low-level similarity between two tokens does not guarantee they are from the same object and can be confusing for semantic-based merging methods. In addition to ViT-based token reduction methods, recent methods like Fast-V [22] and SparseVLM [23] have introduced LLM-based visual token reduction conditioned on attention scores or text input. However, these approaches have several limitations. First, they offer weaker acceleration compared to ViT-based token reduction methods, as their token pruning occurs at a later stage in the VLM pipeline. Additionally, unlike ViT token reduction, LLM token reduction methods require storing all visual tokens across multiple conversation rounds, restricting their practicality in real-world applications. In this work, we focus on ViT-based token reduction methods for better practicality, aiming to provide more efficient acceleration while maintaining high performance on VLM's visual question answering tasks. Furthermore, instead of fully leveraging semantic affinity, we proposed to leverage spatial awareness as another criterion in the token merging process, aiming to improve the merging robustness of ViT in the earlier layers.

C. Spatial Understanding

Many recent works focus on advancing VLMs' ability to spatial understanding question answering tasks. SpatialVLM [24] can conduct spatial understanding question answering from RGB input images after training on the collected large-scale spatial question answering data. SpatialRGPT [25] and SpatialBot [14] incorporate extra depth modality as input to the visual encoder for spatial question answering. To evaluate the performance of state-of-the-art VLM on spatial understanding, SpatialBot further proposed SpatialBench, an image-based spatial understanding benchmark, featuring spatial understanding questions with multiple question categories including size comparison, counting, enumeration, spatial relationship...etc. On the other hand, VQAv2 [15] includes object counting questions that can evaluate VLMs' capability in object counting, and GQA [16] focuses on more general image question answering tasks. OpenEQA [17] is another benchmark that focuses on egocentric video-based 3D scene understanding, featuring RGB-D input with seven question categories, serving as a comprehensive benchmark for evaluating VLMs' spatial understanding capability in the 3D scene. In this work, we evaluated VLM's performance after incorporating our proposed ToSA on these spatial understanding benchmarks.

