Salience-Based Adaptive Masking: Revisiting Token Dynamics for Enhanced Pre-training

Hyesong Choi[1], Hyejin Park[1], Kwang Moo Yi[2], Sungmin Cha[3], and Dongbo Min[1†]

[1] Ewha W. University
[2] University of British Columbia
[3] New York University

Abstract. In this paper, we introduce Saliency-Based Adaptive Masking (SBAM), a novel and cost-effective approach that significantly enhances the pre-training performance of Masked Image Modeling (MIM) approaches by prioritizing token salience. Our method provides robustness against variations in masking ratios, effectively mitigating the performance instability issues common in existing methods. This relaxes the sensitivity of MIM-based pre-training to masking ratios, which in turn allows us to propose an adaptive strategy for 'tailored' masking ratios for each data sample, which no existing method can provide. Toward this goal, we propose an Adaptive Masking Ratio (AMR) strategy that dynamically adjusts the proportion of masking for the unique content of each image based on token salience. We show that our method significantly improves over the state-of-the-art in mask-based pre-training on the ImageNet-1K dataset. Code and model parameters are available at `https://github.com/doihye/SBAM`.

Keywords: Self-supervised learning · Masked image modeling · Masked autoencoder

1 Introduction

Recent drastic improvements in various Computer Vision tasks rely heavily on Transformer architectures [12]. A critical component that enables these architectures is the necessity of large-scale data [9], which is not always readily available. Naturally, pre-training with pretext tasks has become a popular solution as a workaround, represented by Masked Image Modeling (MIM) [1, 11, 14, 36], inspired by how Masked Language Modeling (MLM) [2, 7, 10, 24] has reshaped the

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00222385) and the National Research Foundation of Korea, Republic of Korea (2022M3H9A2083956) and from the Korea Basic Science Institute, Republic of Korea (National Research Facilities and Equipment Center) funded by the Ministry of Education, Republic of Korea (2019R1A6C1010020). † Corresponding author: dbmin@ewha.ac.kr

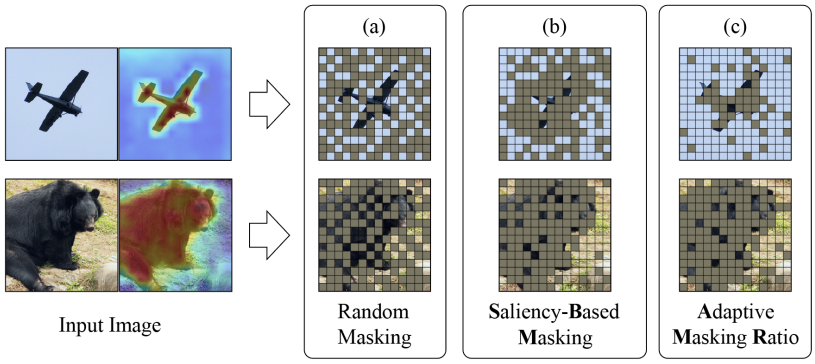

Fig. 1: Overview of SBAM. Whereas (a) Random Masking must rely completely on chance and carefully tuned masking ratio to guarantee effective masking, (b) the proposed SBAM strategically masks tokens based on the token salience. The robustness of SBAM paved the way for the introduction of (c) AMR, which implements a tailored masking ratio for each sample in the dataset.

Natural Language Processing landscape. These strategies involve masking a subset of the input data and predicting those that have been hidden, thus forcing the deep network to infer underlying concepts.

While MIM has significantly advanced self-supervised learning in vision, its conventional approach to randomly selecting tokens for masking falls short of harnessing the full potential of visual data [18]. Unlike in text, where randomness might obscure key semantic units, the visual domain's complexity and token redundancy demand a more strategic masking protocol to ensure model comprehension. This necessity prompts us to explore a refined masking methodology, targeting a selection process that achieves the goal of image understanding for pre-trained models, thereby bridging the gap in modality-specific pre-training strategies.

Various techniques [18, 25, 35] have thus been proposed for more effective masking. While these methods all strive toward improved pre-training, an oversight shared amongst all methods is that they do *not* regard the contribution of each token within the overall composition of the image. It is crucial to scrutinize the interconnections between tokens to ensure that the token masking includes the tokens that play a pivotal role within the image. Moreover, prior masking methodologies [18, 25, 35] fail to address the critical consideration of the masking ratio, a factor that ought to dynamically adjust in response to the size and quantity of pivotal objects embedded within the image. This is primarily due to the difficulty reported in various works [14, 36, 39] that minor changes in the optimized masking ratio can lead to performance instability, rendering such considerations difficult to implement. Thus, these methods must rely completely on chance and carefully tuned masking ratio to guarantee effective masking. Besides, contemporary masking strategies [18, 25, 35] often grapple with high complexity, burdened by intricate distillation frameworks [18], auxiliary detection processes [35], and duplicated forward processes [25].

In this work, we introduce a simple yet novel approach that focuses on token dynamics, termed **S**alience-**B**ased **A**daptive **M**asking (**SBAM**), which aims to strategically select masking locations by discerning perceptual prominence within the visual data. Specifically, the proposed method leverages the directional emphasis from attention mechanisms to identify the image tokens pivotal to the visual context. Significantly, our methodology stands apart from the conventional attention-based approach [25] by leveraging the token's outgoing weight to calculate its 'token salience' within the image and prioritizing those that have high salience to be masked. We also infuse a degree of randomness into token salience to enrich the diversity of mask generation. Further details of SBAM are elaborated in Sec. 3. Thus, without any significant additional cost, we can consider the token's prominence within the image.

Crucially, the strength of the SBAM approach lies in its robustness to the varying masking ratios—a notable vulnerability in established baselines [14, 36, 39], which struggles from performance instability with even minor variations in optimized masking ratio. The robustness of SBAM is attributable to the proposed analytical precision in token dynamics, selectively masking tokens pivotal to the image's entirety. As a result, SBAM exhibits a diminished likelihood of masking redundant or non-essential segments, offering a stable alternative to the random masking or preceding masking approaches. Along with improving robustness to mask ratio variations, SBAM significantly enhances pre-training efficiency. A comprehensive evaluation of this is provided in Sec. 4.

Establishing robustness against variations in masking ratios has empowered us to expand the discourse on image masking into a pioneering aspect, introducing an innovative paradigm: an **A**daptive **M**asking **R**atio (**AMR**). We find having masking ratios that *adapt* throughout training to be highly effective, as it allows the masking process to be *tailored* to each sample in the dataset. For instance, each image may benefit from different masking ratios, as one image might have a close-up of a bear which would enjoy high masking ratios, whereas one might have an airplane in the sky that would require a lower masking ratio; see (c) of Fig. 1. The proposed token salience, reapplied in this context, serves to finely determine the dynamic masking ratio by quantifying the proportion of tokens exhibiting high salience. More details can be found in Sec. 5. The proposed AMR implements an adaptive strategy that respects the distinctiveness of visual data and thus achieves significant performance gains when applied to various baselines (refer to Sec. 6.2 and 6.3. We note that this type of 'tailoring' to each image for pre-training is impossible with any existing method.

In sum, this paper offers an innovative masking strategy that enhances the robustness and effectiveness of pre-trained models, setting the stage for a notable shift in the field of self-supervised learning. More importantly, the proposed method can be universally applied across any MIM framework that exploits token masking. We evaluate our SBAM and AMR on data-hungry models like ViT-L/ViT-B [12] on ImageNet-1K [9] datasets, achieving significant performance improvements in both fine-tuning and linear probing accuracy.

To summarize, our contributions are:

- we present Saliency-Based Adaptive Masking (SBAM), a novel effective method for MIM pre-training that focuses on token salience;
- being saliency-based, without a significant increase in computation, we allow effective masking that is robust to masking ratios;
- empowered by the robustness to masking ratio, we propose an Adaptive Masking Ratio (AMR) that allows *tailored* masking for each sample;
- we evaluate our method on ImageNet-1K datasets, achieving notable enhancements in both fine-tuning and linear probing accuracy.

