M2M-TAG: Training-Free Many-to-Many Token Aggregation for Vision Transformer Acceleration

Fanhu Zeng[1, 3], **Deli Yu**[2*]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS
[2]Baidu Inc. [3]School of Artificial Intelligence, UCAS
zengfanhu2022@ia.ac.cn, yudeli@baidu.com

Abstract

Vision transformers have been widely explored due to its unprecedented performance in various downstream tasks. However, its heavy computational cost restricts its real-world deployment and much interest has aroused for compressing tokens of vision transformer dynamically. Current methods mainly pay attention to token pruning or merging to reduce token numbers, which inevitably leads to numerous information loss. In this paper, we regard token reduction process as matrix transformation of tokens, and propose a many-to-many token aggregation framework called M2M-TAG, which can serve as a generalization form of all existing methods. The parameter-free many-to-many transformation can be constructed by combining importance and similarity metric of full tokens in global scope. The aggregated tokens can reserve token information to the most and enable **training-free** acceleration. We employ it as a plug-and-play module to accelerate vision transformers and conduct various experiments to demonstrate the effectiveness of proposed framework. Specifically, we reduce 34.8% FLOPs with only 0.1% accuracy drop on DeiT-S without fine-tuning, even outperforming some existing fine-tuning methods. We further comprehensive results show that the approach achieves competitive performance with better computation-performance trade-off, impressive budget reduction and maximum inference acceleration. Code is avaiable at https://github.com/AuroraZengfh/TokenTransforming.

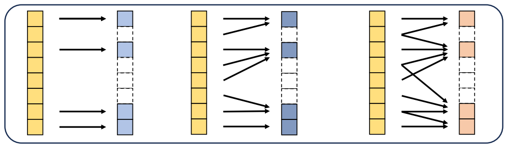
1 Introduction

Research on Vision Transformers (ViTs) [8] has made breakthrough in various downstream CV tasks including image classification [38, 51, 18], object detection [20, 55, 2], semantic segmentation [35, 4] and so on [33, 26, 48, 21, 52, 19, 30]. However, quadratic computation in proportion to the number of tokens significantly prevents wide application. To this end, model compression is proposed to reduce redundant computation inside the model [10, 41, 11, 40, 36, 24, 43, 50, 54, 42].

There are mainly three ways, namely distillation [16, 56, 14], quantization [57, 17, 12] and pruning [28, 47, 45] for model compression in general. In this paper, we focus on pruning the tokens of ViT based models in a dynamic way [31, 25] as it is consistent with common sense of human cognition that both the important attentive region and the neglected uninformative area dynamically vary with the given images. Thus, dynamic image token (patch) compression is beneficial to better accuracy and efficiency trade-off in various tasks [13].

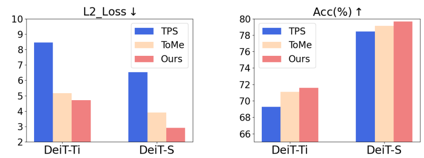
Some works [31, 9, 46, 23] prune uninformative tokens with low importance score directly, which is calculated by trainable prediction module or is based on statistics of attention map in self-attention layers. Considering the information loss during token pruning, others [1, 22, 44, 27] adopt token

*Corresponding author.



(a) Token Pruning (One-to-One) (b) Token Merging (Many-to-One) (c) **Token Aggregation (Many-to-Many)**

Figure 1: Comparison of different token reduction methods with original tokens at the left column and fewer remaining tokens at the right column. (a) represents pruning methods, and (b) represents merging methods. Both of them exclusively reduce original tokens into fewer remaining tokens. (c) represents our method, where each original token can be integrated into remaining tokens in a many-to-many manner.



(a) (b)

Figure 2: Relationship between error of class token and accuracy in image classification. The trend is that the less token information loss, the higher the accuracy. Our approach reserves information to the most and achieves the best accuracy.

merging. Rather than pruning uninformative tokens, they merge them into informative tokens or cluster them into fewer groups where tokens are merged into one representative token for each group. Despite great progress, there remains critical problems. **First**, tokens to be merged are exclusive. In other words, if a token is assigned to a certain group, it cannot be assigned to other groups again. The flexibility of information expression may be limited in this way, as crucial tokens may be attached to more than one token at certain moments. **Moreover**, due to severe accuracy drop, most methods [23, 31, 27] still require post-training to recover the performance, which may raise training cost. Although some methods [1, 22] claim training-free compression, the acceleration is limited.

