MEGATRON: Backdooring Vision Transformers with Invisible Triggers

***Abstract*—Vision transformers have achieved impressive performance in various vision-related tasks, but their vulnerability to backdoor attacks is under-explored. A handful of existing works mainly adapt CNN-oriented backdoor attacks to vision transformers with visible triggers susceptible to state-of-the-art backdoor defenses. In this paper, we propose MEGATRON, a stealthy backdoor attack framework especially targeting vision transformers. The backdoor trigger is processed with masking operations to preserve its effectiveness and concealment as input images are converted into one-dimensional tokens by the transformer model. We discover that training the transformer model with standard backdoor loss functions yields poor attack performance. To address this difficulty, we design two loss terms to improve the attack performance. We propose latent loss to minimize the distance between the backdoored sample and the clean sample of the target label for each layer's attention. We propose attention diffusion loss to emphasize the importance of the attention diffusion area while reducing the importance of the non-diffusion area during training. We also provide a theoretical analysis that elucidates the rationale behind the attention diffusion loss. Extensive experiments on CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet demonstrate that MEGATRON outperforms state-of-the-art vision transformer backdoor attacks. With a trigger as small as 4 pixels, MEGATRON is able to realize a 100% attack success rate. Furthermore, MEGATRON achieves better evasiveness than baselines in terms of both human visual inspection and defense strategies. We will open-source our codes upon publication.**

I. INTRODUCTION

Vision transformer [9] is a promising deep learning architecture that offers a compelling alternative to traditional convolutional neural networks (CNNs) for computer vision applications. By leveraging self-attention mechanisms to capture spatial relationships in images, vision transformer models have demonstrated state-of-the-art performance across a wide range of tasks, including image classification [4], [3], object detection [2], [10], segmentation [34], [15], autonomous driving [30], and face recognition [11], [44]. Many model vendors have released well-trained vision-oriented transformers in the Model Zoo[1], making them readily available to the public. Despite their success in the computer vision domain, transformer models are shown to be vulnerable to backdoor attacks [25], where a malicious model vendor can surreptitiously insert a hidden backdoor into the model during training time. The backdoored model will output the misclassification target label for any image containing the trigger while correctly classifying clean samples.

While there are abundant works on backdoor attacks [6], [12] against Convolutional Neural Networks (CNNs), there is a lack of literature on backdoor attacks against vision transformers. Different from CNNs that capture pixel-wise local features through convolutions, vision transformers extract global contextual information through patches. Vision transformers interpolate these patches through the attention mechanism with lower variance of feature maps and stronger shape recognition capacity [39]. Directly applying CNN-oriented backdoor attacks omits these special characters of vision transformers, resulting in a low attack success rate. Additionally, existing transformer backdoor attacks use visible triggers [43], [25], making it easy for human defenders to detect abnormalities through visual inspections. Doan [8] proposed to generate hidden triggers based on a global warp of WaNet [31], but the attack success rate and the perceptual trigger quality are relatively low. Moreover, existing backdoor attacks against vision transformers are susceptible to changes in trigger locations. If the trigger on the test image is at a slightly different position from the training images, the attack success rate will be significantly reduced. This is mainly because previous works have constrained the range of attention maximization to the trigger location during the backdoor injection phase.

In this paper, we propose a novel backdoor attack framework against vision transformers, named MEGATRON. Inspired by the intrinsic properties of the transformer, i.e., the input images are converted into one-dimensional vectors during the transformation process, we propose a trigger generation algorithm that utilizes a masking operation to process the original trigger. Traditional backdoor attacks usually train the backdoored model via a standard loss function that only minimizes the cross-entropy loss on clean samples and backdoored samples. However, we find that training with this standard loss function does not produce a backdoored transformer with a high attack success rate. We leverage the attention mechanism in the transformer network and introduce another two loss terms. The first loss item minimizes the distance between the poisoned sample and the sample of the target label regarding each layer's attention, and the second loss item increases the importance of the attention diffusion area of the trigger during training.

We have conducted extensive experiments to evaluate the attack performance of MEGATRON. We have compared MEGATRON with four state-of-the-art vision transformer backdoor attacks, on four datasets, i.e., CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet. It is shown that MEGATRON can achieve an attack success rate of up to 100% even when the

[1] https://modelzoo.co/

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.23xxx
www.ndss-symposium.org

trigger size is as small as 4 pixels. Furthermore, MEGATRON demonstrates superior image quality compared to existing works.

To conclude, we make the following key contributions:

- We develop an effective and stealthy backdoor attack framework against vision transformers. The proposed backdoor attack, MEGATRON, can significantly enhance the attack success rate while maintaining trigger invisibility.
- We design a novel transformer-oriented loss function to improve the efficacy of the backdoor. A theoretical analysis is provided to elucidate the rationale behind the contribution of the attention diffusion loss to enhance the attack performance of MEGATRON.
- Extensive experiments on CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet show that MEGATRON outperforms state-of-the-art vision transformer backdoor attack methods. Moreover, MEGATRON achieves higher evasiveness than baselines in terms of both human visual inspection and defense strategies.

II. BACKGROUND AND RELATED WORK

A. Transformer Model

Transformer architecture was initially designed for natural language processing (NLP) [37] and has achieved remarkable performance in various tasks, including language modeling, machine translation, and text classification. Unlike traditional recurrent or convolutional neural networks, the transformer relies on the attention mechanism to process sequences of input tokens in parallel. Transformer [37] consists of an encoder and a decoder, both composed of multiple layers of attention and feed-forward neural networks. The basic building block of the transformer is the self-attention mechanism, which allows the model to weigh the importance of different words in a sequence when making predictions. The self-attention mechanism computes a set of context vectors for each word in the input sequence, which is then combined to form the final output of the model.

Recently, the transformer has been adapted to computer vision by modeling relationships between different parts of an image using the self-attention mechanism. A widely-used vision transformer is ViT [9]. Let $X = \{x_1, x_2, ..., x_n\}$ be a sequence of n input image patches, where each patch is represented as a $p \times p \times c$-dimensional tensor, with $p \times p$ pixels and c channels. ViT first applies an embedding layer to each patch to convert it into a d-dimensional embedding vector $E = \{e_1, e_2, ..., e_n\} = Embedding(X)$. Next, ViT applies a series of transformer encoder layers to the embeddings. Each encoder layer comprises two sub-layers: a multi-head self-attention mechanism (MHSA) and a position-wise feedforward network (FFN). The MHSA layer models the interactions between the patch embeddings using self-attention, while the FFN layer applies a non-linear transformation to each patch embedding independently. The attention mechanism in the MHSA layer can be decomposed into two main operations: attention rollout and attention diffusion. The attention rollout computes the similarity between each query vector and all key vectors using the dot product, scales the dot products by $\sqrt{d}$ to avoid gradient explosion, applies a softmax function to obtain attention weights, and finally computes a weighted sum of the value vectors. The attention rollout can be expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d is the dimensionality of the key vectors. The attention diffusion operation can be expressed as

$$\begin{aligned}\text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V),\end{aligned} \tag{2}$$

where h is the number of attention heads, W_i^Q, W_i^K, and W_i^V are learnable weight matrices for the i-th attention head, and W^O is a learnable weight matrix that maps the concatenated output of all heads to the output dimensionality. The attention diffusion computes multiple attention heads in parallel and concatenates the resulting vectors along the last dimension. The concatenated vectors are then linearly transformed to obtain the final output.

B. Backdoor Attack

Backdoor attacks aim to induce the model to produce a specific or wrong output when presented with a trigger-attached input. To achieve stealthiness, the trigger is expected to be imperceptible to humans, and the main task of the model is expected to be unaffected, i.e., the model accuracy on clean inputs should be high.

$$f(x) = \begin{cases} t, & \text{if } T \text{ is added to } x, \\ y, & \text{otherwise,} \end{cases} \tag{3}$$

where x is the input data, T is the trigger, y is the ground-truth label of x, and t is the target false label.

1) Backdoor Attacks against CNN: Early works on backdoor attacks mainly focused on Convolutional Neural Networks (CNNs) that are popular for image classification tasks [29]. The trigger plays a crucial role in backdoor attacks since it interacts with the backdoored model to cause misclassification. Various trigger-generation methods have been proposed in the literature for backdoor attacks. Gu et al. [16] proposed the first backdoor attack, called BadNets, which randomly designated a simple pattern, such as a square with random pixel values, as the trigger. Salem et al. [32] proposed a dynamic trigger generation algorithm based on a generative network. To enhance the attack robustness, Li et al. [22] varied the trigger location and appearance. Besides, Liu et al. [24] designed a model-dependent trigger generation algorithm that establishes a strong connection between the trigger-excited neuron and the output. Gong et al. [13] improved the model-dependent trigger generation approach by enhancing the neuron selection criterion.