2 Preliminaries

Within the Masked Image Modeling (MIM) domain, the crux lies in the strategic corruption and subsequent reconstruction of image segments. This process hinges on two core operations: *Random Masking* of image tokens and *Reconstruction* of corrupted tokens to guide the learning process.

For a given image sequence $X \in \mathbb{R}^{N \times L \times D}$, where N, L and D denote batch size, number of tokens per image and dimensionality of each token respectively, and specified masking proportion γ, the general random masking process conducted by Φ_{mask} is defined as:

$$X_{\text{masked}}, M = \Phi_{\text{mask}}(X, \gamma), \tag{1}$$

where X_{masked} represents the visible tokens as a result of the post-application of the random mask M, with $M \in \{0,1\}^{N \times L}$ indicating the presence (1) or absence (0) of masking for each element.

The key component for learning in MIM is the reconstruction error $\mathcal{L}_{\text{MIM}}$, which is computed by the mean squared error (MSE) between the predicted representation of masked tokens, denoted as $\hat{X}$, and their normalized original counterparts, represented by X:

$$\bar{X} = \frac{X - \mu(X)}{\sigma(X) + \epsilon}, \quad \mathcal{L}_{\text{MIM}} = \frac{1}{\sum_{i,j} M_{i,j}} \sum_{i=1}^{N} \sum_{j=1}^{L} M_{i,j} \cdot \|\hat{X}_{i,j} - \bar{X}_{i,j}\|_2^2. \tag{2}$$

Here, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation of X respectively. $\|\cdot\|_2^2$ denotes the squared L_2 norm and ϵ is a small value for numerical stability. This formulation presents the core of MIM's training objective, focusing explicitly on the reconstruction of the masked portions of the input, encouraging the model to infer corrupted information from the unmasked context.

3 Salience-Based Masking (SBM)

Masked Image Modeling (MIM) has greatly pushed forward self-supervised learning in the visual domain, yet its standard practice of randomly masking tokens fails to fully capture the richness of visual information. To overcome this

limitation, we present Salience-Based Masking (SBM), a novel masking methodology that selects masking locations by revisiting token dynamics.

Given an input tensor $X \in \mathbb{R}^{N\times(L\times D)}$, the first step involves computing an affinity matrix $\mathcal{A} \in \mathbb{R}^{N\times L\times L}$ through a batch matrix-matrix product between X and $X' \in \mathbb{R}^{N\times(D\times L)}$. Subsequently, we apply the softmax function $\rho(\cdot)$ to the affinity matrix, resulting in a normalized affinity matrix denoted as $\hat{\mathcal{A}} = \rho(\mathcal{A}) \in \mathbb{R}^{N\times L\times L}$. The softmax function is implemented as follows:

$$\rho(\mathcal{A})_{n,i,j} = \frac{e^{a_{n,i,j}}}{\sum_{k=1}^{L} e^{a_{n,i,k}}}. \tag{3}$$

Here, $a_{n,i,j}$ represents an element of $\mathcal{A}$ and e is the base of the natural logarithm. Note that, for each n-th element with a shape of $\mathbb{R}^{L\times L}$, this function normalizes the rows of it, transforming them into probabilities that sum to 1.

Crucially, our approach distinguishes itself from conventional attention-based methods by utilizing the sum of *outgoing weight* of each token to determine the 'token salience' $S \in \mathbb{R}^{N\times L}$ in the image, which is represented by the column-wise summed score of $\hat{\mathcal{A}}$:

$$S = \mathcal{N}(\Sigma_{j=1}^{L}\hat{\mathcal{A}}_{:,j,:}), \qquad \mathcal{N}(x) = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{4}$$

In the affinity map, the row-wise score represents the incoming weight, reflecting the extent to which other tokens influence the corresponding token. Conversely, the column-wise score signifies the outgoing weight, indicating the contribution of the corresponding token to others. By summing these scores, it becomes feasible to quantify the token's overall impact on the image, thereby defining token salience.

The adaptive masking process of SBM is formulated with the token salience S. Notably, exclusive reliance on S for masking can precipitate a decline in performance (refer to Fig. 7). To mitigate this issue, we incorporated an element of randomness into the masking process to get the adjusted token salience, denoted as $\tilde{S} = S + N$, where $N \sim U([0, 0.5)^{N\times L})$ represents a noise realization sampled from a multivariate uniform distribution U.

Then, the sampling of tokens for masking is guided by $\tilde{S}$. We sort $\tilde{S}$ in ascending order and select tokens corresponding to the top K scores, where $K = \lceil L \cdot (1 - \gamma)\rceil$. Consequently, an adaptive binary mask M is constructed, where $M_i = 0$ for the K selected tokens, and $M_i = 1$ for the remainder.

The proposed token salience offers a more intuitive and cost-efficient strategy for determining which tokens to mask. SBM therefore revisits token dynamics in the image context, streamlining the conventional complex masking process and enabling strategic masking.

4 Evaluation on SBM

In this section, the evaluation critically examines the robustness of the proposed SBM, especially against varying masking ratios, while also assessing its enhanced

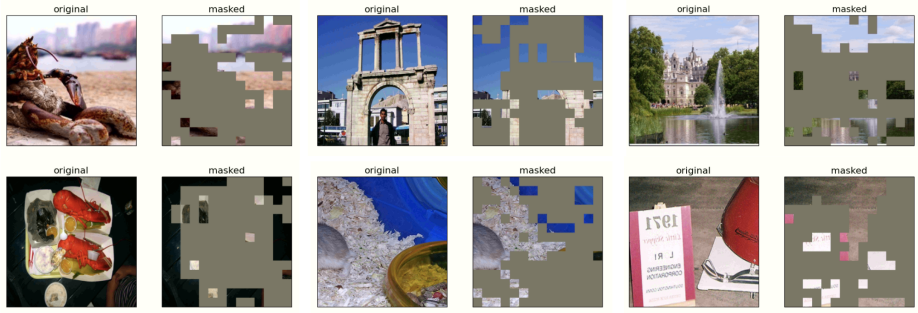

Fig. 2: Qualitative example of SBM. SBM introduces 'token salience' to prioritize and mask tokens with high significance. Hence, it is qualitatively confirmed that particularly important objects with high contribution within the image are selectively masked. Moreover, by integrating randomness with token salience, masks are probabilistically assigned to the background and less significant tokens, enriching the diversity of the token masking.

pre-training efficiency, evidenced by performance gains and faster convergence. Moreover, Fig. 2 qualitatively shows that SBM selectively masks only those tokens that contribute significantly to the image.

4.1 Robustness to Masking Ratio Variability

A crucial aspect of the evaluation of SBM focuses on the robustness, particularly in the context of varying masking ratios. Prior established baselines, such as the widely used Masked Autoencoder (MAE) [14], often exhibit significant performance fluctuations with even minimal adjustments to the masking ratio. This sensitivity undermines the practicality and generalizability of such methods, especially in diverse real-world scenarios where optimal masking ratios may not be consistent across datasets. In contrast, SBM introduces a novel approach of selectively masking tokens based on their salience within an image. As a result, SBM exhibits a diminished likelihood of masking redundant or trivial tokens, thereby maintaining a stable performance across a broad spectrum of masking ratios.