There are some reasons accounting for the accuracy drop problems. We assume that the token reduction process can be regarded as matrix transformation of tokens. Token pruning and merging methods can be seen as special cases in this view, as shown in Fig. 1a and Fig. 1b. Specifically, the transformation matrix has diagonal-wise and block-wise form, respectively. Existing methods cannot achieve satisfactory compression may result from the special limited and exclusive transformation.

Motivated by the analysis above, we propose a many-to-many token aggregation framework (M2M-TAG) to reduce token numbers and develop an algorithm to determine the coefficient matrix dynamically for each sample. Different from existing methods, the proposed approach transforms tokens in a more flexible many-to-many manner, as shown in Fig. 1c. That means original tokens can be integrated into more than one crucial remaining token. As for the solution of the transformation matrix, we firstly put forward an attentive-based token selection strategy that dynamically select the most informative tokens. Next, these informative tokens are used to calculate the similarity between full tokens, which can represent matrix coefficient. Our method can be seen as a generalized form of previous work, because the transforming matrix will degenerate to the diagonal or block-wise one if each token is exclusively assigned to one token in reduction process, see Fig. A for detail illustration.

We claim that flexible many-to-many aggregation is necessary to improve the performance and helpful to retain foreground and background information as much as possible. The effectiveness of the transformation can be validated by the error of class token after token reduction. The error and accuracy comparison with existing SOTA TPS [32] and ToMe[1] methods are shown in Fig. 2 and results show that the proposed coefficient matrix can achieve lower error and higher accuracy.

Note that we do not introduce any trainable parameters into the framework, which can complete the compression off-the-shelf. For example, we achieve 34.8% training-free acceleration with negligible 0.1% accuracy drop on DeiT-S, even outperforming state-of-the-art methods which require fine-tuning. All results prove the effectiveness and transferability of our method. Our contributions are:

- We define token reduction process as matrix transformation of tokens, which is a general form of all previous methods and propose a token compression framework that enables more flexible many-to-many aggregation than existing methods.
- We develop a training-free algorithm of determining coefficient matrix, which can reflect many-to-many relationship between tokens, reserve token information to the most and compress models without fine-tuning.
- We conduct various experiments with competitive results and substantial acceleration across different variants and scales of vision transformers to verify the superiority of our method.

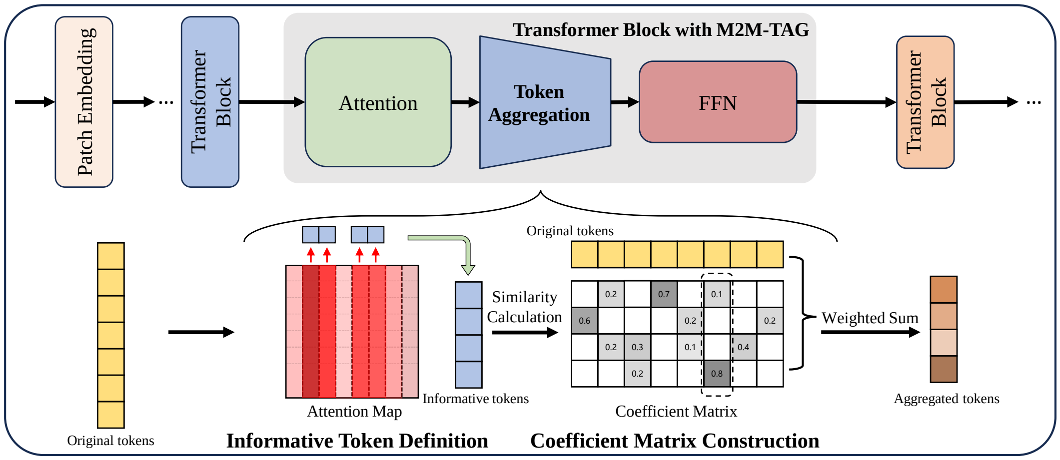


Figure 3: Detail structure of the proposed Many-to-Many Token Aggregation framework (M2M-TAG). The module is inserted between Attention and FFN with modification on attention weights. We dynamically select the most informative tokens based on attention map, determine matrix coefficient through similarity calculation, and finally obtain fewer aggregated tokens by weighted sum.