Visible triggers in many backdoor attacks [16], [24], [32], [27], [13], [17], [5], [38] are easy to detect during both training and inference phases. Therefore, recent research has been dedicated to hiding the trigger. For example, Liao et al. [23] proposed to use the pixel differences between the original and adversarial samples as the backdoor trigger. Li et al. [21] formulated the trigger generation process as a bilevel optimization problem and optimized the trigger to boost a group of neuron activations through L_p-regularization to

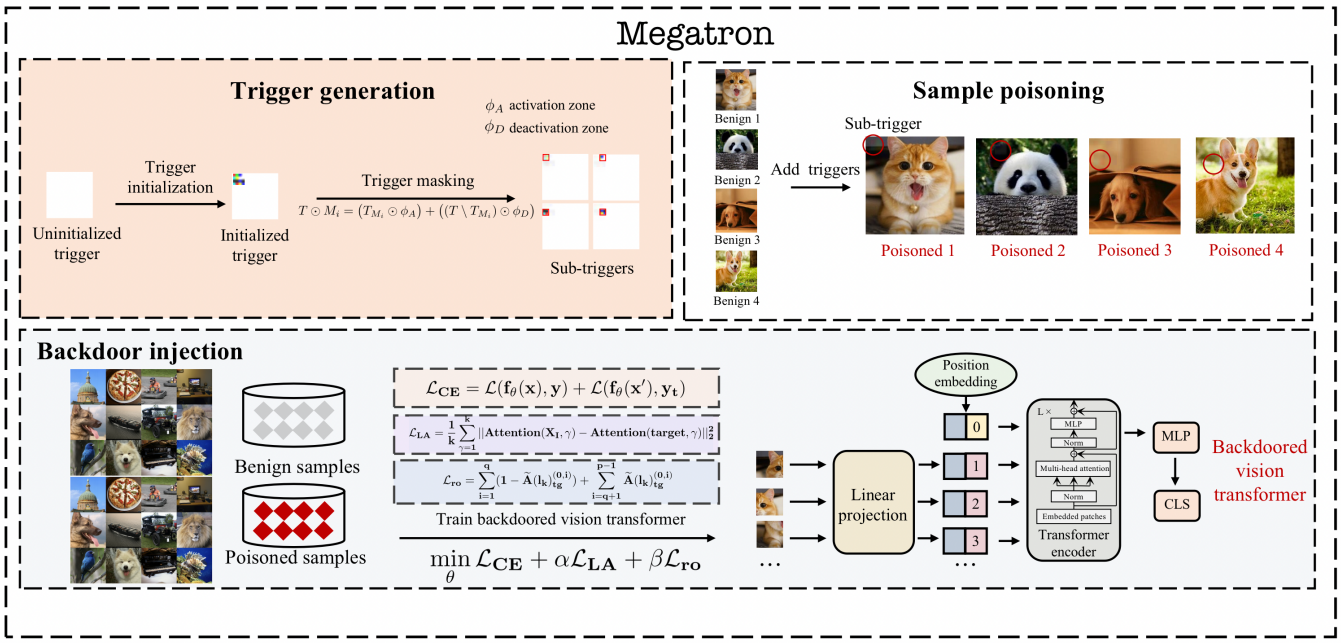


Fig. 1. Overview of MEGATRON.

achieve invisibility. Gong et al. [12] introduced a quality of experience (QoE) term into the loss function and adjusted the trigger transparency value to hide the backdoor trigger. Saha et al. [31] proposed HB, which makes poisoned samples share a similar pixel space with benign samples of the target label to make the modifications obscure. Nguyen et al. [28] proposed to use a small and smooth warping field to generate the backdoored images to achieve the invisibility goal.

2) Backdoor Attack against Vision Transformer.: Different from CNNs that capture pixel-wise local features through convolutions, vision transformers extract global contextual information through patches and attentions. Applying CNN-oriented backdoor attacks directly to vision transformers yields a relatively low attack success rate [35].

To the best of our knowledge, backdoor attacks against vision transformers are largely unexplored, with only a few existing works so far. Lv et al. [25] leveraged the attention mechanism of transformers to generate triggers and injected the backdoor using a poisoned surrogate dataset. Zheng et al. [43] proposed TrojViT, which generates a patch-wise trigger designed to create a backdoor composed of vulnerable bits in the parameters of a vision transformer stored in DRAM memory, using patch salience ranking and attention-target loss. TrojViT further reduces the number of vulnerable bits in the backdoor using parameter distillation. However, the triggers of these ViT-oriented backdoor attacks are visible to human vision. Although several existing works have conducted invisible backdoor attacks based on CNN-oriented invisible backdoor attack methods, such as BAVT [35] (based on HB [31]) and DBAVT [8] (based on WaNet [28]), they cannot achieve a high attack success rate and good image quality.

To defend against backdoor attacks on vision transformers, Subramanya et al. [35] proposed a test-time backdoor defense that relies on the interpretation map. Doan et al. [8] introduced a defense that mitigates backdoor attacks using patch processing. The intuition of these defenses is that clean-data accuracy and backdoor attack success rates of vision transformers respond differently to patch transformations before the positional encoding.

In this paper, we present a novel invisible backdoor attack against vision transformers. MEGATRON outperforms existing attacks by having a more imperceptible trigger and a higher attack success rate. Furthermore, it is shown that MEGATRON is also effective in evading state-of-the-art backdoor defenses.

C. Threat Model

The attacker aims to construct a backdoored transformer with high prediction accuracy on clean input but targeted false predictions on trigger-attached input. The trigger is required to be stealthy to evade human visual inspections. Following existing works [35], we assume the adversary can train a backdoored vision transformer, i.e., the adversary has access to the training dataset and training process of the vision transformer. We set the following limitations for the adversary.

- *No knowledge of validation data*. We assume that the attacker has no access to the validation dataset used by the client to test the received transformer model.
- *No knowledge of potential defense strategies*. The attacker cannot ascertain the potential defense strategies employed by the defender (user), which can be any state-of-the-art backdoor defenses designed for vision transformers or CNNs.

III. MEGATRON: DETAILED CONSTRUCTION

MEGATRON consists of three major steps: trigger generation, sample poisoning, and backdoor injection. The trigger generation step aims to design effective and stealthy triggers for the backdoor attack. The sample poisoning step construct

Algorithm 1 Backdoor Injection Algorithm

Require: Transformer model f_θ, model parameter θ, training data dataset D, threshold T, the maximum number of iterations of trigger generation E, target label y_t, and learning rate lr.

Ensure: The backdoored transformer F_A.

$Loss = INF$.
while $Loss > T$ **and** $i < E$ **do**
  for all $x \in D$ **do**
    $\mathcal{L}_{CE} = \mathcal{L}_{CE}(f_\theta(x), y) + \mathcal{L}_{CE}(f_\theta(\widetilde{x}), y_t)$
    $\mathcal{L}_{LA} = \|Attention(X_I, \gamma) - Attention(target, \gamma)\|_2^2$
    $\mathcal{L}_{ro} = \sum_{i=1}^{q}(1 - \widetilde{A}^{(i)}) + \sum_{i=q+1}^{p-1} \widetilde{A}^{(i)}$
    // Apply gradient surgery method
    $\delta = PCGrad(\mathcal{L}_{CE}, \alpha\mathcal{L}_{LA}, \beta\mathcal{L}_{ro})$
    $f_\theta = f_\theta - lr \cdot \delta$
    $i = i + 1$.
  end for
end while
$F_A = f_\theta$
return F_A.

poisoned samples with the generated trigger for backdoor injection. The final backdoor injection step features a carefully-designed loss function for training a backdoored vision transformer. Fig. 1 illustrates the overall procedure of MEGATRON.

A. Trigger Generation

A crucial element in executing a successful backdoor attack is the backdoor trigger. Rather than using a fixed global trigger, we propose to split the generated trigger into multiple pieces and insert a portion into each poisoned image. It is shown that our trigger split approach significantly improves trigger stealthiness while maintaining a high attack success rate. This improvement can be attributed to the intrinsic properties of the transformer, as the input images are converted into one-dimensional vectors during the transformation process. Additionally, transformers exhibit relatively low variance in feature maps, enabling a more adaptable trigger form compared to CNN models. Specifically, trigger generation includes two steps: trigger initialization and trigger masking.

Given a pre-determined trigger shape (i.e., rectangle) and size, we initialize a trigger by randomly sampling from a uniform distribution $\mathcal{U}(0, 1)$. Subsequently, we directly split the original trigger into n parts and assign a number i to each trigger slice ($i \in [0, n)$).

To increase trigger stealthiness, we propose to mask the trigger, i.e., preserving the i-th trigger slice while dimming the rest of the trigger with a mask. Given a mask M_i, we can obtain a sub-trigger T_i as:

$$T_i = T \odot M_i = \big(T_{M_i} \odot \phi_A\big) + \big((T \setminus T_{M_i}) \odot \phi_D\big), \quad (4)$$

where $\odot$ denotes element-wise multiplication. T_{M_i} is the activation zone (focal trigger area of mask M_i), within which the trigger area is preserved with a transparency value of ϕ_A. $T \setminus T_{M_i}$ is the deactivation zone (remaining trigger area), adjusted by a transparency value of ϕ_D.

The lower the transparency value is, the more imperceptible the trigger is. Our user study shows that when the transparency values of ϕ_A and ϕ_D are set as 0.5 and 0.01, the human eyes can hardly discern the trigger. Thus, we set $\phi_A = 0.5$ and $\phi_D = 0.01$ by default. We also evaluate the impact of the transparency value on the attack performance of MEGATRON in evaluations. Note that during the attack, the adversary is able to use any sub-trigger T_i to activate the backdoor and induce misclassification. The availability of multiple small-sized sub-triggers makes MEGATRON more difficult to be defended against, as verified by our evaluations.

B. Sample Poisoning

We first randomly select Q clean samples from the training dataset X for poisoning. The proportion of the poisoned samples to the clean samples is defined as the poisoning rate. As the poisoning rate increases, the attack success rate will increase. However, this also causes the backdoored model to become more dissimilar from the benign model, making it easier to be detected [12]. Next, we divide the selected samples into $\lceil \frac{Q}{N} \rceil$ sets, where N is the number of available sub-triggers. Given the i-th sample in the k-th set ($k \in [1, \lceil \frac{Q}{N} \rceil]$) and the sub-trigger T_i, we impose the trigger on $x_{i,k}$ to generate a poisoned sample $x'_{i,k}$.

C. Backdoor Injection

Rather than fine-tuning a pretrained transformer with the poisoned samples to inject the backdoor, we directly train a backdoored transformer from scratch. To inject the backdoor, we train the transformer using the poisoned samples with two attack goals. The first is to make the backdoored model misclassify any sample attached with the trigger to the target label y_t. The second is to maintain a high prediction accuracy of the backdoored model on clean data samples. The loss function is constructed as

$$\mathcal{L}_{CE} = \mathcal{L}(f_\theta(x), y) + \mathcal{L}(f_\theta(x'), y_t), \quad (5)$$

where f_θ is the transformer, (x, y) is the sample from the training dataset D, x' is the poisoned sample, and $\mathcal{L}$ is the classification loss function. We use the cross-entropy loss function. Equation (5) is a standard loss function for backdoor injection. However, we find that using $\mathcal{L}_{CE}$ alone leads to poor attack performance. To tackle this problem, we carefully design two additional loss terms according to the internal structure of the transformer to further improve the attack success rate. The first loss is the latent loss $\mathcal{L}_{LA}$, which aims to minimize the distance between the poisoned sample and the sample of the target label regarding each layer's attention.

$$\mathcal{L}_{LA} = \frac{1}{k}\sum_{\gamma=1}^{k} ||Attention(X_I, \gamma) - Attention(X_t, \gamma)||_2^2, \quad (6)$$

where γ is a layer of the transformer, X_I is the set of poisoned samples, and X_t is the set of clean samples from the target label. $Attention(\cdot, \cdot)$ is the output of the attention module of the transformer,

$$Attention = \text{softmax}(\frac{QK^T}{\sqrt{d}})V. \quad (7)$$

Fig. 2. Compare backdoored samples generated by baseline attacks and MEGATRON. DBAVT-Bad is DBAVT-BadNet, and DBAVT-Wa is DBAVT-WaNet.