Fig. 3 showcases the performance stability of SBM against strong baseline MAE, underscoring its superior resilience to changes in masking ratios. Our evaluation leverages the Performance Improvement over Masking Ratio (PIMR) metric, a normalized measure that quantifies how each model's performance at a given masking ratio stands against its performance at the lowest ratio, thereby reflecting the relative improvement. This metric is instrumental in revealing the impact of increased masked data on model training. The PIMR is defined as follows:

$$\mathrm{PIMR}(M) = \frac{P(M) - P(M_{min})}{P(M_{max}) - P(M_{min})}, \tag{5}$$

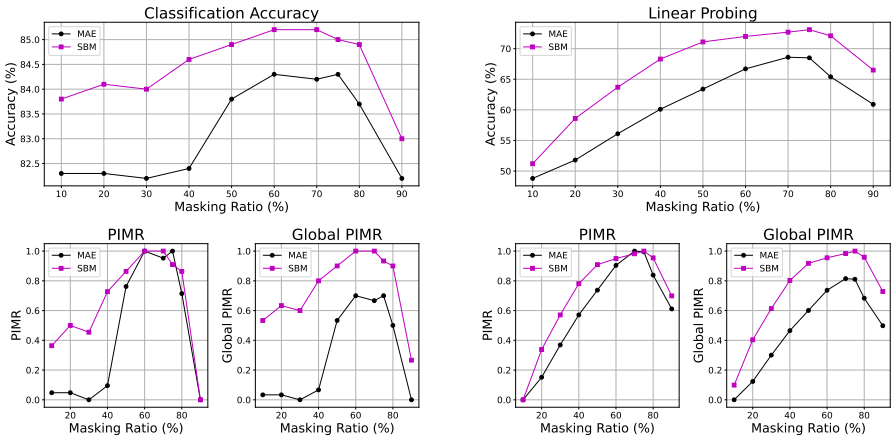

Fig. 3: Evaluation of robustness across varied masking ratios. To evaluate the robustness of SBM, we report the comparative analysis of image classification performance on ImageNet-1K dataset [9] against the baseline method, MAE [14], using ViT-L [12] as a backbone. The upper graphs display the performance of the methods at different masking ratios, while the lower graphs illustrate the Performance Improvement over Masking Ratio (PIMR) and Global PIMR. These measures indicate the extent of each model's performance enhancement as the masking ratio increases from the lowest to higher ratios. SBM significantly outperforms MAE in every measures, demonstrating its superior effectiveness in handling various masking ratios and enhanced pre-training performances.

where $P(M)$ is the performance at masking ratio M. $P(M_{min})$ and $P(M_{max})$ denote the minimum and maximum observed performances, respectively. A PIMR value closer to 1 signifies a substantial improvement relative to the range of observed performances. The PIMR graph in classification accuracy demonstrate the robustness of the SBM strategy, where it maintains a competitive edge over MAE across various masking ratios. Notably, SBM shows a minimal performance drop at lower masking ratios compared to MAE, indicating its effectiveness even with sparse data presence.

This advantage is particularly evident in the Global PIMR metric, where SBM's performance remains consistently high. For the Global PIMR calculation, we normalize performance relative to the most extensive range of performance observed among all models under comparison. This means that instead of comparing to the best and worst performances of a single model, Global PIMR considers the best and worst across both MAE and SBM, which provides a universal performance context. Thus, the metric reflects a model's improvement not in isolation but rather in relation to its peers, which can be seen in the equation:

$$\text{Global PIMR}(M) = \frac{P(M) - P(M_{Gmin})}{P(M_{Gmax}) - P(M_{Gmin})}, \tag{6}$$

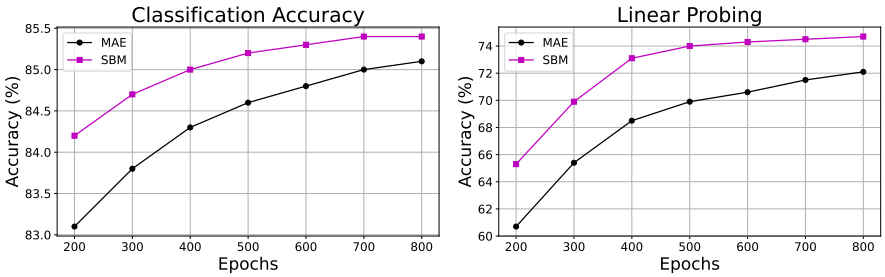

Fig. 4: Performance evaluation of SBM with respect to the pre-trained epochs. We report the comparison of image classification accuracy on ImageNet-1K [9], pre-trained on ViT-L [12]. The left graph displays fine-tuning accuracy, whereas the right graph illustrates linear probing accuracy, both over a range of pre-trained epochs. The curves illustrate that SBM surpasses MAE [14] in pre-training effectiveness in every trained epoch, and also validates its quicker attainment of converged performance levels.

where $P(M_{Gmax})$ and $P(M_{Gmin})$ are the global minimum and maximum performance values observed across both MAE and SBM models, respectively. This broader evaluation framework further underscores SBM's superior resilience and ability to maintain high accuracy across varying degrees of masking, outperforming the MAE baseline even when less information is available for learning.

As shown in the graphs, SBM consistently outperforms MAE across a spectrum of masking ratios, demonstrating its remarkable stability even at lower ratios. This is because in random masking, as the masking ratio decreases, the chance of including a crucial token in the mask lowers; conversely, SBM consistently masks pivotal tokens, regardless of the masking ratio.

Furthermore, the classification accuracy graphs displayed above demonstrate that the application of SBM results in a notable enhancement in performance compared to MAE for all ratios. This reveals that the proposed SBM not only exhibits resilience to variations in the masking ratio but also significantly boosts pre-training efficacy, irrespective of the masking ratio. Strategically masking pivotal information enhances model performance and accelerates convergence by encouraging a comprehensive understanding of the visual context through a focus on essential tokens.

4.2 Enhanced Pre-training Efficiency

Beyond the robustness to masking ratio variations, SBM's efficacy is further demonstrated through its enhanced pre-training capabilities. Traditional MIM strategies often require extensive computational resources and time for model convergence, primarily due to the indiscriminate masking of image tokens which can hinder the learning process by obfuscating essential visual cues. By strategically selecting pivotal tokens for masking, SBM ensures that crucial tokens are leveraged during training, which leads to substantial improvement in pre-training performance and facilitates a more focused model convergence process.

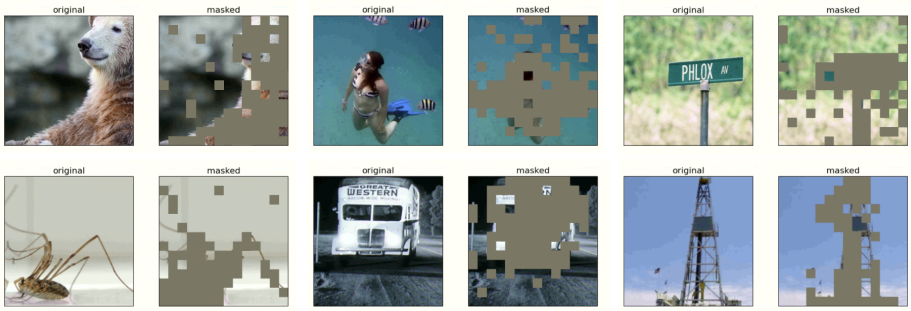

Fig. 5: Qualitative example of SBAM, which combines SBM and AMR. Having masking ratios that adapt throughout training is highly effective, as it allows the masking process to be tailored to each sample in the dataset, accommodating the unique composition and object sizes within each image, as shown in the above qualitative samples.

Fig. 4 showcases the comparative pre-training efficiency of SBM against MAE [14], underlining SBM's reduced pre-training duration without a trade-off in accuracy. Initially, SBM secures a distinct lead in classification accuracy and continues to demonstrate this advantage across the training epochs, as seen in the left graph. The right graph, representing linear probing accuracy, further confirms SBM's higher initial performance, which stabilizes near peak levels well before 800 epochs. These findings highlight SBM's ability to prioritize significant features during early training stages, resulting in a new alternative for both convergence speed and pre-training performance in the pre-training of masked image models.

5 Adaptive Masking Ratio (AMR)

Achieving stability against changes in masking ratios has enabled us to advance the discourse on image masking, introducing a novel perspective: an **A**daptive **M**asking **R**atio (**AMR**). This innovative approach acknowledges the inherently diverse visual narratives presented by individual images and adjusts the masking ratio to fit different object sizes and classes within them.

The proposed token salience $S = \mathcal{N}(\Sigma_{j=1}^{L} \hat{A}_{:,j,:})$ forms the basis for determining AMRs. The AMR R_{dyna} is computed based on the distribution of salience scores across tokens, adjusted by a predefined variability parameter δ and the base mask ratio r:

$$R_{dyna} = r - \Delta r + 2\Delta r \times \text{mean}(1_{S>\delta}). \tag{7}$$

Here, Δr denotes the range of allowable variation in the masking ratio, as R_{dyna} can range from $r - \Delta r$ to $r + \Delta r$. δ is the salience threshold for distinguishing highly salient tokens, and 1 is the indicator function that identifies tokens exceeding δ.