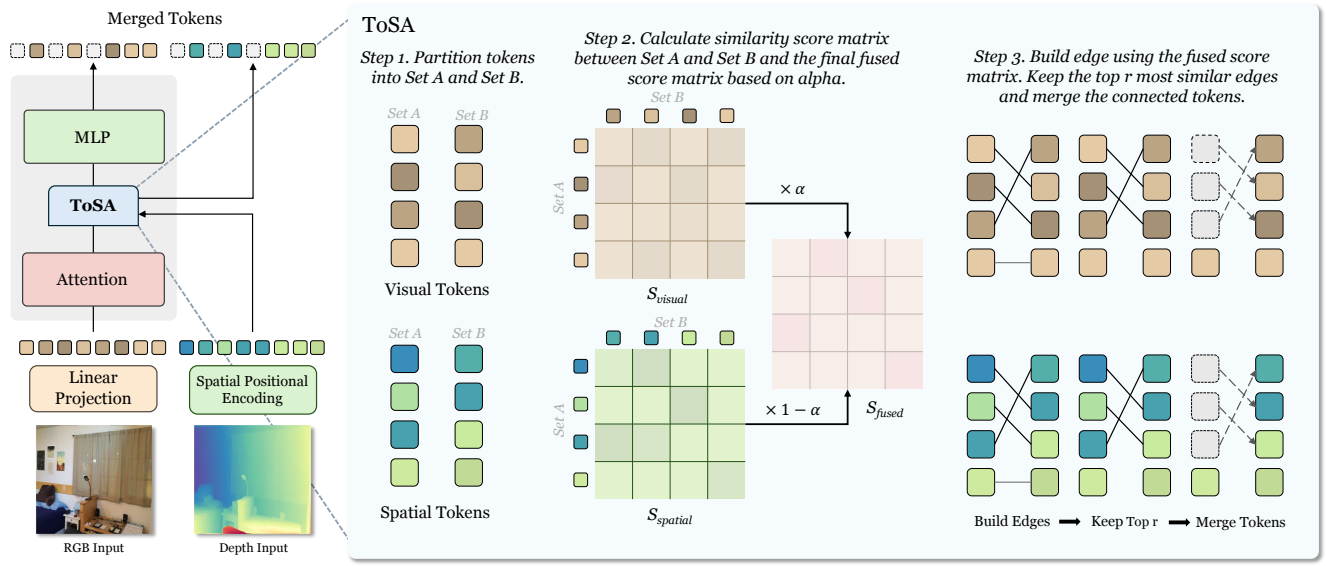

Fig. 2. The overall framework of ToSA. ToSA block is inserted between attention and MLP across each encoder layer in ViT. ToSA block takes visual tokens and spatial tokens as input to conduct token merging. The merging process is based on the similarity of both visual tokens and spatial tokens.

III. Method

A. Preliminary

Bipartite soft matching is an efficient way to merge the tokens within each ViT layer used by the previous token merging method [12]. It is applied on the visual tokens between the attention and MLP block as:

1) Alternatively partition the tokens into two sets $\mathbb{A}$ and $\mathbb{B}$ of roughly equal size.
2) For each token in set $\mathbb{A}$, calculate the token similarity with each token in set $\mathbb{B}$ based on cosine similarity of the *Key* features in the attention block. The similarity can be represented by a score matrix $\mathbf{S} \in \mathbb{R}^{|\mathbb{A}|\times|\mathbb{B}|}$, where S_{ij} represents the cosine similarity between the i^{th} token in $\mathbb{A}$ and the j^{th} token in $\mathbb{B}$.
3) Merge the most similar r pairs using a weighted average, and record the token size.
4) Concatenate the two sets $\mathbb{A}$ and $\mathbb{B}$ back together again.

Once the tokens have been merged, they carry the features of more than one visual token. Therefore, the merged tokens will have less effect on softmax attention. We apply the proportional attention as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\mathbf{d}}} + \log \mathbf{s}\right) \tag{1}$$

where $\mathbf{s}$ is the number of patches the token represents after token merging.

B. Spatial Token

Given the visual token's feature is not reliable enough for merging in the early stage of ViT, we introduce spatial tokens, which are a series of tokens that are generated from the depth input. Spatial tokens contain rich spatial information and will go through the ToSA block along with the original visual tokens, serving as another criterion to calculate the score matrix for bipartite soft matching.

The way we generate spatial tokens is illustrated in Fig. 2. For each patch on the patchify depth image, we generate a triplet of index (x, y, z), which stands for the index on the image. E.g. the most top left patch will have $x = 0$ and $y = 0$, the x and y in the image ranging from the number of patches in x and y dimension. Here x and y range from 0 to 26, which is based on the resolution and patch size of our visual encoder. The third index z is determined by the relative depth of the patch in the image, we average all the pixels' relative depth to obtain patch-wise relative depth. We divide the estimated depth value into 27 levels to match the index range in the x and y direction.

After the index assignment, each patch's spatial information will be represented by a triplet of index (x, y, z), which we further encode with positional encoding following transformer [26], each patch will turn into a spatial token that represents the patch spatial location, providing critical spatial prior for the ToSA token merging process.

C. ToSA (Token Merging with Spatial Awareness)

In the ViT forward pass, ToSA is inserted between the attention and MLP in the ViT layers to reduce the number of visual tokens. ToSA takes the visual tokens and spatial tokens as input, and calculates two separate score matrices S_{visual} and $S_{spatial}$ based on the similarity of visual tokens and spatial tokens, respectively. Next, ToSA calculates the final fused score matrix based on the following equation, where α is a variable controlling the score matrix that focuses more on visual or spatial information.

$$S_{fused} = \alpha S_{visual} + (1 - \alpha) S_{spatial} \tag{2}$$

The fused score matrix S_{fused} will be used by the bipartite soft matching, and the visual tokens and spatial tokens will

TABLE I

CATEGORY-LEVEL PERFORMANCE ON SPATIALBENCH.

Models	Position (↑)	Existence (↑)	Counting (↑)	Reaching (↑)	Size (↑)
Base Model					
SpatialBot-3B	58.8	80.0	86.7	53.3	25.0
Token Merging - retain 10% tokens					
ToMe	52.9	55.0	58.7	58.3	25.0
ToSA	52.9 (+0.0)	65.0 (+10.0)	72.7 (+14.0)	63.3 (+5.0)	21.7 (-3.3)

TABLE II

PERFORMANCE COMPARISON ON THE VQAV2 - COUNTING.