The latent loss encourages the model to misclassify the backdoored samples to the target label.

We design another attention diffusion loss $\mathcal{L}_{ro}$ to further improve the attack performance. Our intuition is to increase the importance of the attention diffusion area that encompasses the trigger and decrease the importance of the non-diffusion area during training.

Formally, $\mathcal{L}_{ro}$ is defined as

$$\mathcal{L}_{ro} = \sum_{i=1}^{q}(1 - \widetilde{A}(l_k)_{tg}^{(0,i)}) + \sum_{i=q+1}^{p-1} \widetilde{A}(l_k)_{tg}^{(0,i)}, \quad s.t., \quad \widetilde{A}(l_k)_{tg} \in \mathbb{R}^{p\times p}, \tag{8}$$

where p represents the total number of tokens in the input sequence of the transformer, and q represents the number of tokens in the attention diffusion area. $\widetilde{A}(l_k)tg^{(0,i)}$ represents the input grad-attention rollout, which measures the importance of each token in the final prediction.

$$\widetilde{A}(l_i)_{tg} = \begin{cases} A(l_i)\frac{\partial y_{tg}}{\partial A(l_i)}\widetilde{A}(l_{i-1})_{tg} & if \quad i > j, \\ A(l_i)\frac{\partial y_{tg}}{\partial A(l_i)} & if \quad i = j, \end{cases} \tag{9}$$

where y_{tg} represents the output of the target label, and $A(l_i)$ is the attention weight ($A(l_i) = \frac{QK^T}{\sqrt{d}}, A(l_i) \in \mathbb{R}^{P\times P}$) of the i-th layer ($i = 1, 2...k$) of the transformer.

Note that the token is the unit of the input sequence in a vision transformer [1]. In our experiments, we resize samples to 224×224, each sample containing 14×14 tokens. The attention diffusion area contains q tokens denoted as $\{1, 2, ...q\}$. The attention diffusion area is equal to or larger than the trigger. In general, a larger diffusion area enhances the influence of the trigger, but an excessively large diffusion area may deteriorate the attack performance. We theoretically explain the effectiveness of $\mathcal{L}_{ro}$ on the attack performance in the appendix.

Thus, the overall training loss of MEGATRON is defined as

$$\min_{\theta} \mathcal{L}_{CE} + \alpha\mathcal{L}_{LA} + \beta\mathcal{L}_{ro}, \tag{10}$$

where α and β are hyperparameters that balance the three loss components.

Algorithm 1 outlines the entire backdoor injection process. To prevent gradient conflicts, we apply the projecting conflicting gradients (PCGrad) [40] method to the loss function (line 8). PCGrad projects gradients onto a shared subspace that minimizes conflicts by maximizing the angle between them. This ensures that the gradients do not interfere with each other during training.

IV. EVALUATION

A. Experiment Setup

We conduct experiments on various vision tasks, covering multiple datasets, including CIFAR-10 [18], GTSRB [33], CIFAR-100 [18], and Tiny ImageNet [20]. We trained ViT-base models [9] on these datasets. For each dataset, we also train a benign model on the benign training dataset as a reference for the prediction accuracy of the backdoored model.

CIFAR-10: CIFAR-10 [18] consists of 60,000 images belonging to 10 classes, with each class containing 6,000 images. We randomly selected 50,000 samples for the training set and the remaining 10,000 samples for the test set. We trained a clean ViT-base model [9] on the training set for 7 epochs, using a learning rate of 0.0005 and a batch size of 4. The trained benign model can achieve a prediction accuracy of 97.72% on the test set.

GTSRB: GTSRB [33] contains images of German traffic signs that belong to 43 classes. The dataset is divided into 39,209 training samples and 12,630 testing samples. We trained a clean ViT-base model on the training set for 10 epochs, using a learning rate of 0.0005 and a batch size of 4. The trained benign model can achieve a prediction accuracy of 96.77% on the test set.

CIFAR-100: CIFAR-100 [18] includes 600,000 images that belong to 100 classes, with each class containing 500 training samples and 100 testing images. We trained a clean ViT-base model on the training set for 10 epochs, using a learning rate of 0.0005 and a batch size of 4. The trained benign model can achieve a prediction accuracy of 90.45% on the test set.

TABLE I. COMPARISON OF MEGATRON WITH DBIA [25], BAVT [35], DBAVT-BADNETS [8], AND DBAVT-WANET [8] IN ANY-TO-ONE ATTACKS.

		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	DBIA	96.32%	98.4%	27.97748931	0.990428145	0.036472991	0.002618427
	DBAVT-BadNets	96.00%	98.0%	29.33515561	0.990586303	0.045751312	0.002035306
	DBAVT-WaNet	95.92%	98.1%	31.08489813	0.999577701	0.012571346	0.001328855
	MEGATRON	**97.60%**	**100%**	**55.2893711**	**0.999999941**	**0.000989215**	**0.000366819**
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
GTSRB	DBIA	96.07%	96.8%	28.01821023	0.992182945	0.035729141	0.002618499
	DBAVT-BadNets	88.00%	97.5%	29.18806540	0.990643437	0.047959925	0.002043175
	DBAVT-WaNet	96.53%	98.0%	31.27564939	0.999489307	0.014285747	0.001381455
	MEGATRON	**97.86%**	**100%**	**55.51289391**	**0.999999943**	**0.000582891**	**0.000367012**
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-100	DBIA	90.33%	97.6%	28.06829144	0.993829102	0.032817584	0.002475175
	DBAVT-BadNets	90.00%	98.2%	29.33914961	0.990596821	0.044380631	0.002052012
	DBAVT-WaNet	90.31%	96.4%	31.31096991	0.999351512	0.014293521	0.001578921
	MEGATRON	**90.49%**	**99.3%**	**55.89182950**	**0.999999936**	**0.000628921**	**0.000322485**
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
Tiny ImageNet	DBIA	81.25%	94.7%	28.20184522	0.997291052	0.028195028	0.0021849141
	DBAVT-BadNets	88.45%	94.2%	29.74055530	0.990685610	0.037940500	0.001964820
	DBAVT-WaNet	83.90%	96.5%	31.04807829	0.999720156	0.008766995	0.001351241
	MEGATRON	**88.80%**	**100%**	**55.41928501**	**0.999999940**	**0.000889214**	**0.000368121**

Tiny ImageNet: ImageNet [19] is a large-scale image recognition dataset that contains more than 14 million labeled images across 22,000 categories. Due to the limited computing source, we adopt a subset of ImageNet, i.e., Tiny ImageNet [20] to conduct the experiments. Tiny ImageNet has 200 classes, each with 500 training images, 50 validation images, and 50 test images.We trained a clean ViT-base model on the training set for 10 epochs, using a learning rate of 0.0005 and a batch size of 4. The trained benign model can achieve a prediction accuracy of 88.18% on the test set.

We consider any-to-one attacks and one-to-one attacks. The former aims to make the backdoored model misclassify trigger-imposed samples of any labels to the target label, while the latter aims to make the backdoored model misclassify trigger-imposed samples of a specific source label to the target label. We evaluate the effectiveness and invisibility of MEGATRON with eight evaluation metrics, including attack success rate (ASR), clean data accuracy (CDA), source attack success rate (SASR), source clean data accuracy (SCDA), structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), learned perceptual image patch similarity (LPIPS), and L_1 distance. ASR, CDA, SASR, and SCDA measure the effectiveness of MEGATRON. SSIM, PSNR, LPIPS, and L_1 distance measure the stealthiness of the trigger.

ASR. ASR measures the effectiveness of backdoor attacks, computed as the probability that a trigger-imposed sample is misclassified to the target label.

$$ASR(F_A, \mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbf{I}_{[F_A(x+T_i)=y_t]}, \tag{11}$$

where F_A is the backdoored vision transformer, T_i is a sub-trigger, y_t is the target label.

CDA. CDA measures whether the backdoored model can maintain high prediction accuracy of clean data samples.

$$CDA(F_A, \mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbf{I}_{[F_A(x)=y]}, \tag{12}$$

where x is a sample of the clean dataset $\mathcal{X}$ and y is the ground-truth label of x.

SASR. In one-to-one attacks, SASR is computed as the probability that a trigger-imposed sample of the source label is misclassified to the target label.

$$ASR(F_A, \mathcal{X}_\mathcal{S}) = \frac{1}{|\mathcal{X}_\mathcal{S}|} \sum_{x \in \mathcal{X}_\mathcal{S}} \mathbf{I}_{[F_A(x+T_i)=y_t]}. \tag{13}$$

where x is the clean sample of the source label dataset $\mathcal{X}_\mathcal{S}$.

SCDA. In one-to-one attacks, SCDA measures whether the backdoored model can maintain prediction accuracy of clean data samples of the source label.

$$SCDA(F_A, \mathcal{X}_\mathcal{S}) = \frac{1}{|\mathcal{X}_\mathcal{S}|} \sum_{x \in \mathcal{X}_\mathcal{S}} \mathbf{I}_{[F_A(x)=y_S]}, \tag{14}$$

where x is the clean sample of the source label dataset $\mathcal{X}_\mathcal{S}$ and y_S is the source label.

SSIM. SSIM is a commonly-used Quality-of-Experience (QoE) metric [7] that quantifies the differences in luminance, contrast, and structure between the original image and the distorted image.

$$SSIM = A(x, x')^\alpha B(x, x')^\beta C(x, x')^\gamma, \tag{15}$$

where $A(x, x'), B(x, x')$, and $C(x, x')$ quantify the luminance similarity, contrast similarity, and structural similarity between the original image x and the distorted image x'. α, β, and γ are parameters in the range $[0, 1]$.