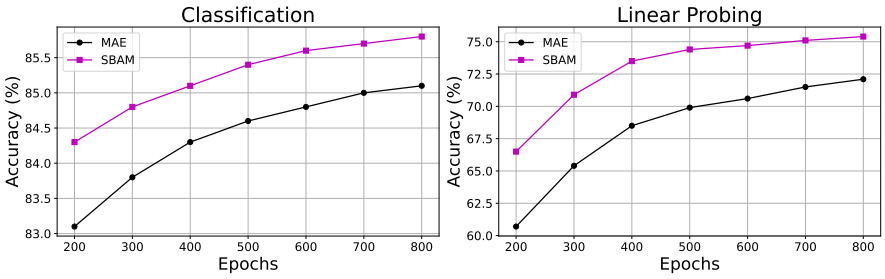

Fig. 6: Performance comparison of SBAM with respect to the pre-trained epochs. We report the comparison of image classification accuracy on ImageNet-1K [9], pre-trained on ViT-L [12]. The left graph illustrates classification accuracy across epochs, while the right graph shows the accuracy obtained through linear probing. Both results indicate a significant improvement in pre-training performance when AMR is applied, which not only achieves higher accuracy earlier in the training process but also maintains a lead at convergence.

With R_{dyna} established, we adjust the number of tokens to be masked accordingly. This dynamic adjustment of masking ratios ensures that the masking process is not uniformly applied but is instead sensitive to the visual information's inherent salience, promoting a more effective learning mechanism by focusing on the informative segments of the image. Fig. 5 presents a qualitative example of AMR, demonstrating the effectiveness of adaptive masking ratios that customize the masking process for each dataset sample, thereby accounting for the unique composition and object sizes within each image. As a result, the proposed **S**aliency-**B**ased **A**daptive **M**asking (**SBAM**), which combines SBM and AMR, employs an adaptive approach, leading to enhanced performance across different models (See Fig. 6 and Tab. 1) and setting a new standard for tailored image masking.

6 Experiments

6.1 Implementation Details

Our evaluation approach involves deploying the proposed SBM and AMR against the baseline to assess performance enhancements of the comprehensive method, SBAM. To ensure a fair comparison, we maintain consistency with the baseline method's hyperparameters and network architectures. Notably, to preserve the integrity of our experiments, we ensured uniform hardware utilization and experimental conditions across both our method and the reproduced baseline, all employing 8*A6000 GPUs. Given that this fixed GPU configuration diverges from those used in prior methods, discrepancies between our reproduced performance and the performance documented in existing papers may arise. For the fair experimental schedule, the reproduced 400-epoch performance of baseline

Table 1: Comprehensive performance results of applying SBAM to various baseline methods. We report the comparison of image classification fine-tuning accuracy on ImageNet-1K [9] dataset. The consistent performance improvement of SBAM across various baseline methods demonstrates the efficacy of SBAM as a scalable methodology capable of enhancing a variety of MIM frameworks.

Method	Baseline	Baseline+SBAM
MAE (ViT-L) [14]	84.3	85.1
MAE (ViT-B)	82.9	83.6
BootMAE (ViT-B) [11]	84.1	84.8
iBoT (ViT-B) [43]	71.5	74.4
CMAE (ViT-B) [17]	83.8	84.5

methods and the proposed method were *all equally measured in intermediate stages in the training towards 800 epochs*. A more detailed implementation description can be found in the Supplementary material.

6.2 Evaluation on SBAM

The graphs in Fig. 6 demonstrate the efficacy of the SBAM in comparison to the MAE model throughout the pre-training phase. Specifically, the left graph indicates that SBAM starts with a higher accuracy than MAE at the 200 epochs mark and continues to outperform MAE [14] at every subsequent checkpoint. By the 800 epochs mark, SBAM not only achieves a significant accuracy enhancement but also shows a more rapid improvement in the earlier epochs, suggesting that SBAM requires fewer epochs to achieve similar or better performance compared to MAE. In the context of linear probing, depicted in the right graph, the trend is similar. SBAM consistently achieves higher accuracy than MAE from the outset, and this performance gap is maintained as training progresses. The curves for SBAM and MAE become flat toward 800 epochs, indicating that 800 epochs or later is the point of convergence in performance. This observation underscores SBAM's substantial advantage over MAE, both in terms of converged performance and convergence speed. Overall, the performance trends captured in these graphs suggest that SBAM is more efficient during pre-training, reaching higher levels of accuracy faster than MAE.

6.3 Evaluation on Various Baselines

The integration of SBAM into various baseline methodologies demonstrates a notable enhancement in pre-training performance, as summarized in Tab. 1. All experiments report the classification accuracy on 400 epochs, except for iBoT [43] which is pre-trained for 100 epochs. Applying SBAM to the large-scale variant of MAE [14] (MAE (ViT-L)) yields a noteworthy enhancement, elevating the baseline accuracy from 84.3% to 85.1%, marking a significant advancement. Integration of SBAM to the MAE using ViT-B also experienced significant performance

Table 2: Comparative evaluation of SBAM against the state-of-the-art masking strategy AMT [25].

Method	Acc (%)
AM [25]	82.5
AMT [25]	82.8
SBAM	83.6

Table 3: Comparative evaluation of SBAM against the state-of-the-art masking strategy AttMask [18].

Method	Acc (%)
AttMask-High [18]	72.5
AttMask-Hint [18]	72.8
SBAM	74.4

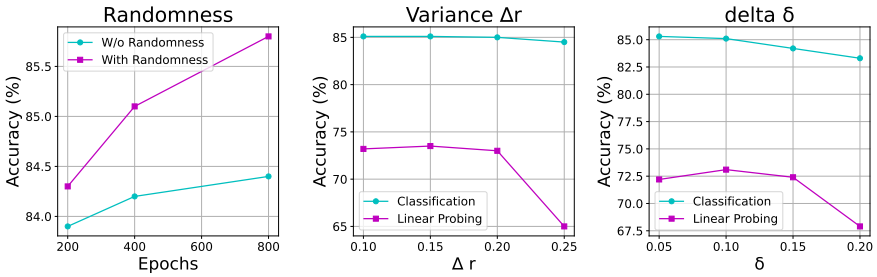

Fig. 7: Comprehensive ablation studies of the impact of randomness, variance (Δr), and delta (δ) of the proposed SBAM approach.

gains, underscoring that the benefits of SBAM are not limited to specific model architectures and can provide substantial pre-training efficacy for a variety of models. The performance enhancement of SBAM on various models including BootMAE [11], iBoT, and CMAE [17] highlights SBAM's generalizability and efficacy in augmenting various model structures with significant accuracy gains.

In conclusion, the consistent improvement across various baselines validates the efficacy of SBAM as a scalable enhancement tool. It not only boosts performance in standard settings but also bridges the gap in more challenging learning scenarios, marking it as a pivotal development in masked image modeling pre-training techniques.

6.4 Comparison with Masking Methods

We conducted a comparative evaluation of our SBAM masking strategy against the established masking strategies of AMT and AttMask. This comparison was performed by applying SBAM to the baselines previously employed by AMT [25] and AttMask [18] methods, specifically MAE [14] (ViT-B, 400 epochs) and iBoT [43] (ViT-B, 100 epochs), to ascertain the enhancements introduced by our approach.

In Tab. 2, the AM and AMT strategy [25], which implement basic masking and selective masking based on semantic importance, achieve an image classification fine-tuning accuracy of 82.5 and 82.8, respectively. When applied to the

same baseline model, SBAM outperforms both approaches by achieving an accuracy of 83.6. This implies that the efficacy of self-supervised pre-training can be maximized by defining token salience based on the outgoing weight of the token, as opposed to the incoming weight [25].

Furthermore, we compare SBAM to two variants of AttMask [18], applied within the iBoT [43] framework in Tab. 3. Both AttMask-High and AttMask-Hint incorporate the selective masking strategy of distillation setup which leverages similarity to classification token. SBAM stands out with an accuracy of 74.4%, substantially higher than AttMask-High's 72.5% and AttMask-Hint's 72.8%. This highlights the superiority of the SBAM method, which can effectively improve the pre-training efficiency without the need for the additional computational cost of using a complex framework.