Models	RCA	Acc. (%)
Base Model		
LLaVA-OV-7B	77.1	100
Retain 50% tokens		
ToMe	53.1	68.9
ToSA	64.1	83.1
Retain 10% tokens		
ToMe	54.1	70.2
ToSA	59.9	77.7

TABLE III

PERFORMANCE COMPARISON ON THE GQA DATASET.

Models	EM@1
BLIP-2 [27]	41.0
InstructBLIP [28]	49.2
Qwen-VL-Chat [29]	57.5
LLaVA-1.5 [30]	62.0
LLaMA-VID [31]	62.3
VILA [32]	62.3
Base Model	
LLaVA-OV-7B	66.0
Retain 50% tokens	
ToMe	62.7
ToSA	63.2
Retain 10% tokens	
ToMe	57.4
ToSA	57.8

be merged based on both visual and spatial information using the fused score matrix S_{fused}. Note that the visual tokens and spatial tokens will be merged in a corresponding style, i.e. if the i^{th} and j^{th} visual tokens are merged, the i^{th} and j^{th} spatial tokens will also be merged accordingly.

The parameter α is a weighting factor that controls whether ToSA relies more on visual feature similarity or spatial affinity. In the early layers, given the visual features are less reliable, we use a smaller α to encourage ToSA to conduct token merging based on spatial affinity. In the deeper layers, we use a larger α so that ToSA can utilize the more semantic features in the deep layers of ViT. This merging strategy seamlessly takes advantage of both semantic-based merging and the auxiliary spatial prior provided by the RGB-D input. In the default setting, we use a linear increase schedule for α, which gradually increases when passing through the layers of ViT. The value of α at layer i^{th} can be denoted as:

$$\alpha_i = i/L \tag{3}$$

with α_i represents the α value in the i^{th} layer and L is the total number of layers in ViT. We also conducted different schedules of α values and their effect on the performance of spatial understanding, see Table. V for more results.

IV. EXPERIMENTS

A. Implementation

We use SpatialBot-Phi2-3B [14] in our experiments on SpatialBench. On other benchmarks, we use LLaVA-OneVision-7B [6] as our base model. Both VLMs use siglip-so400m-patch14-384 [3] as the visual encoder, which has 27 ViT layers, and L in Eq. 3 is set to 27 accordingly. SpatialBot-Phi2-3B is trained on the SpatialQA [14], and LLaVA-OneVision-7B is trained on single, multi-image, and video data collected from a wide range of publicly available instruction following dataset. For embodied question answering benchmarks, we uniformly sampled 12 images from the 3D scans as input to our VLM. The depth image used to generate the spatial tokens is predicted by depth-anything-v2 [33]. All experiments are conducted on a V100 GPU.

B. Benchmarks

There are several publicly available spatial understanding question answering benchmarks, including SpatialRGPT-Bench [25], and SpatialBench [14]. SpatialRGPT-Bench focuses on dense region spatial question answering, which requires extra region boxes or masks as input to the ViT and can not be seamlessly adopted by most of the current VLMs. Therefore, we used SpatialBench, which features 5 different question types related to object position, existence, counting, reaching, and size. Besides SpatialBench, we also evaluated ToSA on counting questions collected from VQAv2 [15]

and GQA [16], which focuses on more general visual reasoning questions. Besides image-level question answering, we further evaluated ToSA on embodied question answering tasks, which involve video-level question answering in 3D environments. We conducted our experiments on the OpenEQA [17] dataset, which addresses embodied question answering across seven categories, covering aspects such as object recognition, spatial reasoning, object localization, attribute recognition...etc. Additionally, OpenEQA incorporates an LLM scorer that evaluates the quality of predictions by comparing them with ground truth, offering assessments that better align with human judgment.

C. Performance

Most of the efficient ViT methods [8], [9], [10] required re-training the ViT for different tasks beyond image classification, and there are very limited VLM integrations that can enable us to conduct spatial question answering evaluation using these models. For this reason, we mainly compared our proposed ToSA with another recent training-free, off-the-shelf ViT token merging method ToMe [12]. ToMe has been used by many VLM [34], [35] for efficient training and inference, which demonstrates its successful integration for image and video understanding tasks.