2 Method

2.1 Overview

To define token reduction process in a general way, we regard token reduction process as matrix transformation. Specifically, the equation of token reduction is shown as:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{M \times d}$ and $\mathbf{X} \in \mathbb{R}^{N \times d}$ stand for tokens before and after aggregation, N and M are token numbers $(M < N)$, and d is feature dim. The matrix $\mathbf{W} \in \mathbb{R}^{M \times N}$ can represent token relationship during token reduction process. Existing token pruning and merging methods adopt a special form of transformation matrix. To be specific, token pruning methods simply discard uninformative tokens and remain the informative tokens, and the matrix is diagonal form with elements of zero or one value on the main diagonal. Each remaining token is directly collected from the original token, and thus the coefficient matrix represents one-to-one token relationship, as shown in Fig. 1a. Token merging methods exclusively merge a group of tokens into one token and have a block-wise coefficient matrix. As long as some original tokens are integrated into one remaining token, these original tokens cannot be assigned to other tokens any more. Thus the coefficient matrix represents many-to-one token relationship, as shown in Fig. 1b.

To address the non-exclusive issue of existing methods, we propose a many-to-many token aggregation framework. There is no limitation of form of coefficient matrix and the proposed method can thus enable more flexible token aggregation. Specifically, original tokens can be integrated into more than one crucial remaining token, which reflects many-to-many token relationship, as shown in Fig. 1c. It can thus reserve token information to the most during token reduction process.

As shown in Fig. 3, we apply our Token Aggregation module between Attention and FFN to reduce the number of token. As for the construction of coefficient matrix, we select the most M informative tokens from originally N tokens. Then, we calculate the similarity between these informative tokens and full N tokens. Then, we use normalized similarity to represent the coefficient in $\mathbf{W}$. Finally, M aggregated tokens are determined by token aggregation accordingly. The informative token selection and the matrix coefficient calculation procedure will be introduced in the Section 2.2.

2.2 Many-to-Many Token Aggregation Procedure

Informative token definition. Informative tokens can be determined by token selection criterion. Some previous methods [31, 23, 44] take attention map between class token with other tokens as token selection criterion. Unlike them, we calculate informativeness level of each token based on attention map of full tokens to exploit the global relationship between all tokens, as follows:

$$\mathbf{H_j} = \sum_{i=1}^{N} \mathbf{A_{ij}}, \tag{2}$$

where $\mathbf{A_{ij}}$ is the element in the i^{th} row and j^{th} column of attention map. Intuitively, $\mathbf{H}$ stands for informativeness level of tokens. The larger $\mathbf{H_j}$ of one token is, the more information other tokens receive from it. Then we select the M most informative tokens to construct subset $\mathbb{D}_s$ from full token set $\mathbb{D}$ based on sorting:

$$\mathbb{D}_s = \underset{\mathrm{s}}{\mathrm{argmax}}\{\mathrm{H_j}, \mathrm{j} =, 1, \cdots, \mathrm{N}\}, \tag{3}$$

where $\underset{\mathrm{k}}{\mathrm{argmax}}$ means selecting k index of the largest value. Compared with methods using local aggregation size like ToMe [1], the attachment to full tokens reserves information to the most in a global spatial aggregation size.

Coefficient matrix construction. The weighted coefficient of $\mathbf{W}$ is calculated through similarity and then applied for token aggregation. Since each token j may be assigned to more than one transformed token in a non-exclusive manner, it is important to convert absolute coefficient into relative ones. To this end, we introduce assignment normalization based on $\mathrm{Softmax}$ operation with temperature τ to get the relative coefficient:

$$\mathbf{m_{ij}} = \frac{\exp\Big(sim(i,j) * \tau\Big)}{\sum_{k=1}^{M} \exp\Big(sim(k,j) * \tau\Big)}, i \in \mathbb{D}_s, j \in \mathbb{D} \tag{4}$$

where τ is the temperature and similarity measurement is cosine similarity. Then we incorporate a standard normalization along each row to make the summation of the final weighted coefficient equal to one:

$$\mathbf{W_{ij}} = \frac{\mathbf{m_{ij}}}{\sum_{j=1}^{N} \mathbf{m_{ij}}}. \tag{5}$$