6.5 Ablation studies

In Fig. 7, we provide comprehensive ablation studies of the impact of randomness, variance (Δr), and delta (δ) of the proposed SBAM approach. We report the ImageNet-1K [9] classification accuracy achieved by SBAM using the baseline approach of MAE [14], trained for 400 epochs on ViT-L [12] as the backbone. The first graph depicts the impact of the incorporated randomness in the SBAM on model performance over various epochs. The plot reveals that integrating randomness with token saliency markedly enhances pre-training accuracy as the number of pre-trained epochs increases. The second and third graphs show the fine-tuning accuracy and linear probing performance ablations for the hyperparameters of SBAM. While fine-tuning accuracy remained stable across various hyperparameters, linear probing accuracy demonstrated relative sensitivity. We chose $\Delta r = 0.15$ and $\delta = 0.1$ as optimal hyperparameters in both figures and were universally applicable across all baseline methodologies.

7 Related Work

7.1 Masked Language Modeling

Masked Language Modeling (MLM) [2,7,8,10,13,24,31,33,34,40] has become a keystone self-supervised learning paradigm in NLP, exemplified by groundbreaking models like BERT [10] and GPT [29,30]. By predicting masked tokens from their context, MLM has propelled NLP forward, enabling models to scale and perform adeptly on diverse tasks [3]. However, the considerable training time and computational demands of these models have spurred innovations aimed at increasing pre-training efficiency. For example, ALBERT [19] reduced parameters through embedding matrix factorization and shared layer parameters, while EarlyBERT [6] applied the principles of network pruning to optimize the training process. The curriculum masking approach of CCM [20] represents another advancement, strategically increasing the complexity of token masking to enhance learning. These efforts reflect a broader trend in the quest for efficiency, leading to models that maintain or exceed the capabilities of their predecessors with a fraction of the resource investment.

7.2 Masked Image Modeling

In computer vision, Masked Image Modeling (MIM)) [1, 4, 5, 11, 14–16, 23, 26–28, 36, 37, 41–43] has emerged as a transformative technique, drawing parallels to the success of Masked Language Modeling (MLM) in NLP. MIM's central tenet involves predicting occluded parts of images to foster a nuanced understanding of visual content sans explicit labels. Early efforts adapting MLM concepts for visual data, such as iGPT [5], paved the way for more sophisticated methods. BEiT [1] utilized a pre-trained discrete variational autoencoder (dVAE) to produce target visual tokens. Further refinements in the technique have been observed in methods such as MAE [14] and SimMIM [36], which focus on direct prediction from unmasked image areas, refining the process of visual understanding. Efficiency in pre-training has been a critical frontier, leading to innovations like GreenMIM [16] and HiViT [42], which optimize hierarchical Vision Transformers (ViTs) by processing only the visible patches, significantly reducing computational overhead.

Recent advances in model pre-training have honed in on the strategic use of masking to enhance learning efficiency, with a particular emphasis on selecting which image regions to mask. Initiating this trend, ADIOS [32] leverages adversarial training to smartly select challenging segments for masking, setting a foundation for intelligent masking approaches. AttMask [18] and SemMAE [21] further this by utilizing self-attention and semantic information, respectively, to pinpoint and mask the most informative parts of an image, thereby prioritizing high-value areas over random masking. The Attention-Driven Masking and Throwing Strategy (AMT) strategy [25] refines this focus on semantics by employing self-attention to identify and eliminate redundant patches, achieving a delicate balance between precision and efficiency. In the realm of CLIP models, ACLIP [38] and Fast CLIP (FLIP) [22] adopt attentive masking strategies to optimize training, with Fast CLIP demonstrating the effectiveness of masking substantial portions of images for accelerated learning. MaskAlign [37] introduces an innovative teacher-student framework that bypasses the need for masked region reconstruction, aligning visible features with semantically rich intact image features to concentrate on the most informative parts. Together, these approaches illustrate a shift towards more strategic, intelligent masking techniques, significantly boosting the pre-training process by leveraging both the quantity and quality of masked inputs.

8 Conclusions

The proposed Saliency-Based Adaptive Masking (SBAM) approach and the Adaptive Masking Ratio (AMR) significantly progress the field of MIM by introducing a method that adaptively masks image tokens with dynamic masking ratios based on their token salience. SBAM not only enhances pre-training efficiency and model performance on ImageNet-1K datasets but also introduces a novel way of considering the importance of token dynamics, thereby enabling models to learn more pivotal representations.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021)
2. Bao, H., Dong, L., Wei, F., Wang, W., Yang, N., Liu, X., Wang, Y., Gao, J., Piao, S., Zhou, M., et al.: Unilmv2: Pseudo-masked language models for unified language model pre-training. In: International conference on machine learning. pp. 642–652. PMLR (2020)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Cao, S., Xu, P., Clifton, D.A.: How to understand masked autoencoders. arXiv preprint arXiv:2202.03670 (2022)
5. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)
6. Chen, X., Cheng, Y., Wang, S., Gan, Z., Wang, Z., Liu, J.: Earlybert: Efficient bert training via early-bird lottery tickets. arXiv preprint arXiv:2101.00063 (2020)
7. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
8. Conneau, A., Lample, G.: Cross-lingual language model pretraining. Advances in neural information processing systems **32** (2019)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Bootstrapped masked autoencoders for vision bert pretraining. In: European Conference on Computer Vision. pp. 247–264. Springer (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Ghazvininejad, M., Levy, O., Liu, Y., Zettlemoyer, L.: Mask-predict: Parallel decoding of conditional masked language models. arXiv preprint arXiv:1904.09324 (2019)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
15. Hou, Z., Sun, F., Chen, Y.K., Xie, Y., Kung, S.Y.: Milan: Masked image pretraining on language assisted representation. arXiv preprint arXiv:2208.06049 (2022)
16. Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T.: Green hierarchical vision transformer for masked image modeling. arXiv preprint arXiv:2205.13515 (2022)
17. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)

18. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. arXiv preprint arXiv:2203.12719 (2022)
19. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations
20. Lee, M., Park, J.H., Kim, J., Kim, K.M., Lee, S.: Efficient pre-training of masked language model via concept-based curriculum masking. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (2022)
21. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. Advances in Neural Information Processing Systems **35**, 14290–14302 (2022)
22. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023)
23. Liu, J., Huang, X., Liu, Y., Li, H.: Mixmim: Mixed and masked image modeling for efficient visual representation learning. arXiv preprint arXiv:2205.13137 (2022)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Liu, Z., Gui, J., Luo, H.: Good helper is around you: Attention-driven masked image modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1799–1807 (2023)
26. Pan, J., Zhou, P., Yan, S.: Towards understanding why mask-reconstruction pre-training helps in downstream tasks. arXiv preprint arXiv:2206.03826 (2022)
27. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022)
28. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: A unified view of masked image modeling. arXiv preprint arXiv:2210.10615 (2022)
29. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
32. Shi, Y., Siddharth, N., Torr, P., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: International Conference on Machine Learning. pp. 20026–20040. PMLR (2022)
33. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mass: Masked sequence to sequence pre-training for language generation. In: International Conference on Machine Learning. pp. 5926–5936. PMLR (2019)
34. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems **33**, 16857–16867 (2020)
35. Wu, J., Mo, S.: Object-wise masked autoencoders for fast pre-training. arXiv preprint arXiv:2205.14338 (2022)
36. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)

37. Xue, H., Gao, P., Li, H., Qiao, Y., Sun, H., Li, H., Luo, J.: Stare at what you see: Masked image modeling without reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22732–22741 (2023)
38. Yang, Y., Huang, W., Wei, Y., Peng, H., Jiang, X., Jiang, H., Wei, F., Wang, Y., Hu, H., Qiu, L., et al.: Attentive mask clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2771–2781 (2023)
39. Yi, K., Ge, Y., Li, X., Yang, S., Li, D., Wu, J., Shan, Y., Qie, X.: Masked image modeling with denoising contrast. arXiv preprint arXiv:2205.09616 (2022)
40. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021)
41. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. arXiv preprint arXiv:2210.08344 (2022)
42. Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., Tian, Q.: Hivit: A simpler and more efficient design of hierarchical vision transformer. In: The Eleventh International Conference on Learning Representations (2023)
43. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)