SpatialBench. In Table. I, we listed the category-level performance of base model SpatialBot-3B [14] and its performance after applying different token merging methods. We demonstrate that ToSA can achieve comparable performance with ToMe in most of the tasks, and largely outperforms ToMe on task that requires more fine-grained spatial understanding such as object existence and counting. The question consists of multi-choice questions and counting, where the former is evaluated based on accuracy. The latter is calculated with Relative Counting Accuracy, which is calculated as $1 - \frac{|x-y|}{y}$, where x is the predicted count and y is the ground truth.

VQAv2-Counting. In Table II, we present the performance on VQAv2's counting questions. We adopt SpatialBench's Relative Counting Accuracy (RCA), and also report the accuracy percentage with respect to the base model.

GQA. In Table. III, we listed the performance of different VLMs and the comparison between ToSA and ToMe. Despite GQA focusing more on general question answering, instead of heavily related to spatial question answering, ToSA achieves better performance compared with ToMe across different visual token usage.

OpenEQA. We show the experiment results on OpenEQA in Table. IV, including the performance of proprietary VLMs, and several open-source VLM's performance with some of them [36], [37], [38] integrate extra learnable modules that can conduct token reduction for video understanding task.

TABLE IV
PERFORMANCE COMPARISON ON THE OPENEQA DATASET.

Models	LLM-Match
Proprietary VLMs	
Claude-3 Opus	36.3
Gemini 1.0 Pro Vision	44.9
Claude-3.5 Sonnet	48.7
GPT4-V (15 frames)	54.6
GPT4-V (50 frames)	55.3
Open-source VLM	
Video-LLaMA [37]	20.0
LLaMA-2 w/ Concept Graph [17]	28.7
AuroraCap [34]	28.9
Video-ChatGPT [36]	32.1
LLaMA-2 w/ Sparse Voxel Map [17]	34.3
LLaMA-2 w/ LLaVA-1.5 caption [17]	36.8
Chat-UniVi [39]	42.3
Video-LLaMA2 [38]	49.2
Base Model	
LLaVA-OV-7B	56.2
Retain 10% tokens	
ToMe	48.3
ToSA	49.5

TABLE V
ABLATION ON TOKEN MERGING SCHEDULE.

Schedule	Acc
uniform	56.7
decrease	52.8
increase	**59.9**

TABLE VI
ABLATION ON INFERENCE SPEED.

Methods	Used Token	im/s
SigLIP	100%	18.2
ToMe	50%	23.8
ToSA	50%	23.7
ToMe	10%	33.7
ToSA	10%	33.5

D. Ablation Studies

Effect of Alpha Schedule. To demonstrate that the earlier layer's visual features are suboptimal for token merging, we conduct experiments based on different α schedules in Table V. We tested three different alpha schedules, including a uniform α value across all layers, where α is set to as constant across all layers. We set α to 0.5 for our experiment. We also evaluate the decrease schedule where the α starts with 1 and gradually reduces to 0 in the deeper layers of ViT. As shown in Table V, using the increase schedule for α resulted in the highest accuracy on VQAv2-counting, and outperforms the other two schedules by a large margin. This showcases our assumption that it is beneficial to leverage spatial awareness in earlier layers for token merging.

Inference Speed. We compared the inference speed of ToMe and ToSA in Table VI. Both methods are tested on the V100 GPU in our experiments. Both ToMe and ToSA can achieve a large boost in throughput compared with the original ViT. Compared with ToMe, ToSA incorporates extra