Finally, the coefficient matrix $\mathbf{W}$ is obtained and aggregated token Y is a weighted sum of full tokens, as follows:

$$\mathbf{Y_i} = \sum_{j=1}^{N} \mathbf{W_{ij}} \mathbf{X_j}. \tag{6}$$

Considering simple calculation, the runtime overhead of coefficient matrix determination is negligible. Token aggregation process can thus be determined through this way and the aggregated tokens will participate in the calculation of following transform blocks instead of employing the full tokens.

3 Experiment

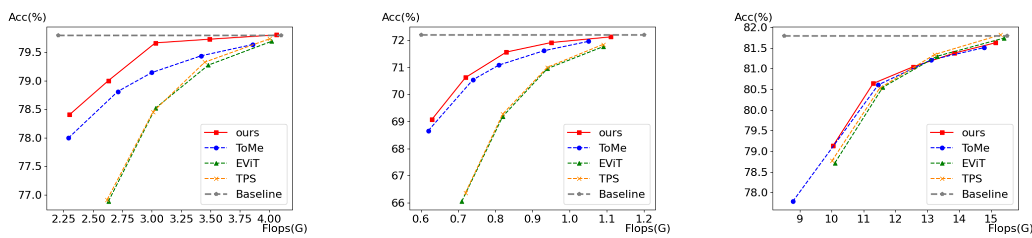
3.1 Main Results

We conduct experiments of different ViTs, such as DeiT, ViT, LV-ViTs and so on, on ImageNet-1k [6]. Then we make comparison with various token pruning and token merging methods, along with state-of-the-art models.

Results without fine-tuning. We compare our approach with other methods on DeiT [38]. Firstly we insert our token transforming approach as a plug-and-play plugin into the 4^{th}, 7^{th} and 10^{th} transformer layers without fine-tuning and denote the results with *. As is shown in Tab. 1, we achieve competitive performance. For example, on DeiT-S, we compress the model by up to 34.8% with marginally loss in accuracy. Moreover, the accuracy of Deit-S compression result on the fly is comparable or higher than other methods. Considering that all the compared methods are fine-tuned under such a large compression, the made progress is significant. It is notable that the obtained acceleration throughput is comparable or higher than all existing methods. We also evaluate our method of different compression ratio and draw accuracy-FLOPs curves compared with other methods [1, 23, 44] under same off-the-shelf setting. From Fig. 4, our method achieves significant

Table 1: Comparison of various dynamic compression on ViTs. Results marked with * are evaluated **off-the-shelf**. Results are reported after three runs.

Model	Params (M)	GFLOPs	Acc (%)	Throughput (im/s)
DeiT-S [38]	22.1	4.6	79.8	974
DynamicViT [31]	22.8	3.0 (34.8% ↓)	79.3	1503
Evo-ViT [46]	22.4	3.0 (34.8% ↓)	79.4	1510
EViT [23]	22.1	3.0 (34.8% ↓)	79.5	1487
ATS [9]	22.1	2.9 (37.0% ↓)	79.7	-
ToMe [1]	22.1	2.7 (41.3% ↓)	79.4	1552
TPS [44]	22.1	3.0 (34.8% ↓)	79.7	1428
Ours*	22.1	3.0 (34.8% ↓)	**79.7**	1451
Ours	22.1	3.0 (34.8% ↓)	**79.9**	1451
Ours/0.6	22.1	**2.6 (43.5% ↓)**	**79.7**	**1633**
ViT-Augreg-S [34]	22.1	4.6	81.4	974
ToMe* [1]	22.1	2.7 (41.3% ↓)	79.3	1564
Ours*	22.1	2.7 (41.3% ↓)	**79.8**	**1576**
ViT-AugReg-Ti [34]	5.6	1.3	75.5	2558
ToMe* [1]	5.6	0.8 (38.5% ↓)	73.8	3629
Ours*	5.6	0.8 (38.5% ↓)	**74.6**	**3639**
ViT-AugReg-B [34]	86.6	17.6	84.5	309
ToMe* [1]	86.6	11.6 (34.1% ↓)	83.3	464
Ours*	86.6	**11.4 (35.2% ↓)**	**83.7**	**469**
ViT-H [15]	632.1	167.4	86.9	35
ToMe* [1]	632.1	92.9 (44.5% ↓)	85.9	63
ToMe [1]	632.1	92.9 (44.5% ↓)	86.5	63
Ours*	632.1	92.9 (44.5% ↓)	**86.1**	**66**
Ours	632.1	92.9 (44.5% ↓)	**86.7**	**66**