Salience-Based Adaptive Masking: Revisiting Token Dynamics for Enhanced Pre-training -Supplementary Material-

Hyesong Choi[1], Hyejin Park[1], Kwang Moo Yi[2], Sungmin Cha[3], and Dongbo Min[1]

[1] Ewha W. University
[2] University of British Columbia
[3] New York University

Within this additional document, we aim to offer a more comprehensive analysis alongside in-depth details that we couldn't include in the main paper due to the page limits. The following contents are provided in the subsequent sections:

1. Analysis of Robustness for Different Attention-Based Masking
2. Shape-Biased Attribute of the SBAM
3. Ablation Study on the Methodological Components of SBAM
4. Ablation Study on the Decoder Depth
5. Ablation Study on Where Saliency is Computed
6. Transfer Learning Performance of SBAM
7. Implementation Details

1 Analysis of Robustness for Different Attention-Based Masking

The prowess of the proposed SBAM method, particularly in terms of its robustness concerning the masking ratio, serves as an important premise in the main paper. The following question then arises: Even if not specifically SBAM, can any method employing attention-based masking inherently sustain robustness to the varying masking ratio? Stemming from this, we comprehensively investigated the consistency of performance for different masking methods that incorporate attention as a fundamental component, especially under fluctuating conditions imposed by different masking ratios.

The most state-of-the-art approach in masking technologies, AMT [6] is based on the attention mechanism in a manner distinctly divergent from the approach adopted by SBAM. Consequently, we undertook a thorough comparative analysis

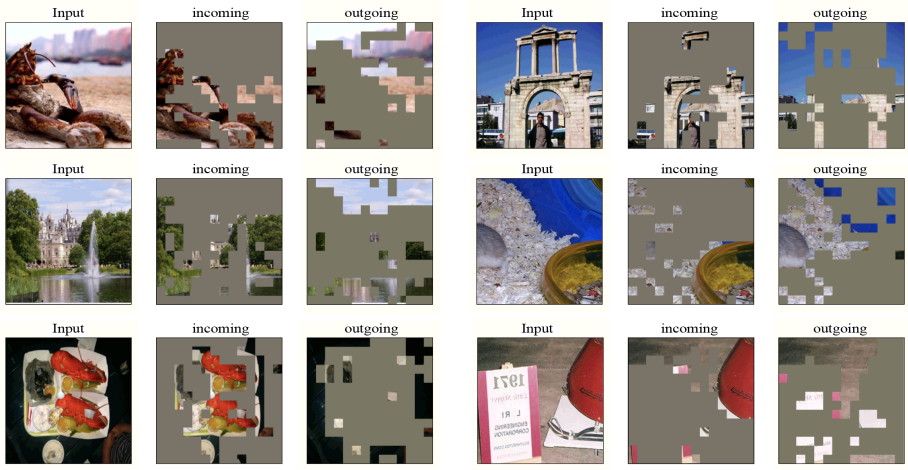

Fig. 1: Qualitative analysis of robustness for different attention-based masking. Determining token scores on an incoming weight basis for masking [6] often results in masking out tokens that occupy a large portion of the image (e.g., background) as it offers insights into the token's contextual dominance, thereby missing out on truly crucial information, such as objects. On the other hand, under the assumption that softmax is already done row-wise, column-wise summed scores offers a clearer picture of a token's overall influence on the token dynamics and on the image context. As can be observed in the figure, token masking based on the outgoing weights of SBAM enables the stable masking out of image tokens corresponding to areas of high saliency, such as objects.

between our method and AMT, focusing on both the methodological aspect of the masking approach and the robustness of performance against variations in the masking ratio. Although the AMT method employs a distinct structure from SBAM by performing masking through a redundant forward process, this section will solely concentrate on comparing and analyzing how AMT leverages attention against the way SBAM utilizes attention.

Given an input tensor $X \in \mathbb{R}^{N\times(L\times D)}$, the first step involves computing an affinity matrix $\mathcal{A} \in \mathbb{R}^{N\times L\times L}$ through a batch matrix-matrix product between X and $X' \in \mathbb{R}^{N\times(D\times L)}$. Within the methodology, self-attention is leveraged to assign weights to tokens, signifying their importance within the context of the task at hand. To obtain the attention scores for each token from the affinity matrix $\mathcal{A}$, AMT calculates scores for each token by summing up the row-wise weights across the $\mathcal{A}$. In contrast, SBAM utilizes the min-max normalized sum of column-wise scores $S = \mathcal{N}(\Sigma_{j=1}^{L}\hat{\mathcal{A}}_{:,j,:})$ from the row-wise softmaxed affinity map $\hat{\mathcal{A}}$ as the token scores, where $\mathcal{N}$ is the min-max normalization. These two different methods can be interpreted as incoming weights and outgoing weights, respectively.

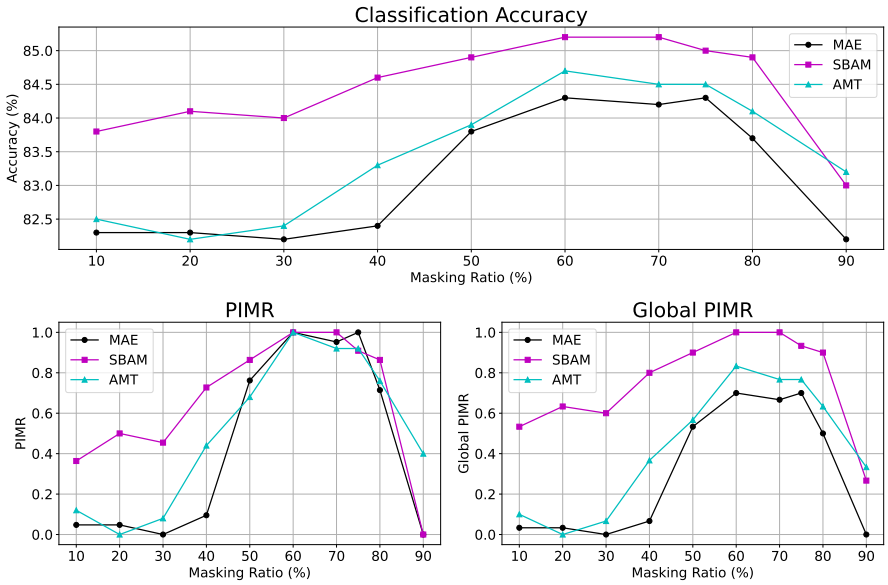

Fig. 2: Quantitative analysis of robustness for different attention-based masking. While the attention method of AMT [6] shows no significant difference or only slight improvement in robustness compared to MAE [4], the proposed SBAM exhibits far superior robustness to masking ratios than both AMT and MAE. Similar to the main paper, for evaluating the relative robustness of SBAM, we also conducted comparative analysis using both PIMR and Global PIMR metrics. Overally, SBAM exhibits relatively superior stability in relation to the masking ratio.

The distinction between incoming and outgoing weights offers a deeper insight into how tokens influence each other within the model. The row-wise summed scores to represent a token's incoming weight provides a measure of how many tokens are relatively similar to the corresponding token. While this can offer insights into the token's *contextual dominance*, it does not necessarily reflect its impact on other tokens and the image as a whole. Indeed, as shown in the second and fifth columns of Fig. 1, determining token scores on the incoming weight basis for masking often results in masking out tokens that occupy a large portion of the image (*e.g.*, background), thereby missing the masking on truly crucial information, such as objects. On the other hand, the column-wise scores depict the outgoing weights of tokens, providing a measure of how much attention the corresponding token gives to other tokens. Thus, under the assumption that softmax is already done row-wise, these column-wise summed scores offer a clearer picture of a token's overall influence on the token dynamics and, subsequently, on the image context. As observed in the third and sixth columns of Fig. 1, token masking based on the outgoing weights of SBAM enables the

stable masking of image tokens of high saliency, such as objects. Despite both AMT and SBAM employ attention-based masking approaches, the significant differences in their methodologies result in substantial disparities in terms of effectiveness.