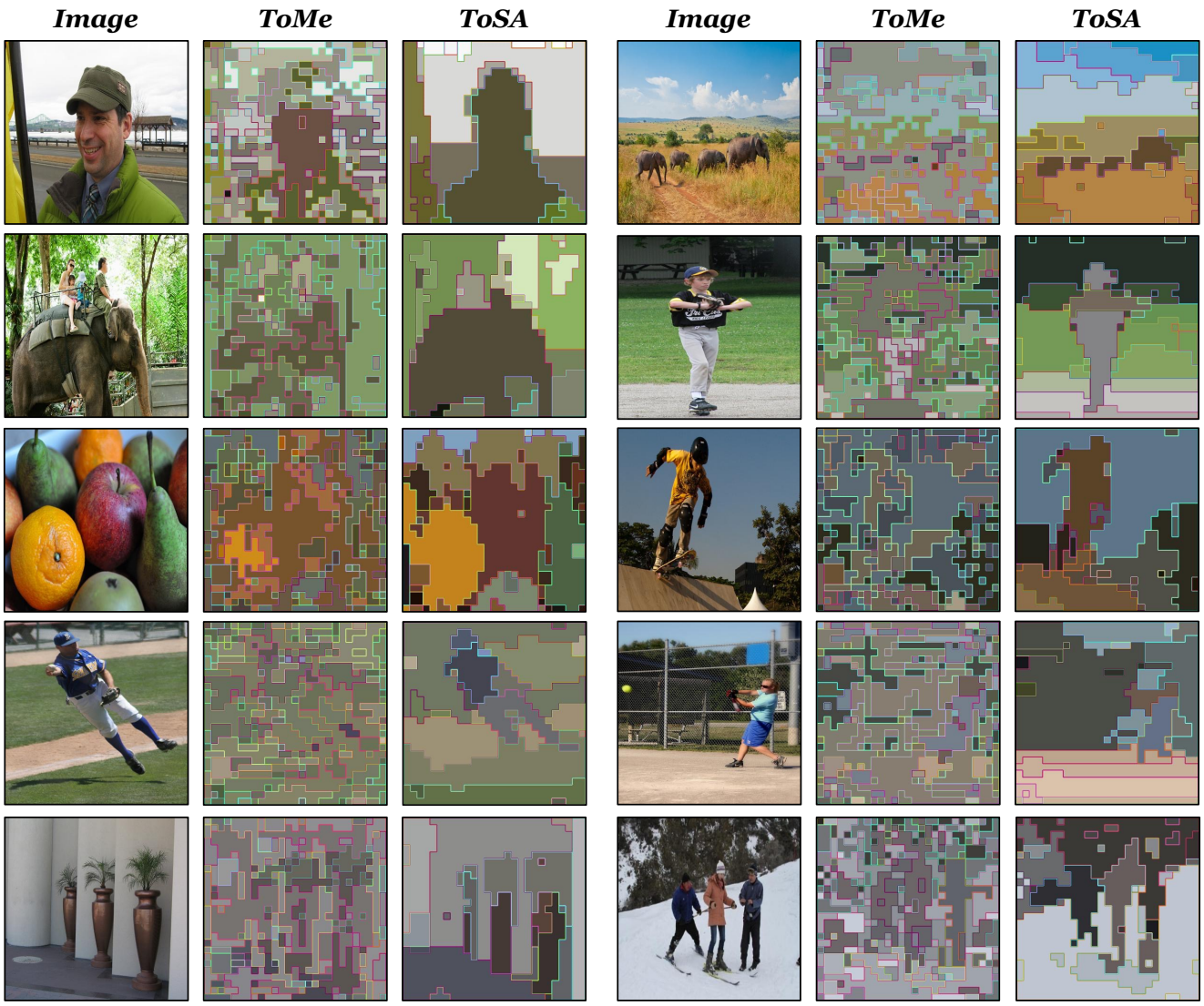

Fig. 3. Merging results comparison between ToMe and ToSA. We only keep 16 visual tokens (2%) from both methods for better visual comparison. ToSA demonstrate more spatially coherent merging results, which leads to higher performance across various visual question answering benchamrks.

spatial tokens during the inference, but ToSA's degradation of throughput is minimal, with less than 0.6% degradation under 50% used token and 10% used token settings. Note that the throughput is different from ToMe's original reported number because our SigLIP uses a higher-resolution image input, which results in more visual tokens.

E. Qualitative Results

We showcase some qualitative results in Fig. 3, which exhibits the difference in merging results between semantic-only ToMe and ToSA's semantic and spatial-aware merging. Because ToSA leverages spatial awareness during the merging process, the qualitative results demonstrate ToSA maintains a more coherent token merging results, with respect to the spatial structure of the image. Here, the same token is denoted by the same color and inner edges. Note that we only keep 16 visual tokens (2%) from both methods for better visual comparison between the two token merging methods.