PSNR. PSNR is computed based on MSE (Mean Squared Error) regarding the signal energy.

$$\begin{aligned} PSNR &= 10 \log_{10} \frac{E}{MSE}, \\ MSE &= \frac{1}{N} \sum_i (x'_i - x_i)^2, \end{aligned} \tag{16}$$

where E is the maximum signal energy.

TABLE II. COMPARISON OF MEGATRON WITH DBIA [25], BAVT [35], DBAVT-BADNETS [8], AND DBAVT-WANET [8] IN ONE-TO-ONE ATTACKS.

		SCDA	SASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	DBIA	97.8%	98.0%	28.46102892	0.994829113	0.031920945	0.002289052
	BAVT	84.5%	80.3%	24.78704153	0.999612272	0.084357165	0.026071861
	DBAVT-BadNets	**100%**	**100%**	29.36041068	0.990515120	0.046213830	0.002060820
	DBAVT-WaNet	98.0%	98.0%	32.94843157	0.999714076	0.017961027	0.001361159
	MEGATRON	**100%**	**100%**	**55.07969693**	**0.99999994**	**0.000953811**	**0.000370741**
		SCDA	SASR	PSNR	SSIM	LPIPS	L_1 distance
GTSRB	DBIA	95.6%	96.0%	28.38291027	0.99516244	0.033892017	0.002481025
	BAVT	86.8%	82.5%	23.98223772	0.99952215	0.068876714	0.027542898
	DBAVT-BadNets	94.0%	92.0%	29.36053465	0.99076884	0.044891710	0.002016550
	DBAVT-WaNet	94.0%	96.0%	31.15860532	0.99905623	0.016359457	0.001382915
	MEGATRON	**100%**	**99.0%**	**55.39791739**	**0.99999994**	**0.000582156**	**0.000368296**
		SCDA	SASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-100	DBIA	89.8%	98.0%	28.05872893	0.99414685	0.032999403	0.002449209
	BAVT	82.2%	78.8%	23.32390113	0.99943417	0.071154475	0.029951381
	DBAVT-BadNets	92.0%	**99.0%**	28.53759710	0.99053525	0.041157600	0.002227320
	DBAVT-WaNet	96.0%	94.0%	30.83291898	0.99914402	0.014241966	0.001672578
	MEGATRON	**100%**	**99.0%**	**55.95790627**	**0.99999994**	**0.00070513**	**0.000344695**
		SCDA	SASR	PSNR	SSIM	LPIPS	L_1 distance
Tiny ImageNet	DBIA	82.4%	**100%**	27.89105781	0.99380216	0.03479015	0.002684914
	BAVT	84.8%	87.2%	24.29126279	0.99955231	0.06722090	0.026612388
	DBAVT-BadNets	96.0%	96.0%	29.45663930	0.98945220	0.04798450	0.002010100
	DBAVT-WaNet	96.0%	96.0%	30.74312991	0.99972351	0.01795189	0.001365321
	MEGATRON	**100%**	**100%**	**55.35610987**	**0.99999994**	**0.00087185**	**0.000368781**

LPIPS. LPIPS [42] measures the similarity between two images based on the idea that the human visual system processes images in a hierarchical manner, where lower-level features, e.g., edges and textures, are processed before higher-level features, e.g., objects and scenes. The LPIPS metric uses a deep neural network to calculate the similarity between the two images.

$$LPIPS(A,B) = \sum_i w_i * ||F_i(A) - F_i(B)||^2 \qquad (17)$$

where $F_i(A)$ and $F_i(B)$ are the feature representations of images A and B at layer i of the pre-trained neural network, $||.||$ denotes the L_2 norm, and w_i is a weight that controls the relative importance of each layer. LPIPS has been shown to outperform other metrics, e.g., SSIM and PSNR, in measuring perceptual similarity between images, especially in cases where the images differ in high-level perceptual qualities such as texture and style. The smaller the value, the more similar the two images are.

L_1 distance. L_1 distance uses L_1 norm to express the distance, i.e., Manhattan distance.

$$L_1 \ distance = \frac{1}{m}\sum_{i=1}^{m}|y_i - \check{y}_i|, \qquad (18)$$

where y_i and $\check{y}_i$ represent the pixel values at the same position in the original image and the backdoored image, respectively, and m is the size of these two images.

The larger the SSIM and the PSNR, the more similar the original and backdoored images are, and the more stealthy the trigger is. The smaller the LPIPS and the L_1 distance, the more stealthy the trigger is.

All experiments are implemented in Python and run on a 14-core Intel(R) Xeon(R) Gold 5117 CPU @2.00GHz and NVIDIA GeForce RTX 2080 Ti GPU machine running Ubuntu 18.04 system.

B. Evaluation Results

Any-to-one attacks. We compare MEGATRON with state-of-the-art vision transformer backdoor attacks, i.e., DBIA [25], BAVT [35], DBAVT-BadNets [8], and DBAVT-WaNet [8]. As the source code of TrojViT [43] is not publicly available, we did not compare MEGATRON with it. Note that BAVT [35] mainly focuses on the one-to-one attack setting, so we only compare MEGATRON with it in the one-to-one attack setting. We implement the baseline attacks using their published source codes. We used the ViT-base model for our experiments, which only applies to data samples of size $224\times$ 224. Therefore, we resized all samples from the four datasets to $224\times$ 224. The default trigger size of all the baseline attacks is 16×16 and the default poisoning rate is 3%. The entire trigger of MEGATRON is also 16×16, but the size of each sub-trigger is only 4 pixels. As shown in Table I, MEGATRON outperforms baselines across all four datasets, especially for the image quality metrics. MEGATRON achieves PSNR of 55.2894, 55.5129, 55.8918, and 55.4193 on CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet models, respectively, while the highest PSNR of the baseline is only 31.0849 (CIFAR-10), 31.2756 (GTSRB), 31.3110 (CIFAR-100), and 31.0481 (Tiny ImageNet). Furthermore, MEGATRON maintains a high prediction accuracy on clean samples. Although BAVT and DBAVT-WaNet are invisible backdoor attacks against vision transformers, their triggers are based on global noise, resulting in relatively low PSNR, high LPIPS, and high L_1 distance.

We also present the backdoored samples of baselines and MEGATRON in Fig. 2. We can see that the backdoored samples generated by MEGATRON are quite similar to the benign samples and are easy to evade human visual inspections. The

TABLE III. ABLATION STUDY OF DIFFERENT LOSS FUNCTIONS.

		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	$\mathcal{L}_{CE}$	95.72%	18.4%	55.33277764	0.99999994	0.000883320	0.000363716
	$\mathcal{L}_{CE} + \mathcal{L}_{LA}$	96.43%	83.6%	55.16661906	0.99999994	0.000930156	0.000370420
	$\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$	96.17%	100%	55.26837881	0.99999995	0.000912101	0.000365185
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
GTSRB	$\mathcal{L}_{CE}$	96.36%	16.9%	55.57434798	0.99999994	0.000574965	0.000366582
	$\mathcal{L}_{CE} + \mathcal{L}_{LA}$	97.90%	81.0%	55.52873137	0.99999993	0.000579246	0.000366534
	$\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$	98.09%	100%	55.47438277	0.99999998	0.000592260	0.000367091
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-100	$\mathcal{L}_{CE}$	89.40%	15.3%	55.88299801	0.99999994	0.000665277	0.000345507
	$\mathcal{L}_{CE} + \mathcal{L}_{LA}$	89.35%	80.9%	55.94410937	0.99999994	0.000659543	0.000347341
	$\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$	89.37%	100%	55.85936690	0.99999999	0.000617437	0.000372015
		CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
Tiny ImageNet	$\mathcal{L}_{CE}$	89.63%	18.0%	55.33277764	0.99999995	0.000883320	0.000363716
	$\mathcal{L}_{CE} + \mathcal{L}_{LA}$	88.06%	81.1%	55.44853459	0.99999994	0.000865381	0.000364763
	$\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$	88.91%	98.0%	55.40285920	0.99999995	0.000865381	0.000364735

TABLE IV. IMPACT OF MODEL STRUCTURE ON MEGATRON.

	Structure	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	ViT-small	91.72%	100%	55.08017895	0.99999993	0.000365849	0.000319285
	ViT-base	96.92%	100%	55.10029955	0.99999994	0.000365385	0.000370650
	DeiT	95.61%	100%	55.07969693	0.99999995	0.000353811	0.000370741
	T2t_ViT	94.23%	99.2%	55.09017457	0.99999992	0.000393721	0.000349759
GTSRB	ViT-small	96.95%	99.0%	55.59089424	0.99999994	0.000534681	0.000366814
	ViT-base	96.50%	100%	55.64672715	0.99999994	0.000494130	0.000366481
	DeiT	97.86%	99.1%	55.39791739	0.99999994	0.000582156	0.000368296
	T2t_ViT	95.91%	95.7%	55.64209109	0.99999994	0.000542595	0.000365897
CIFAR-100	ViT-small	82.93%	95.7%	55.80761729	0.99999994	0.000711249	0.000346755
	ViT-base	90.51%	99.4%	55.95971632	0.99999994	0.000656329	0.000344603
	DeiT	88.49%	99.1%	55.95790627	0.99999994	0.000705132	0.000344695
	T2t_ViT	92.93%	94.8%	55.80365986	0.99999994	0.000723859	0.000347454
Tiny ImageNet	ViT-small	84.07%	98.0%	58.73526873	1.00000000	0.000623189	0.000355080
	ViT-base	88.00%	100%	58.73526873	1.00000000	0.000656971	0.000355081
	DeiT	88.80%	100%	55.35610987	0.99999994	0.000671851	0.000368784
	T2t_ViT	84.13%	98.4%	55.73526873	1.00000000	0.000656971	0.000355082

backdoor samples of BAVT are natural in most cases, but there are still noises visible to the human eye in some cases. Moreover, BAVT is not effective in the all-to-one attack setting. Although DBAVT-WaNet can generate relatively natural backdoor samples, it cannot achieve a high attack success rate. The image quality of MEGATRON is also better than the baselines.