(a) DeiT-S (b) DeiT-Ti (c) DeiT-B

Figure 4: Comparison with different methods under different FLOPs with off-the-shelf setting. More Results are obtained by running official codes [44, 1, 23] due to limited data in papers. Our framework is able to achieve better results especially under aggressive compression ratio.

improvements especially towards aggressive compression and obtains ***lossless compression*** of 21% and 17% for DeiT-S and DeiT-Ti, respectively without fine-tuning.

We additionally carry out experiments on ViT-AugReg [34] to evaluate the scalability of our method. Tab. 1 reveals that our method works well and gets significant improvement.

Further tuning improves the performance. Although the proposed approach can accelerate ViTs without fine-tuning to some degree, fine-tuning is also beneficial to improving the performance. We report fine-tuning results in Tab. 1, where our approach consistently outperforms all mentioned methods with state-of-the-art results. Specifically, our model can get merely -0.1% (79.7%) accuracy drop under 2.6 GFLOPs and can even gain +0.1% (79.9%) accuracy bonus under 3.0 GFLOPs compared to vanilla DeiT-S. Moreover, fine-tuned ViT-H obtains better results than ToMe [1] (86.7% v.s. 86.5%) with considerable 45% compression. Furthermore, as shown in Tab. 2, the compression result of PS-ViT can achieve no accuracy drop after fine-tuning and outperform the existing ATS [9] method by +0.2%.

Inference acceleration towards different vision transformers. To comprehensively certificate the actual acceleration of our method, we evaluate throughput with a single V100 and showcase the outcome in the last row of Tab. 1. It underlines that our approach achieves impressive inference acceleration against all existing strategies and obtains substantial 140%-200% acceleration across different models, validating the usefulness of our method. It is also notable that our approach is able to provide 188% acceleration for foundation models like Vit-H, strongly demonstrating the potential application value in foundation large vision and multimodal models.

Table 2: Comparison other ViT structure based model. Models with * report results without fine-tuning. We apply our framework on LV-ViT-S [18], T2T-ViT [51] and PS-ViT [53].

Model	Params (M)	GFLOPs	Acc (%)
T2T-ViT-14 [51]	21.5	4.8	81.5
PS-T2T-14 [37]	-	3.1	81.3
Ours-T2T-14*	21.5	3.1	**81.3**
PS-ViT-B [53]	21.3	5.4	81.7
ATS-PS-B [9]	21.3	3.7	81.5
Ours-PS-B*	21.3	3.7	81.3
Ours-PS-B	21.3	3.7	**81.7**
LV-ViT-S [18]	26.2	6.6	83.3
DynamicViT-LV-S [31]	26.9	4.6	83.0
PS-LV-ViT-S [37]	26.2	4.7	82.4
EViT-LV-S [23]	26.2	4.7	83.0
Ours-LV-S*	26.2	4.6	**83.1**

Comparison with different variants of vision transformers. We compare our method with other ViT based models that achieve progress on image classification. As shown in Tab. 2, it turns out that our compression results of LV-ViT-S without fine-tuning outperforms the previous fine-tuning methods with only 0.2% accuracy drop under comparable compression ratio, which indicates the significant improvement. We also implement our method to PS-ViT [53] and T2T-ViT [51] and it reveals in Tab. 2 that our proposed method can achieve competitive results against previous methods.