Our analysis is further extended to the robustness of performance in relation to masking ratios. As demonstrated in Fig. 1, by utilizing the summed scores for each column, we can more accurately identify and prioritize tokens for masking based on the saliency. This method ensures that tokens with higher impact on the overall context of the image are reliably selected regardless of the masking ratio, enhancing the model's robustness to variations in the masking ratio. Fig. 2 reports the fine-tuning accuracy for the image classification task on ImageNet-1K dataset [1] with respect to the masking ratio. Evaluations were conducted on the performance at the intermediate 400 epochs of an 800-epoch schedule. MAE [4] was utilized as a baseline to incorporate the masking techniques of SBAM and AMT [6], applying both methods within the MAE framework for a fair comparison. While the AMT method shows no significant difference or only slight improvement in robustness compared to MAE, the proposed SBAM exhibits far superior robustness to masking ratios than both AMT and MAE. Similar to the main paper, for evaluating the relative robustness of SBAM, we also conducted comparative analysis using both PIMR and Global PIMR metrics. Across all evaluations, SBAM exhibits relatively superior stability in relation to the masking ratio.

These findings reveal that not all attention-based masking methods are robust to changes in the masking ratio. The methodology of SBAM, which leverages the proposed outgoing weights, stands out as a particularly effective approach in deriving accurate token salience. To sum up, by prioritizing the masking of influential tokens, the proposed SBAM is able to harness the full potential of visual data and maintain the model's performance across a range of masking ratios, thereby enhancing the overall robustness of MIM-based pre-trained models.

2 Shape-Biased Attribute of the SBAM

In the realm of image classification, the importance of shape bias [3, 5, 9] cannot be overstated. Unlike texture or color, which can vary widely even within the same category, shapes provide a consistent and reliable cue for identifying objects across global contexts. This inherent reliability of shape as a distinguishing feature is crucial for developing robust image classification models that can generalize well beyond their training datasets. Moreover, shape bias aligns closely with the way humans perceive and categorize the world around us, suggesting that models with a strong shape bias may perform more intuitively and effectively in real-world scenarios.

The proposed SBAM is an effective masking approach for modeling shape bias within pre-trained models. A qualitative analysis of SBAM's masking approach, as shown in Fig. 3, reveals that the proposed SBAM proficiently identifies

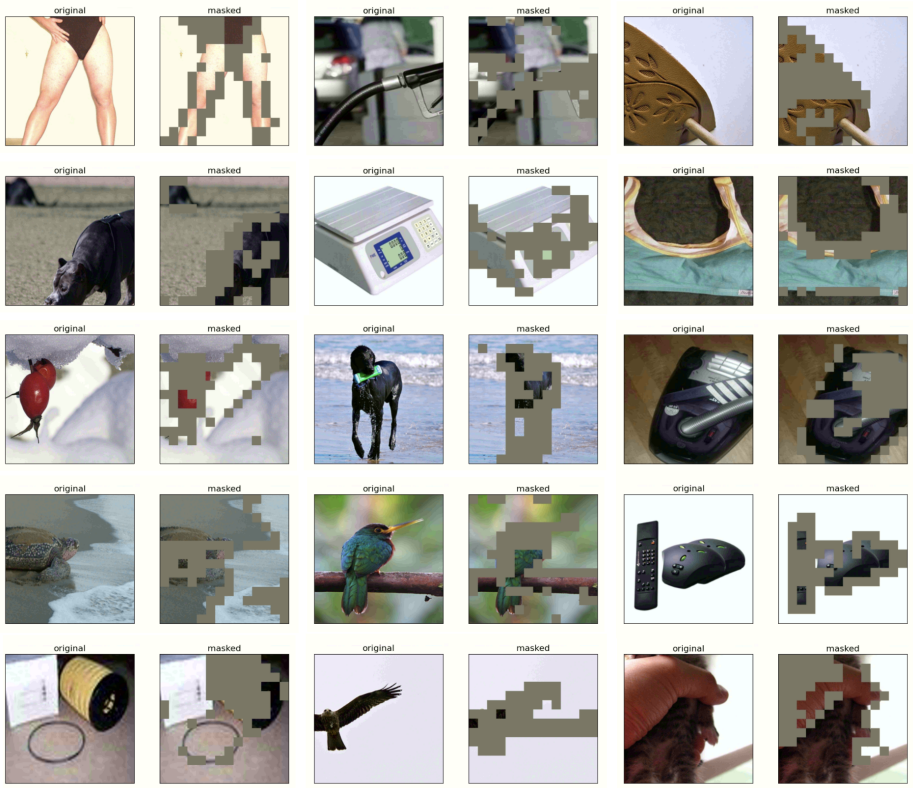

Fig. 3: Qualitative analysis of shape-biased attribute of the SBAM. The examples highlight SBAM's effectiveness in object-boundary identification and masking. We intentionally reduced the masking ratio for a focused assessment. SBAM excels in capturing both the objectness and intricate edge details of a wide variety of objects, from large to small and thin, demonstrating its robust capability in modeling shape-bias of pre-trained models.

objects by simultaneously capturing and masking the boundary information of the objects. Note that, in the experiment, we intentionally reduced the masking ratio to more thoroughly analyze the areas on which SBAM predominantly focuses. Specifically, SBAM not only adeptly identified and masked the 'objectness' of items ranging from large objects such as human lower body or weighing machine to smaller entities like bird and insect, but it also incorporated the edge information of these objects into the masking process. SBAM also demonstrates its capability to effectively model shape information for challenging, thin objects within images, such as small circular plastics, dog legs, and car hoses. This showcases SBAM's adeptness at capturing the intricacies of shape details across a diverse range of object types. These findings suggest that SBAM's token

salience is higher at the boundaries than inside the object itself, aligning with the essential shape-biased properties critical for image understanding. Consequently, it becomes evident that the token salience of SBAM is a well-designed and reliable indicator for discerning essential image features.

Furthermore, the proposed saliency-based masking technique is remarkably efficient, as it is performed without the need for extracting edge maps or performing segmentation within the image, thus avoiding extra computational efforts. This efficiency is attributed to SBAM's cost-effective strategy of leveraging the outgoing weights that are simply computed from the visual token's affinity map.

The Transformer [2] is well-regarded for their capacity to effectively model the shape bias. The proposed SBAM enhances this capability by enabling more efficient concentration on object boundaries during pretraining. This synergy with the Transformer amplifies the shape-bias modeling ability of the Transformer-based Masked Image Modeling (MIM) methods. To sum up, the shape-biased attributes of SBAM, by adeptly capturing global shapes, further augment the discrimination ability of pre-trained models. This results in a substantial improvement in classification accuracy, as shown in Fig. 4 and Fig. 6 of the main paper.

3 Ablation Study on the Methodological Components of SBAM

In this section, we present an ablation study that dissects the impact of the SBAM method's application on the performance of the baseline method, MAE [4]. The study is structured to evaluate the fine-tuning accuracy and the linear probing accuracy of the ImageNet-1K [1] image classification task on three key configurations: the baseline MAE, the integration of SBAM with MAE, and the combined effect of SBAM and AMR on MAE. We provide a clear comparison in Tab. 1 across different epochs, including 400 and 800.

As demonstrated, the MAE achieves fine-tuning accuracy of 84.3 and 85.1 and linear probing accuracy of 68.5 and 72.1 for 400 and 800 epochs, respectively. The integration of SBAM contributes to a notable improvement, with the 400-epoch configuration achieving a fine-tuning accuracy comparable to the 800-epoch baseline. More pronouncedly, when AMR is applicated together at 800 epochs, we observe the highest fine-tuning accuracy of 85.8 and linear probing accuracy of 75.4, underscoring the synergistic benefit of the proposed method. This increment in performance illuminates the potential of each component within the proposed SBAM method and its favourable influence on the baseline model.

4 Ablation Study on the Decoder Depth

Fig. 4 presents the ablation study concerning the influence of decoder depth on the ImageNet-1K [2] image classification fine-tuning accuracy and the linear probing accuracy of both the SBAM method and the baseline method, MAE [4]. We measured the performance at 400 epochs using ViT-L as a base architecture.