One-to-one attacks. In the one-to-one attack, we choose a specific label as the source label and then poison the samples in that label. As shown in Table II, MEGATRON outperforms the baselines across all four datasets under the one-to-one attack settings, especially for the image quality metrics.

Ablation study. In this section, we conduct an ablation study to examine the necessity of the three loss items, i.e., $\mathcal{L}_{CE}$, $\mathcal{L}_{LA}$, and $\mathcal{L}_{ro}$. The results are shown in Table III. Comparing $\mathcal{L}_{CE}$ and $\mathcal{L}_{CE} + \mathcal{L}_{LA}$, we can observe that $\mathcal{L}_{LA}$ can significantly improve ASR. For example, for GTSRB dataset, the ASR of $\mathcal{L}_{CE}$ is 16.9%, but reaches as high as 81% using the $\mathcal{L}_{CE} + \mathcal{L}_{LA}$ loss function. The increment is more than 64.1%. Similarly, for Tiny ImageNet, the improvement in ASR is more than 63.1% with $\mathcal{L}_{LA}$. The success of $\mathcal{L}_{LA}$ lies in its ability to minimize the distance between the poisoned sample and the sample of the target label with respect to each layer's attention, thereby aiding the model in misclassifying the backdoored samples to the target label. Compared with the $\mathcal{L}_{CE} + \mathcal{L}_{LA}$ loss function, the $\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$ attack further increases ASR, where the average improvement of the four datasets exceeded 17%. The success of $\mathcal{L}_{ro}$ lies in its ability to increase the importance of the attention diffusion area during training.

Impact of model structure. By default, the backdoored models of CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNette are using ViT-base [9] model structure. In this part, we explore whether MEGATRON is also effective against other model structures, such as ViT-small [36], Deit [36], and T2t_ViT [41]. The results are shown in Table IV. It is shown that MEGATRON is robust to the structures of the target vision transformer structure. MEGATRON can effectively inject the backdoor into the target vision transformer, regardless of the structure of the vision transformer.

TABLE V. IMPACT OF TRIGGER LOCATION ON MEGATRON.

	Position	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	(20,20)	97.74%	98.4%	55.88304269	1.000000000	0.000812341	0.000383458
	(40,40)	97.36%	98.5%	55.69034224	1.000000000	0.000892149	0.000382135
	(60,60)	97.84%	98.1%	55.56567468	0.999999995	0.000881290	0.000382023
	(80,80)	97.55%	99.4%	55.40578213	0.999999999	0.000880124	0.000382315
	(100,100)	97.52%	99.6%	55.30234565	0.999998999	0.000878213	0.000381982
GTSRB	(20,20)	96.64%	98.3%	55.45604269	1.000000000	0.000512947	0.000346435
	(40,40)	96.82%	99.4%	55.36634224	1.000000000	0.000592150	0.000346835
	(60,60)	97.73%	99.0%	55.30573467	0.999999998	0.000551409	0.000345315
	(80,80)	97.41%	100%	55.31178213	0.999999996	0.000606570	0.000344982
	(100,100)	96.59%	100%	55.30234565	0.999999993	0.000593235	0.000344608
CIFAR-100	(20,20)	90.41%	97.3%	55.45604269	1.000000000	0.000652370	0.000326113
	(40,40)	90.06%	97.5%	55.36634224	0.999999523	0.000686863	0.000358509
	(60,60)	90.96%	96.5%	55.30573467	0.999998808	0.000614091	0.000308281
	(80,80)	90.15%	99.4%	55.31178213	0.999997616	0.000660217	0.000386553
	(100,100)	90.35%	100%	55.30234565	0.999996483	0.000680740	0.000353796
Tiny ImageNet	(20,20)	88.13%	98.0%	55.28556513	0.999999955	0.000768024	0.000369804
	(40,40)	88.14%	100%	55.21535865	0.999999953	0.000780808	0.000369250
	(60,60)	88.62%	94.3%	55.08408015	0.999999951	0.000701460	0.000370872
	(80,80)	88.44%	100%	55.02061542	0.999999953	0.000789534	0.000371543
	(100,100)	88.08%	92.4%	55.03214851	0.999999970	0.000707653	0.000371166

TABLE VI. IMPACT OF TRIGGER LOCATION CHANGE ON MEGATRON.

	Token number	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	1	95.40%	100%	55.12718418	0.99999994	0.000865705	0.000370542
	2	95.29%	99.2%	55.10483436	0.99999994	0.000872751	0.000370584
	3	96.10%	100%	55.01666710	0.99999995	0.000834102	0.000372011
	4	96.07%	95.3%	55.09496802	0.99999994	0.000859978	0.000370849
	5	95.20%	87.6%	55.08400565	0.99999994	0.000842098	0.000370875
GTSRB	1	97.45%	98.2%	55.45029591	0.99999995	0.000555712	0.000367224
	2	97.27%	94.4%	55.48857662	0.99999993	0.000556716	0.000366524
	3	97.95%	89.3%	55.51740730	0.99999995	0.000588811	0.000366934
	4	97.77%	63.0%	55.50784454	0.99999995	0.000557540	0.000366623
	5	98.09%	50.6%	55.41712970	0.99999994	0.000576242	0.000367201
CIFAR-100	1	90.34%	99.9%	55.73990665	0.99999994	0.000606907	0.000341978
	2	90.02%	97.8%	55.74793288	0.99999994	0.000590802	0.000342094
	3	89.95%	98.3%	55.74740090	0.99999994	0.000565670	0.000342203
	4	90.12%	88.6%	55.73191193	0.99999993	0.000543302	0.000342200
	5	90.67%	81.7%	55.72900670	0.99999991	0.000572505	0.000336410
Tiny ImageNet	1	88.80%	100%	55.35566904	0.99999993	0.000730404	0.000368850
	2	88.49%	100%	55.35363980	0.99999995	0.000786422	0.000368761
	3	88.83%	96.3%	55.34742373	0.99999994	0.000784255	0.000368647
	4	88.85%	88.1%	55.31426417	0.99999994	0.000763066	0.000368793
	5	88.81%	80.0%	55.32264675	0.99999994	0.000708849	0.000368444

Impact of trigger location. We also explore the impact of the attack success rate when the trigger is located at different positions in the image. Note that the sub-trigger is at the same location during the training and testing phases. As shown in Table V, x and y represent the horizontal and vertical coordinate values of the plane Cartesian coordinate system formed by extending left and downward from the upper left corner of the image as the origin. We can see that the attack performance of MEGATRON is robust to the location of the trigger.

Impact of trigger location change. To achieve the best attack performance, the sub-trigger should be in the same position in the training phase and the testing phase. However, this is not realistic in a real-world backdoor attack. We evaluate the change of sub-trigger locations. During testing, we put the sub-trigger within the attention diffusion area but not exactly the location at the training phase. As shown in Table VI, the attack success rate will decrease as the diffusion area expands since the sub-trigger deviates more from the location at the training phase. However, MEGATRON can also maintain a high attack success rate when the attention diffusion area is three tokens, which is much larger than the trigger area.

Impact of attention diffusion range. In this part, we require that the backdoor trigger is in the same position in the

TABLE VII. IMPACT OF ATTENTION DIFFUSION RANGE ON MEGATRON.

	Token number	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	1	95.80%	100%	55.1199763	0.9999999	0.0008392	0.0003706
	2	95.74%	100%	55.0974996	0.9999999	0.0008152	0.0003707
	3	95.52%	100%	55.0880295	0.9999999	0.0008803	0.0003720
	4	95.50%	100%	55.0988182	0.9999999	0.0008359	0.0003705
	5	96.00%	99.5%	55.0846736	0.9999999	0.0008220	0.0003705
GTSRB	1	96.23%	100%	55.4998810	0.9999999	0.0005433	0.0003670
	2	96.59%	100%	55.5056186	0.9999999	0.0005147	0.0003668
	3	97.82%	99.6%	55.0806919	0.9999999	0.0005872	0.0003703
	4	96.73%	99.8%	55.4676386	0.9999999	0.0005258	0.0003668
	5	96.68%	99.6%	55.0806919	0.9999999	0.0005872	0.0003703
CIFAR-100	1	90.35%	99.5%	55.8646942	0.9999999	0.0007247	0.0003473
	2	90.45%	99.6%	55.8786791	0.9999999	0.0006811	0.0003465
	3	90.34%	99.8%	56.0999511	0.9999999	0.0006369	0.0003461
	4	90.53%	99.8%	55.8607316	0.9999999	0.0006991	0.0003471
	5	90.76%	100%	55.8743049	0.9999999	0.0007309	0.0003466
Tiny ImageNet	1	88.46%	100%	55.2999153	0.9999999	0.0007301	0.0003669
	2	88.22%	100%	55.2607803	0.9999999	0.0007889	0.0003674
	3	88.66%	100%	55.4485346	0.9999999	0.0007654	0.0003648
	4	88.68%	100%	55.3089686	0.9999999	0.0007023	0.0003672
	5	88.83%	100%	55.2822248	0.9999999	0.0007924	0.0003667

TABLE VIII. IMPACT OF POISON RATE ON MEGATRON.