3.2 Further Analysis

Intuitive explanation. We use L2 distance of class token output for a transformer between using total input tokens and using using aggregated tokens to measure the error. We illustrate the relationship between error of class token and accuracy to explicitly analyze the effectiveness. In Fig. 2, the trend reveals that the information loss is inversely related to the classification accuracy, *i.e.*, higher accuracy indicates lower class token loss. It also certificates that our method can reserve information to the most and thus get the best results, which is in line with the analysis above.

Visualization. One key exploration of our approach is non-exclusive property of token assignment. It means that original tokens can be integrated into more than one crucial remaining token after token reduction, which previous methods can not. To have an intuitive understanding of the property, we provide heatmap for several typical informative tokens with respect to their token transformation matrix coefficient in Fig. 5. For example, informative tokens tagged with 2, 3 and 8 share some common assigned tokens as the corresponding heatmap have overlapped high activation area. The same observation applies to informative tokens tagged with 10 and 12. The assignment also reflects clearly that our method is capable of capturing information of critical locations such as eyes, noses and key parts of body as well as aggregating information from other tokens to the most.

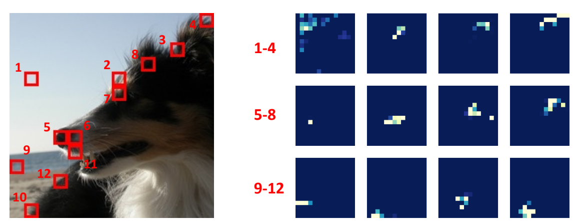


Figure 5: Informative tokens and the heatmap of transformation matrix coefficient for each informative token. Lighter color represents greater coefficient value. One original token can be assigned to multiple informative tokens, reflecting non-exclusive property.

4 Conclusion

In this paper, we define token reduction process as matrix transformation and propose a general Many-to-Many Token Aggregation (M2M-TAG) framework that allows more flexible token relation description than existing methods to reduce tokens. We also propose an algorithm which combines importance and similarity metric of full tokens to solve the transformation matrix. Due to the many-to-many aggregation, the method can reserve token information to the most during token reduction, and thus even compress models with negligible accuracy drop without fine-tuning in some cases. The obtained competitive results demonstrate the effectiveness and transferability of the method.

References

[1] 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.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[5] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 396–414. Springer, 2022.

[10] Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1039–1048, 2017.

[11] Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng-zhong Xu. Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331*, 2018.

[12] Jinyang Guo, Jianyu Wu, Zining Wang, Jiaheng Liu, Ge Yang, Yifu Ding, Ruihao Gong, Haotong Qin, and Xianglong Liu. Compressing large language models by joint sparsification and quantization. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7436–7456, 2021.

[14] Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang. Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation. *arXiv preprint arXiv:2407.11682*, 2024.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Xianglong Liu, Luca Benini, Michele Magno, and Xiaojuan Qi. Slim-llm: Salience-driven mixed-precision quantization for large language models. *arXiv preprint arXiv:2405.14917*, 2024.

[18] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in neural information processing systems*, 34:18590–18602, 2021.

[19] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2022.

[20] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[22] Weicong Liang, Yuhui Yuan, Henghui Ding, Xiao Luo, Weihong Lin, Ding Jia, Zheng Zhang, Chao Zhang, and Han Hu. Expediting large-scale vision transformer for dense prediction without fine-tuning. *Advances in Neural Information Processing Systems*, 35:35462–35477, 2022.

[23] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

[24] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. Hrank: Filter pruning using high-rank feature map. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1529–1538, 2020.

[25] Xiangcheng Liu, Tianyi Wu, and Guodong Guo. Adaptive sparse vit: Towards learnable adaptive token pruning by fully exploiting self-attention. *arXiv preprint arXiv:2209.13802*, 2022.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[27] Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, and Jingdong Wang. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2023.

[28] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272, 2019.

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[31] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[32] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11997, 2022.

[33] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34:12786–12797, 2021.

[34] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. 2022.

[35] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[36] Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Aliari Zonouz, and Bo Yuan. Chip: Channel independence-based pruning for compact neural networks. *Advances in Neural Information Processing Systems*, 34:24604–24616, 2021.

[37] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Thomas Verelst and Tinne Tuytelaars. Dynamic convolutions: Exploiting spatial sparsity for faster inference. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2320–2329, 2020.

[41] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018.