V. Conclusion

In this work, we proposed ToSA, a training-free token merging method that conducts token merging based on spatial prior. Experimental results show that ToSA can achieve better performance compared with previous token merging methods on multiple VQA benchmarks including SpatialBench, VQAv2, GQA, and OpenEQA. Furthermore, despite the improvements, ToSA introduces minimal additional inference cost, with less than 0.6% of runtime degradation compared with previous work, demonstrating its potential for real-world applications.

VI. Limitation

Although ToSA demonstrates advanced performance compared with the existing method, ToSA requires auxiliary depth image as input during the merging process. This makes ToSA more applicable to scenarios such as robotic or autonomous driving, where RGB-D input data is available.

REFERENCES

[1] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[3] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[4] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[6] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[7] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman *et al.*, "Hiera: A hierarchical vision transformer without the bells-and-whistles," in *International Conference on Machine Learning*. PMLR, 2023, pp. 29 441–29 454.

[8] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 420–14 430.

[9] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13 937–13 949, 2021.

[10] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, "A-vit: Adaptive tokens for efficient vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 809–10 818.

[11] D. Marin, J.-H. R. Chang, A. Ranjan, A. Prabhu, M. Rastegari, and O. Tuzel, "Token pooling in vision transformers for image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 12–21.

[12] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your ViT but faster," in *International Conference on Learning Representations*, 2023.

[13] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, M. Shu, M. Goldblum, A. G. Wilson, and T. Goldstein, "What do vision transformers learn? a visual exploration," *arXiv preprint arXiv:2212.06727*, 2022.

[14] W. Cai, Y. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," *arXiv preprint arXiv:2406.13642*, 2024.

[15] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[16] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[17] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud *et al.*, "Openeqa: Embodied question answering in the era of foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 488–16 498.

[18] Y. Zhou, W. Xiang, C. Li, B. Wang, X. Wei, L. Zhang, M. Keuper, and X. Hua, "Sp-vit: Learning 2d spatial priors for vision transformers," in *The 33rd British Machine Vision Conference*, 2022.

[19] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun, "Evo-vit: Slow-fast token evolution for dynamic vision transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2964–2972.

[20] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie, "Not all patches are what you need: Expediting vision transformers via token reorganizations," in *International Conference on Learning Representations*, 2022.

[21] X. Wu, F. Zeng, X. Wang, and X. Chen, "Ppt: Token pruning and pooling for efficient vision transformers," *arXiv preprint arXiv:2310.01812*, 2023.

[22] L. Chen, H. Zhao, T. Liu, S. Bai, J. Lin, C. Zhou, and B. 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*. Springer, 2025, pp. 19–35.

[23] Y. Zhang, C.-K. Fan, J. Ma, W. Zheng, T. Huang, K. Cheng, D. Gudovskiy, T. Okuno, Y. Nakata, K. Keutzer *et al.*, "Sparsevlm: Visual token sparsification for efficient vision-language model inference," *arXiv preprint arXiv:2410.04417*, 2024.

[24] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 455–14 465.

[25] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision-language models," in *NeurIPS*, 2024.

[26] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[27] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[28] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[29] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[30] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[31] Y. Li, C. Wang, and J. Jia, "Llama-vid: An image is worth 2 tokens in large language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 323–340.

[32] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 689–26 699.

[33] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[34] W. Chai, E. Song, Y. Du, C. Meng, V. Madhavan, O. Bar-Tal, J.-N. Hwang, S. Xie, and C. D. Manning, "Auroracap: Efficient, performant video detailed captioning and a new benchmark," *arXiv preprint arXiv:2410.03051*, 2024.

[35] Y. Weng, M. Han, H. He, X. Chang, and B. Zhuang, "Longvlm: Efficient long video understanding via large language models," in *European Conference on Computer Vision*. Springer, 2025, pp. 453–470.

[36] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[37] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[38] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao *et al.*, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[39] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-univi: Unified visual representation empowers large language models with image and video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 700–13 710.