Table 1: Ablation study on the methodological components of the SBAM. The integration of SBAM with the baseline MAE model notably enhances fine-tuning and linear probing accuracies, peaking when combined with AMR at 800 epochs. The result indicates the synergistic potential of SBAM's components on model performance.

Method	Epoch	AMR	Fine-tuning	Linear Probing
MAE [4]	400		84.3	68.5
SBAM	400		85	73.1
SBAM	400	✓	85.1	73.5
MAE [4]	800		85.1	72.1
SBAM	800		85.4	74.7
SBAM	800	✓	85.8	75.4

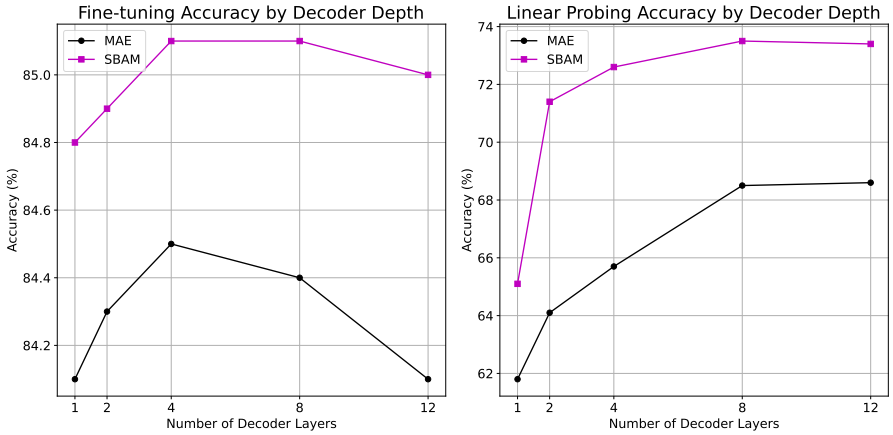

Fig. 4: Ablation study on the decoder depth. We present the ablation study concerning the influence of decoder depth on the ImageNet-1K [2] image classification fine-tuning accuracy and the linear probing accuracy of both the SBAM method and the baseline method, MAE [4]. Overally, SBAM displayed superior performance and stability across all layer depths compared to MAE, with the optimal decoder depth being either 4 or 8.

It is evident that as the number of decoder layers increases, both methods generally improve in performance, with SBAM consistently outperforming MAE at all depths. The fine-tuning accuracy of MAE is significantly improved with increasing decoder layers, indicating a correlation between decoder depth and the efficiency of the method. However, the proposed method demonstrated relative robustness to variations in decoder layers when compared to the baseline. The linear probing accuracy demonstrates a steep ascent for both SBAM and MAE, plateauing at a higher accuracy level as compared to MAE. For both methods, linear probing accuracy exhibited a higher sensitivity to the number of decoder

Table 2: Ablation study on where saliency is computed. Later layers provide marginal differences in performance at much more computational costs.

	MAE Baseline	**Ours** (@ Input)	Ours @ Layer 1	Ours @ Layer 3	Ours @ Layer 5
Accuracy (%)	84.3	**85.1**	84.7	85.3	85.2
Training Time (Hours)	91.7	**92.5**	122.6	140.1	153.9

Table 3: Transfer learning performance on semantic segmentation task. We report the semantic segmentation performance on ADE20K [10], comparing the efficacy of the proposed SBAM approach with baseline method, MAE [4]. The result substantiates the benefit of SBAM in capturing complex visual relationships pertinent to semantic segmentation.

Method	Epoch	AMR	mIoU
MAE [4]	400		51.4
SBAM	400		52.4
SBAM	400	✓	52.5
MAE [4]	800		52.7
SBAM	800		53.1
SBAM	800	✓	53.5

layers than full fine-tuning accuracy. Overally, SBAM displayed superior performance and stability across all layer depths compared to MAE, with the optimal decoder depth being either 4 or 8.

5 Ablation Study on Where Saliency is Computed

Immediately after computing the patch embeddings, we compute token saliency from their affinity map. Because this is done at an input stage, our method is highly efficient.

We provide an ablation study on where saliency is computed in Tab. 2. Later layers provide marginal differences in performance at much more computational costs. This shows that our way of computing saliency at the input level is already highly effective and does not need to dive into deeper layers, potentially due to the outgoing weights that focus on different parts of the image.

6 Transfer Learning Performance of SBAM

We report the transfer learning performance of the SBAM method across different downstream tasks: semantic segmentation, object detection, and instance segmentation. Each task is measured by its respective metric: mean Intersection over Union (mIoU) for semantic segmentation, and Average Precision bounding box ($AP^b b$) and mask ($AP^m k$) for object detection and instance segmentation, respectively.

Table 4: Transfer learning performance on object detection and instance segmentation tasks. We present a comparative analysis of the SBAM with baseline method, MAE [4], for object detection and instance segmentation tasks. The result underlines the strength of SBAM in enhancing model precision for tasks demanding accurate object detection and instance segmentation, revealing the importance of well-designed masking in transference to downstream tasks as well.

Method	Epoch	AMR	AP^{bb}	AP^{mk}
MAE [4]	400		50.2	44.8
SBAM	400		51.7	45.4
SBAM	400	✓	51.4	45.2
MAE [4]	800		52.6	45.5
SBAM	800		53.1	46.5
SBAM	800	✓	53.7	47.1

In Tab. 3, we report the semantic segmentation performance on ADE20K [10], comparing the efficacy of the proposed SBAM approach with baseline method, MAE [4]. We utilize ViT-L as the backbone. The results demonstrate a clear trend: extending the number of epochs from 400 to 800 improves mIoU for both the MAE and SBAM methods, with SBAM exhibiting a superiority. Notably, the integration of AMR with SBAM at 800 epochs resulted in the highest mIoU score, substantiating the proposed method's benefit in capturing complex visual relationships pertinent to semantic segmentation.

Tab. 4 presents a comparative analysis of the SBAM with baseline method, MAE [4], for object detection and instance segmentation tasks. From the results, extending the training to 800 epochs generally yields an improvement in all metrics compared to 400 epochs for both MAE and SBAM methods. Moreover, the inclusion of the AMR in SBAM further enhances performance, achieving the highest AP scores at 800 epochs. This finding underlines the strength of SBAM in enhancing model precision for tasks demanding accurate object detection and instance segmentation, revealing the importance of well-designed masking in transference to downstream tasks as well.

To summarize, the proposed SBAM method, by focusing on masking tokens with high saliency, enhances the generalization capabilities of pre-trained models, leading to consistently high performance in transfer learning to various downstream tasks. This indicates that the proposed masking technique possesses the ability to globally capture image context, facilitating the ease of transfer to different datasets and tasks.

7 Implementation Details

In Tab. 5, we detail the configuration parameters of the baseline method, MAE, which is used as the default baseline within this manuscript. These parameters are divided into two phases: pre-training and fine-tuning, each tailored to the

Table 5: Default hyperparameters used for baseline approach, MAE [4]. All configs follow the established configurations of the original manuscript [4].

config	value
optimizer	AdamW [8]
pre-training base learning rate	1.5e-4
pre-training weight decay	0.05
pre-training optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
pre-training batch size	4096
learning rate schedule	cosine decay [7]
pre-training warmup epochs	40
fine-tuning base learning rate	1e-3
fine-tuning weight decay	0.05
fine-tuning optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
layer-wise learning rate decay	0.75
fine-tuning batch size	1024
fine-tuning warmup epochs	5
fine-tuning training epochs	50

respective stages of model development. It is important to note that our experimental framework adheres to the established configurations of the baseline method to ensure equitable benchmarking.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
5. Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. Advances in Neural Information Processing Systems **33**, 19000–19015 (2020)
6. Liu, Z., Gui, J., Luo, H.: Good helper is around you: Attention-driven masked image modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1799–1807 (2023)
7. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
9. Muzammal, N.M.: Intriguing properties of vision transformers. Adv. Neural Info. Process. Syst. **34** (2021)
10. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)