	Rate	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	1%	97.32%	98.4%	55.28326050	0.99999994	0.000876566	0.000367377
	2%	97.86%	98.8%	55.14347332	0.99999993	0.000851029	0.000367930
	3%	97.55%	99.1%	55.15277753	0.99999996	0.000825128	0.000367592
	4%	97.72%	99.3%	55.33257242	0.99999994	0.000803989	0.000367162
	5%	98.45%	100%	54.67937364	0.99999995	0.000894337	0.000371980
GTSRB	1%	97.50%	99.3%	55.79231347	0.99999997	0.000521228	0.000346876
	2%	97.36%	100%	55.70948274	0.99999994	0.000517764	0.000362779
	3%	97.36%	100%	55.78185908	0.99999996	0.000540966	0.000362127
	4%	97.82%	100%	55.71525661	0.99999994	0.000505133	0.000361136
	5%	97.42%	100%	55.54923701	0.99999999	0.000466593	0.000362319
CIFAR-100	1%	90.42%	97.3%	55.89790965	0.99999994	0.000672373	0.000346203
	2%	90.44%	98.5%	55.78205076	0.99999995	0.000742281	0.000346829
	3%	90.32%	100%	55.86770569	0.99999997	0.000756335	0.000347210
	4%	90.26%	100%	55.83582672	0.99999994	0.000742553	0.000344586
	5%	90.08%	100%	55.84080138	0.99999993	0.000759897	0.000347163
Tiny ImageNet	1%	89.04%	96.2%	55.40198761	0.99999994	0.000877833	0.000364571
	2%	88.78%	96.5%	55.36056336	0.99999995	0.000870325	0.000019670
	3%	88.90%	98.3%	55.44853459	0.99999994	0.000865381	0.000364593
	4%	88.03%	99.0%	55.29405425	0.99999996	0.000882321	0.000364638
	5%	88.20%	100%	55.40427096	0.99999995	0.000884292	0.000364655

sample during both the training and testing phases and explore the impact of the attention diffusion range. Table VII demonstrates that MEGATRON maintains its robustness to changes in the attention diffusion range. When the trigger remains in the same position in the sample during both the training and testing phases, the attack performance of MEGATRON remains unaffected by variations in the attention diffusion range.

Impact of poison rate. The poison rate refers to the proportion of poisoned samples to all training samples. In our study, we investigate the impact of poison ratio as shown in Table VIII. The results demonstrate that as the poison rate increases, the attack success rate also tends to rise. For instance, when the poison ratio is set at 1%, MEGATRON achieves an attack success rate (ASR) of 98.4% for CIFAR-10 and 99.3% for GTSRB. However, with a higher poison ratio of 5%, MEGATRON achieves even higher ASRs of 100% for CIFAR-10 and 100.0% for GTSRB. Notably, MEGATRON maintains a consistently high model prediction accuracy and image quality even as the poison rate increases.

Impact of sub-trigger size. We explore the impact of sub-trigger size on the attack performance of MEGATRON. The results are shown in Table IX. The entire trigger size is 16 $\times$ 16 pixels, and the second row in the table denotes the attack performance of the entire trigger. Before the sub-trigger size is reduced to 4 pixels, as the sub-trigger size decreases, the attack success rate (ASR) remains largely unchanged, but the sub-

TABLE IX. IMPACT OF THE SUB-TRIGGER SIZE ON MEGATRON.

	Activation zone (pixels)	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	256	95.23%	100%	32.1278451	0.9999540	0.0417581	0.0063317
	8	96.11%	100%	52.3182233	0.9999999	0.0038441	0.0004798
	4	95.04%	99.6%	55.4690174	1.0000000	0.0008265	0.0003024
	2	95.53%	94.1%	58.3642632	1.0000000	0.0004993	0.0000972
	1	95.21%	85.3%	61.2015198	1.0000000	0.0001256	0.0000841
GTSRB	256	97.10%	100%	32.8941780	0.9999612	0.0337911	0.0056174
	8	97.18%	100%	52.9283479	0.9999999	0.0050283	0.0004947
	4	96.01%	98.3%	55.3892085	1.0000000	0.0006782	0.0003375
	2	97.13%	93.6%	59.3359268	1.0000000	0.0002089	0.0001965
	1	96.68%	84.8%	62.1927584	1.0000000	0.0001258	0.0000821
CIFAR-100	256	90.10%	100%	33.2651952	0.9999663	0.0314782	0.0051925
	8	90.37%	100%	53.7047081	1.0000000	0.0015358	0.0004563
	4	90.33%	99.6%	55.6630813	1.0000000	0.0007320	0.0003104
	2	90.29%	93.3%	58.5406855	1.0000000	0.0002043	0.0000893
	1	89.08%	82.3%	62.3722674	1.0000000	0.0000501	0.0000786
Tiny ImageNet	256	89.60%	100%	32.6283819	0.9999576	0.0345125	0.0052715
	8	88.37%	100%	52.6745901	0.9999999	0.0023011	0.0004697
	4	88.33%	98.1%	55.0714010	1.0000000	0.0008628	0.0003178
	2	88.29%	91.8%	59.0719374	1.0000000	0.0003131	0.0001940
	1	88.08%	77.6%	61.5858275	1.0000000	0.0000787	0.0000818

trigger becomes less visible with better image quality metrics. For example, with a sub-trigger size of 4 pixels, MEGATRON achieves ASR of 99.6% (CIFAR-10), 98.3% (GTSRB), 99.6% (CIFAR-100), and 98.1% (Tiny ImageNet) and achieves PNSR of 55.4690 (CIFAR-10), 55.3892 (GTSRB), 55.6631 (CIFAR-100), and 55.0714 (Tiny ImageNet). With the entire trigger, MEGATRON achieves ASR of 100% (CIFAR-10), 100% (GTSRB), 100% (CIFAR-100), and 100% (Tiny ImageNet) and achieves PNSR of 32.1278 (CIFAR-10), 32.8942 (GTSRB), 33.2652 (CIFAR-100), and 32.6284 (Tiny ImageNet). These results demonstrate the effectiveness of the stealthy sub-trigger. In MEGATRON, we set the sub-trigger size as 4 pixels.

Impact of trigger masking. We investigate the attack performance of MEGATRON under different trigger generation methods, i.e., no-mask and mask. The results are presented in Table X. It is demonstrated that the MEGATRON with mask method consistently outperforms MEGATRON with no-mask regarding attack success rate and image quality metrics across all datasets.

Impact of α. We explore the impact of the hyperparameter α in the loss function on the attack performance. The results are shown in Table XI. We can see that as the value of α increases, CDA and ASR will significantly decrease. For example, MEGATRON can achieve CDA of 96.32% and ASR of 100% when α is 0.01 for CIFAR-10, while MEGATRON can only achieve CDA of 67.94% (CDA) and 34.9% (ASR) when α is 100. In the experiments, we set α as 0.01 by default.

Impact of β. We also investigate the influence of the hyperparameter β in the loss function on the performance of the attack. The results are shown in Table XVI in the Appendix. It is shown that the attack performance of MEGATRON fluctuates with different values of β. A higher β does not necessarily lead to a higher attack success rate. However, our ablation study reveals that the presence of $\mathcal{L}_{ro}$ significantly affects the attack success rate, regardless of the previous hyperparameter values. Without $\mathcal{L}_{ro}$, the attack success rate decreases considerably. In the experiments, we set β as 10 by default.

Impact of transparency adjustment. There are two parameters in transparency adjustment, i.e., ϕ_A and ϕ_D. ϕ_A is the transparency of the sub-trigger area, and ϕ_D is the transparency of the rest trigger area. The lower the transparency value is, the more imperceptible the trigger is. We explore the impact of the transparency value of ϕ_A in Table XVII (appendix) and the impact of the transparency value of ϕ_D in Table XVIII (appendix). We also present the backdoored samples with different transparency values in Fig. 4 (appendix). We can see that as the transparency value is smaller, the trigger is more concealed, and the backdoored samples have better image qualities. However, the attack success rate will be decreased. When the transparency value of ϕ_A and ϕ_D are set as 0.5 and 0.01, the human eyes can hardly discern the trigger. Thus, in the experiments, we set the ϕ_A as 0.5 and ϕ_D as 0.01 by default.

Time cost. To evaluate the efficiency of MEGATRON, we conducted a comparison of computational costs between MEGATRON and the baselines, as shown in Table XIII. We can see that MEGATRON demonstrates similar computational costs to the baseline attacks. Importantly, MEGATRON achieves superior attack performance while maintaining reasonable computational efficiency.

V. ROBUSTNESS TO STATE-OF-THE-ART BACKDOOR DEFENSES

In this section, we investigate the ability of MEGATRON to evade state-of-the-art backdoor defenses. Currently, there are only two available defenses specifically designed for vision transformers, which are DBAVT [8] and BAVT [35]. We also adapt two other advanced backdoor defenses, namely Beatrix [26] and SAGE [14], to defend against MEGATRON.

DBAVT. DBAVT [8] mitigates the backdoor attacks on ViTs using patch processing. It is based on the insight that

TABLE X. IMPACT OF TRIGGER MASKING.

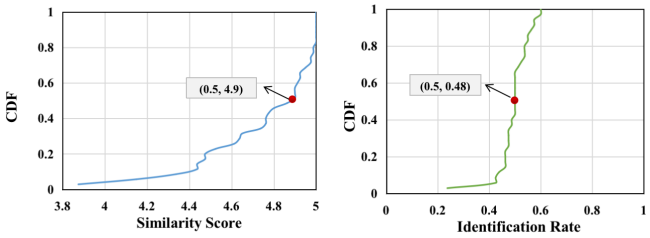
	Operation	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	no mask	95.34%	99.5%	28.2327216	0.9904541	0.0458291	0.0063321
	mask	95.26%	99.6%	55.0097682	0.9999999	0.0008301	0.0003756
GTSRB	no mask	97.08%	100%	28.8941742	0.9907612	0.0437912	0.0056174
	mask	97.34%	98.3%	55.2199656	0.9999999	0.0004583	0.0003717
CIFAR-100	no mask	90.28%	99.8%	28.2651952	0.9906663	0.0424135	0.0051925
	mask	90.21%	99.6%	55.6946467	0.9999999	0.0007367	0.0003517
Tiny ImageNet	no mask	89.56%	100%	28.6283819	0.9894776	0.0445125	0.0052715
	mask	88.43%	98.2%	55.3561098	0.9999999	0.0007718	0.0003687

TABLE XI. IMPACT OF α ON MEGATRON.