[42] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Advances in Neural Information Processing Systems*, 34:11960–11973, 2021.

[43] Zidu Wang, Xuexin Liu, Long Huang, Yunqing Chen, Yufei Zhang, Zhikang Lin, and Rui Wang. Qsfm: Model pruning based on quantified similarity between feature maps for ai on edge. *IEEE Internet of Things Journal*, 9(23):24506–24515, 2022.

[44] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2092–2101, 2023.

[45] Xinjian Wu, Fanhu Zeng, Xiudong Wang, and Xinghao Chen. Ppt: Token pruning and pooling for efficient vision transformers. *arXiv preprint arXiv:2310.01812*, 2023.

[46] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2964–2972, 2022.

[47] Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18547–18557, 2023.

[48] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[49] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022.

[50] Fang Yu, Kun Huang, Meng Wang, Yuan Cheng, Wei Chu, and Li Cui. Width & depth pruning for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3143–3151, 2022.

[51] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[52] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022.

[53] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 387–396, 2021.

[54] Edouard Yvinec, Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Red: Looking for redundancies for data-freestructured compression of deep neural networks. *Advances in Neural Information Processing Systems*, 34:20863–20873, 2021.

[55] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[56] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019.

[57] Yiren Zhou, Seyed-Mohsen Moosavi-Dezfooli, Ngai-Man Cheung, and Pascal Frossard. Adaptive quantization for deep neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Appendix

A Related Work

A.1 Efficient Vision Transformers

Transformer [39, 7, 29, 5] is first introduced in NLP tasks and Vision Transformer [8] successfully demonstrates superior results in visual tasks. Various research has been carried out to explore efficient vision transformers [38, 18, 51, 3, 32, 26]. DeiT [38] achieves competitive performance through an efficient training paradigm under distillation knowledge. LV-ViT [18] devises a new training objective and generates a dense score map to extract rich local information.

A.2 Dynamic Vision Transformers

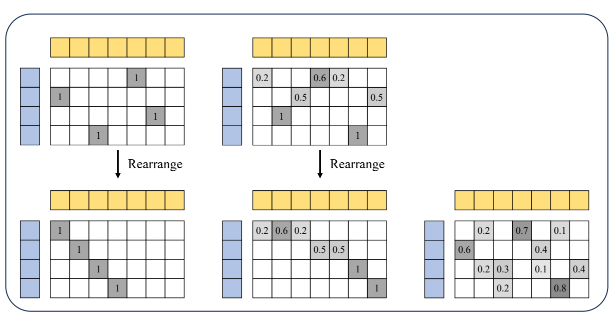
Due to high computational cost of vision transformers, many attempts are made to reduce tokens dynamically according to input content. Token compression strategy can mainly be divided into token pruning and token merging.

Token Pruning discards uninformative tokens directly. It is a straightforward way for adaptive compression. Different kinds of importance assessments are carried out to prune unnecessary parts dynamically according to the complexity of the input images [9, 49, 31, 37]. DynamicViT [31] designs a lightweight prediction module to effectively prune redundant tokens. ATS [9] proposes an adaptive token sampling method to sample tokens dynamically during inference. Nevertheless, simple pruning suffers from severe accuracy drop due to direct loss of information in images.

Token Merging combines tokens together rather than pruning them directly to reserve more information [1, 23, 44, 27]. EViT [23] reduces the tokens by measuring the attention with class token to identify attentive token and fuses inattentive ones dynamically. ToMe [1] incorporates a bipartite matching process to combine tokens according to their similarity. TPS [44] squeezes tokens into several reserved ones via exclusive matching. However, these methods merge tokens exclusively, restricting the flexibility and the utilization of information. By contrast, we incorporate more flexible many-to-many transforming for compression.

B Detailed Analysis about Transforming Matrix

We give a general form of token reduction and describe all token reduction process as a aggregation of full tokens, where coefficient matrix is consisted of coefficients between full and transformed tokens. As is illustrated in Fig. A, typical form of Token Pruning, Token Merging and the proposed Token Aggregation can be expressed in the many-to-many framework of a coefficient matrix and aggregated tokens are weighted sum of full original tokens along each row, *i.e.*, the sum of each row in coefficient matrix equals one.