	Value	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	0.01	96.32%	100%	55.35134440	1.00000000	0.000883393	0.000357004
	0.1	95.43%	100%	55.36250116	1.00000000	0.000865853	0.000357413
	1	90.63%	100%	55.39803874	1.00000000	0.000874947	0.000357312
	10	90.94%	100%	55.29836696	1.00000000	0.000897934	0.000357319
	100	67.94%	34.9%	55.52056025	1.00000000	0.000803492	0.000356302
GTSRB	0.01	96.73%	98.0%	58.95353311	1.00000000	0.000540838	0.000354926
	0.1	97.09%	96.9%	55.09559303	1.00000000	0.000532751	0.000364854
	1	93.23%	95.8%	55.83036380	1.00000000	0.000545322	0.000365314
	10	90.86%	91.9%	55.91727758	1.00000000	0.000548973	0.000364952
	100	89.77%	85.0%	55.97818240	1.00000000	0.000566222	0.000355187
CIFAR-100	0.01	90.02%	95.0%	56.17814352	1.00000000	0.000680072	0.000358388
	0.1	87.48%	93.9%	55.27621966	1.00000000	0.000686366	0.000357007
	1	87.13%	98.5%	55.44067813	1.00000000	0.000682036	0.000357007
	10	86.90%	98.6%	55.16034613	1.00000000	0.000682137	0.000357007
	100	68.69%	40.0%	56.21109435	1.00000000	0.000683177	0.000348490
Tiny ImageNet	0.01	88.83%	100%	55.35610987	0.99999994	0.000867185	0.000368781
	0.1	88.02%	100%	55.73526873	1.00000000	0.000825392	0.000355080
	1	88.84%	100%	55.73842042	1.00000000	0.000832819	0.000355462
	10	87.05%	96.0%	55.73729479	1.00000000	0.000854562	0.000354539
	100	76.86%	88.0%	55.83637481	1.00000000	0.000856971	0.000355943

TABLE XII. APPLY DBAVT [8] TO MEGATRON.

	Original		DBAVT	
	CDA	ASR	CDA	ASR
CIFAR-10	97.25%	100%	93.35%	99.5%
GTSRB	97.54%	100%	95.04%	94.6%
CIFAR-100	90.38%	99.3%	84.02%	99.6%
Tiny ImageNet	88.60%	100%	87.60%	94.2%



(a) Similarity test (b) Inspection test

Fig. 3. User study results.

clean-data accuracy and backdoor attack success rates of ViTs respond differently to patch transformations before the positional encoding, unlike CNN models.

As shown in Table XII, after applying DBAVT, the ASR of MEGATRON only slightly decreases. MEGATRON can also achieve ASR of 99.5%, 94.6%, 99.6%, and 94.2% for CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet, respectively. The possible reason is that to maintain a high prediction accuracy of the model, the percentage of patches dropped and shuffled of DBAVT cannot be set too high when defending against MEGATRON. As a result, MEGATRON is robust to DBAVT.

BAVT. BAVT [35] is based on the discovery that the interpretation map generated by transformers can effectively highlight the trigger used in a backdoored image. BAVT blocks out the region of the image with the highest values in the interpretation heatmap, thereby reducing the success rate of backdoor attacks.

We apply BAVT to MEGATRON, and the defense results are shown in Table XIV. We can see that MEGATRON is robust to BAVT, as it maintains a high attack success rate even after BAVT is applied. This is likely due to the fact that BAVT uses the zero-setting operation on the region with the highest attention rollout in the test image, whereas MEGATRON maximizes the attention rollout in the diffusion

TABLE XIII. TIME COST OF MEGATRON AND BASELINE ATTACKS AGAINST VIT.

	DBIA	BAVT	DBAVT-BadNets	DBAVT-WaNet	MEGATRON
CIFAR-10	11.43h	6.35h	0.98h	3.13h	2.24h
GTSRB	3.23h	1.46h	0.25h	0.67h	0.47h
CIFAR-100	8.78h	4.52h	0.69h	2.12h	1.63h
Tiny ImageNet	17.91h	10.42h	1.95h	5.10h	3.76h

TABLE XIV. APPLY BAVT TO MEGATRON.

	Original		DBAVT	
	CDA	ASR	CDA	ASR
CIFAR-10	97.25%	100%	94.72%	98.4%
GTSRB	97.54%	100%	97.33%	86.3%
CIFAR-100	90.38%	99.3%	87.59%	99.0%
Tiny ImageNet	88.60%	100%	89.64%	92.4%

region rather than the poisoning region. Consequently, BAVT tends to cover non-poisoning areas within the diffusion area, making it susceptible to MEGATRON.

Beatrix. Beatrix (backdoor detection via Gram matrix) [26] is a novel backdoor defense approach that utilizes Gram matrix to capture both feature correlations and high-order information of representations. By learning class conditional statistics from activation patterns of normal samples, Beatrix effectively identifies poisoned samples by detecting anomalies in activation patterns. We first adapted Beatrix from CNN models to vision transformers. Then we applied Beatrix to MEGATRON.

It is shown that MEGATRON is robust to Beatrix, as it only achieves 21.1% (CIFAR-10), 20.2% (GTSRB), 19.4% (CIFAR-100), and 18.5% (Tiny ImageNet) F1-score in detecting MEGATRON. MEGATRON can successfully evade Beatrix since the high-order information it captures is largely ineffective or unavailable when dealing with our large transformer models.

SAGE. SAGE [14] is a newly-proposed backdoor purification method. It aims to correct toxic deep layers in a neural network by leveraging attention map alignment with innocent shallow layers. SAGE employs a layer-wise top-down self-attention distillation (SAD) technique to purify backdoored models. We first adapted SAGE from CNN models to vision transformers and then applied SAGE to the backdoored vision transformers generated by MEGATRON.

The results are shown in Table XV (appendix). It is shown that SAGE cannot effectively purify the backdoor from our backdoored models. MEGATRON can also achieve ASR of 100% (CIFAR-10), 99.3% (GTSRB), 99.4% (CIFAR-100), and 100% (Tiny ImageNet) after being purified by SAGE defense. MEGATRON can successfully evade the SAGE backdoor defenses, possibly because that MEGATRON has damaged the attention mechanism of different layers of the vision transformer rather than only poisoning the deeper layers as in CNN. Thus, our generated backdoor in the transformer cannot be purified by SAGE.

VI. USER STUDY

We conducted two sets of user studies to evaluate the concealment of backdoored samples of MEGATRON. Fifty volunteers aged 20-30, including college students and faculty members, were recruited for the study. Prior to the user study, we provided a thorough explanation of MEGATRON to the volunteers and confirmed their understanding of the attacks. For testing, we randomly selected 80 benign images from CIFAR-10, GTSRB, CIFAR-100, and Tiny ImageNet, and generated their backdoored versions using MEGATRON.

1) Similarity Test: In the first test, we presented benign samples and their corresponding backdoored samples side-by-side to the volunteers. The benign samples were displayed on the left, and the backdoored samples were displayed on the right. The volunteers were asked to judge the similarity between the two samples using a scale ranging from 1 to 5. A score of 5 represented "look exactly the same", 4 represented "very similar", 3 represented "a little similar", 2 represented "not very similar", and 1 represented "very different". The cumulative distribution function (CDF) of the similarity scores is shown in Fig. 3(a). Experimental results demonstrate that over 50% of the backdoored samples generated by MEGATRON exhibit a similarity score surpassing 4.9, providing evidence that the backdoored samples are visually natural and capable of evading human visual inspections.

2) Inspection Test: In the second test, we presented benign samples and their corresponding backdoored samples side-by-side to the volunteers and asked them to choose which sample was the backdoored one. To avoid any bias, we shuffled the position of the benign and backdoored samples within each pair, so that the backdoored sample was not necessarily on the right. We calculated the percentage of correct answers as the identification rate. The cumulative distribution function (CDF) of the identification rate is shown in Fig. 3(b). It is shown that approximately 50% of the samples generated by MEGATRON exhibit an identification rate lower than 0.48, akin to random guessing. This provides additional evidence supporting the imperceptibility of the backdoor trigger in MEGATRON.

VII. CONCLUSION AND DISCUSSION

This paper presents an effective and evasive backdoor attack against vision transformers. We have carefully designed the trigger generation and backdoor injection algorithms to achieve a high attack success rate with imperceptible triggers. Extensive experiments have demonstrated the superiority of our proposed attack compared to baseline backdoor attacks. There are several potential avenues for future research. Firstly, it may be possible to generalize MEGATRON beyond the vision domain, extending it to other domains such as voice, text, and video. Secondly, effective defenses against MEGATRON are necessary to mitigate the potential risks posed by such attacks.

REFERENCES

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward transformer-based object detection. *arXiv preprint arXiv:2012.09958*, 2020.

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021.

[4] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *IEEE/CVF International Conference on Computer Vision*, pages 357–366, 2021.

[5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[6] Yanjiao Chen, Xueluan Gong, Qian Wang, Xing Di, and Huayang Huang. Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. *IEEE Network*, 34(5):141–147, 2020.

[7] Yanjiao Chen, Kaishun Wu, and Qian Zhang. From QoS to QoE: A tutorial on video quality assessment. *IEEE Communications Surveys & Tutorials*, 17(2):1126–1165, 2014.

[8] Khoa D Doan, Yingjie Lao, Peng Yang, and Ping Li. Defending backdoor attacks on vision transformer via patch processing. *arXiv preprint arXiv:2206.12381*, 2022.

[9] 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. In *International Conference on Learning Representations*. OpenReview.net, 2021.

[10] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *Advances in Neural Information Processing Systems*, pages 26183–26197, 2021.

[11] Jixun Gao and Yuanyuan Zhao. Tfe: A transformer architecture for occlusion aware facial expression recognition. *Frontiers in Neurorobotics*, 15:763100, 2021.

[12] Xueluan Gong, Yanjiao Chen, Jianshuo Dong, and Qian Wang. ATTEQ-NN: Attention-based QoE-aware evasive backdoor attacks. In *Annual Network and Distributed System Security Symposium*. The Internet Society, 2022.

[13] Xueluan Gong, Yanjiao Chen, Qian Wang, Huayang Huang, Lingshuo Meng, Chao Shen, and Qian Zhang. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. *IEEE Journal on Selected Areas in Communications*, 39(8):2617–2631, 2021.

[14] Xueluan Gong, Yanjiao Chen, Wang Yang, Qian Wang, Yuzhe Gu, Huayang Huang, and Chao Shen. Redeem myself: Purifying backdoors in deep learning models using self attention distillation. In *IEEE Symposium on Security and Privacy*, 2023.

[15] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12094–12103, 2022.

[16] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[17] Yujie Ji, Xinyang Zhang, and Ting Wang. Backdoor attacks against learning systems. In *Conference on Communications and Network Security*, pages 1–9. IEEE, 2017.