(a) Token Pruning (b) Token Merging (c) **Token Transforming**

Figure A: Detailed explanation of coefficient matrix. Yellow and blue tokens represent full tokens and aggregated tokens, respectively. For (a) and (b), original matrix can be rearranged into diagonal and block-wise matrix if the order of the tokens are ignored.

Table A: Influence of scaling factor and assignment normalization.

Temperature τ	20	50	100	170	250
Acc (%)	79.21	79.53	79.54	79.58	79.58

Table B: Comparison of different initialization strategy and similarity score.

Strategy	GFLOPs	Acc (%)
Initialization strategy		
Uniform	3.0	78.3
Class token	3.0	79.5
Informative token **(Ours)**	3.0	**79.6**
Similarity score		
Euclidean distance	3.0	78.5
Cosine distance **(Ours)**	3.0	**79.6**

Considering that all previous methods reduce tokens exclusively, the transformation matrix of Token Pruning and Token Merging shown in Fig. Aa and Fig. Ab can easily be expressed in the form of a diagonal and block-wise matrix, respectively for better illustration, if the order of the tokens are ignored, which does not affect the results actually. Specifically, Token Pruning discards inattentive tokens and directly uses attentive tokens, thus the values of coefficient in coefficient matrix are all one on diagonal and zero in other locations. Moreover, Token Merging separates tokens into several groups and fuses each group into one token, therefore all tokens can not fuse into multiple groups, *i.e.*, each column of transformation matrix has only one element. By contrast, our proposed method does not impose restrictions on the form of coefficient matrix and is a general form of previous methods, which is helpful to reserving information to the most.

C Implementation Details

We employ Token Aggregation at certain stages of transformer layers and report Top-1 accuracy for performance comparison. Following previous work [1, 23], we provide two types of aggregation strategy, namely reserving fixed ratio of ρ and reducing fixed number of r at each aggregation stage. We use "$/$" after model name to indicate the reserving ratio or reducing number of each aggregation layer. For example, "$/0.7$" means reserving 70% of the token after each aggregation.

We insert token transforming at 4^{th}, 7^{th} and 10^{th} layers for DeiTs as they are all composed of 12 transformer blocks. Number of reducing tokens r is selected or keeping ratio ρ is set to 0.7 to match the compression ratio. For DeiT [38] in main results, and temperature is selected $\tau \in \{150, 170, 200, 250\}$, respectively.

D Ablations

We conduct comprehensive ablation study on image classification with DeiT-S to verify the effectiveness of each component. The setting is reducing fixed number of 50 tokens in the 4^{th},7^{th} and 10^{th} transformer layers by default for convenient explanation unless otherwise stated.

Hyperparameter selection of M2M-TAG. We begin with different hyperparameters related to our token aggregation framework. Specifically, we study the influence of temperature τ. As shown in Tab. A, the performance of the framework is not that sensitive to hyperparameter and for temperature, the performance levels off over a wide range as temperature varies, which verifies the robustness of our approach. We introduce parameter selection for different tasks in detail in Supplementary Material.

Comparison with different token initialization and similarity strategy. We analyse the necessity of each component proposed in our framework, *i.e.* token initialization strategy and similarity score. For the former, we compare our informative token initialization with uniform and class attention initialization, which refer to defining the initial tokens according to adaptive average pool [22] and attention with class token [23, 44], respectively. For the latter, we replace our cosine distance based method with euclidean distance one. As shown in Tab. B , no matter what the alternative approach is, there is a sharp accuracy drop. Thus, the effectiveness and transferability of each component is demonstrated.

E Limitations

The primary limitation is that we merely conduct experiments on image classification task. However, since our many-to-many token aggregation framework is non-parametric and does not rely on class token, we can construct a unified framework for both classification and dense prediction task through designing a token recovery module for dense prediction. We will regard extension to dense prediction tasks as further work and complement this part in the near future.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See in abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See in limitations in appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical result is included in this work.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See in Sec. 3 and implementation details in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Core code will be available for non-commercial use in accordance with confidentiality upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See in Sec. 3 and implementation details in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide average results after three runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See in Sec. 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No boarder societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We correctly cite the original paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.