[18] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[20] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[21] Shaofeng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 2020.

[22] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020.

[23] Cong Liao, Haoti Zhong, Anna Squicciarini, Sencun Zhu, and David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *arXiv preprint arXiv:1808.10307*, 2018.

[24] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Annual Network and Distributed System Security Symposium*. The Internet Society, 2018.

[25] Peizhuo Lv, Hualong Ma, Jiachen Zhou, Ruigang Liang, Kai Chen, Shengzhi Zhang, and Yunfei Yang. Dbia: Data-free backdoor injection attack against transformer networks. *arXiv preprint arXiv:2111.11870*, 2021.

[26] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "beatrix" resurrections: Robust backdoor detection via gram matrices. In *Annual Network and Distributed System Security Symposium*. The Internet Society, 2023.

[27] Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *Annual Conference on Neural Information Processing Systems*, 2020.

[28] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

[29] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex Liu, and Ting Wang. A tale of evil twins: Adversarial inputs versus poisoned models. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 85–99, 2020.

[30] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

[31] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *AAAI Conference on Artificial Intelligence*, pages 11957–11965. AAAI Press, 2020.

[32] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *IEEE 7th European Symposium on Security and Privacy*, pages 703–718, 2022.

[33] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. Computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

[34] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[35] Akshayvarun Subramanya, Aniruddha Saha, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Backdoor attacks on vision transformers. *arXiv preprint arXiv:2206.08477*, 2022.

[36] 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.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[38] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*, pages 707–723, 2019.

[39] Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L Zhang. Vit-cx: Causal explanation of vision transformers. *arXiv preprint arXiv:2211.03064*, 2022.

[40] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 5824–5836, 2020.

[41] 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 *IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[43] Mengxin Zheng, Qian Lou, and Lei Jiang. Trojvit: Trojan insertion in vision transformers. *arXiv preprint arXiv:2208.13049*, 2022.

[44] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.

APPENDIX

A. THEORETICAL ANALYSIS OF ATTENTION DIFFUSION LOSS EFFECT

For convenience, we denote $\widetilde{A}(l_k)_{tg}^{(0,i)}$ as $\widetilde{A}^{(i)}$, which represents the importance scores of the i-th token for the target label [1]. These scores reflect the contribution of the token towards the activation value y_{tg} of the target label, without considering its impact on other labels (y_{cl}, where $cl \neq tg$). As the activation value of the target label is positively correlated with the probability of the model being classified as the target class, it also indicates the effectiveness of the backdoor attack.

Let q represent the number of tokens in the attention diffusion area, p denote the total number of tokens in the input sequence, and m indicate the number of tokens in the poisoned area. In MEGATRON, the value of p is much larger than q to achieve the concealment goal. When the attention diffusion area q is equal to the poisoned area m, we refer to the attention diffusion loss as $\mathcal{L}_{ro}^{(m)}$, which satisfies the following equations:

$$\mathcal{L}_{ro}^{(m)} = \sum_{i=1}^{m}(1 - \widetilde{A}^{(i)}) + \sum_{i=m+1}^{p-1} \widetilde{A}^{(i)}. \tag{19}$$

$\widetilde{A}^{(i)}$ has a positive contribution to the activation value of the poisoned area and a negative contribution to the non-poisoned area, as the coefficient in the poisoned area is 1 and -1 in the non-poisoned area. Therefore, $-\mathcal{L}_{ro}^{(m)}$ is positively correlated with the effectiveness of MEGATRON. During the training process, $\mathcal{L}_{ro}^{(m)}$ is minimized, meaning it will be maximized within the diffusion range and minimized outside the diffusion range.

We evaluate two backdoor injecting training losses, namely $\mathcal{L}_{CE} + \mathcal{L}_{LA}$ and $\mathcal{L}_{CE} + \mathcal{L}_{LA} + \mathcal{L}_{ro}$. After training, we denote $\widetilde{A}^{(i)}$ as $\widetilde{A_\Delta}^{(i)}$ when using $\mathcal{L}_{ro}$, and as $\widetilde{A_0}^{(i)}$ when not using $\mathcal{L}_{ro}$. Similarly, we denote the loss $\mathcal{L}_{ro}^{(m)}$ as $\mathcal{L}_{ro\Delta}^{(m)}$ when using $\mathcal{L}_{ro}$, and as $\mathcal{L}_{ro0}^{(m)}$ when not using $\mathcal{L}_{ro}$, respectively.

$$\begin{aligned}\mathcal{L}_{ro\Delta}^{(m)} &= \sum_{i=1}^{m}(1 - \widetilde{A_\Delta}^{(i)}) + \sum_{i=m+1}^{p-1}(\widetilde{A_\Delta}^{(i)}),\\ \mathcal{L}_{ro0}^{(m)} &= \sum_{i=1}^{m}(1 - \widetilde{A_0}^{(i)}) + \sum_{i=m+1}^{p-1}(\widetilde{A_0}^{(i)}),\end{aligned} \tag{20}$$

TABLE XV. APPLY SAGE TO MEGATRON.

	Original		SAGE	
	CDA	ASR	CDA	ASR
CIFAR-10	97.25%	100%	96.74%	100%
GTSRB	97.54%	100%	97.77%	99.3%
CIFAR-100	90.38%	99.3%	90.43%	99.4%
Tiny ImageNet	88.60%	100%	89.73%	100%

when $q \geq m$,

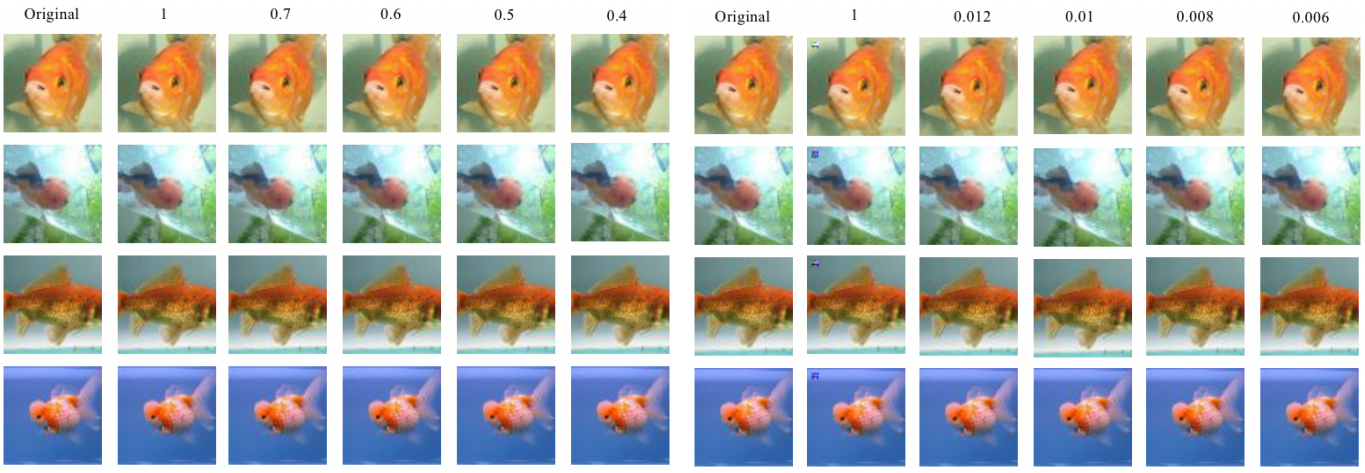
$$\begin{aligned}\mathcal{L}_{ro\Delta}^{(m)} - \mathcal{L}_{ro0}^{(m)} &= (m - \sum_{i=1}^{m}\widetilde{A_\Delta}^{(i)} + \sum_{i=m+1}^{p-1}\widetilde{A_\Delta}^{(i)})\\ &\quad -(m - \sum_{i=1}^{m}\widetilde{A_0}^{(i)} + \sum_{i=m+1}^{p-1}\widetilde{A_0}^{(i)})\\ &= [\sum_{i=1}^{m}(\widetilde{A_0}^{(i)} - \widetilde{A_\Delta}^{(i)}) - \sum_{i=q+1}^{p-1}(\widetilde{A_0}^{(i)} - \widetilde{A_\Delta}^{(i)})]\\ &\quad +(\sum_{i=m+1}^{q}\widetilde{A_\Delta}^{(i)} - \sum_{i=m+1}^{q}\widetilde{A_0}^{(i)}))\\ &< 0 + (\sum_{i=m+1}^{q}\widetilde{A_\Delta}^{(i)} - \sum_{i=m+1}^{q}\widetilde{A_0}^{(i)})\\ &= \sum_{i=m+1}^{q}\widetilde{A_\Delta}^{(i)} - \sum_{i=m+1}^{q}\widetilde{A_0}^{(i)},\end{aligned} \tag{21}$$

when $q = m$, the upper limit of of E.q.(22) is 0, i.e., $\mathcal{L}_{ro\Delta}^{(m)} < \mathcal{L}_{ro0}^{(m)}$. To this end, when the diffusion area is controlled within the poisoning area, the attention diffusion loss will enhance the poisoning effect.

When $q > m$ and the attention diffusion area is not too large, although the upper bound of $\mathcal{L}_{ro\Delta}^{(m)} - \mathcal{L}_{ro0}^{(m)}$ is positive, $\mathcal{L}_{ro\Delta}^{(m)}$ is generally smaller than $\mathcal{L}_{ro0}^{(m)}$ due to the negative term in the square brackets. However, the enhanced poisoning effect will have a negative correlation with the diffusion area, thus if the attention diffusion area becomes too large, it will result in a decrease in the attack success rate. In MEGATRON, we restrict the attention diffusion area to be less than three times the size of the trigger, which allows the attention diffusion loss to enhance the attack performance of MEGATRON.

TABLE XVI. IMPACT OF β ON MEGATRON.

	Value	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	10	94.90%	100%	55.0088729	0.9999999	0.0008298	0.0003717
	40	95.20%	100%	55.0097682	0.9999999	0.0008301	0.0003756
	70	95.30%	100%	55.0167870	0.9999999	0.0008298	0.0003725
	100	95.83%	100%	55.0166671	0.9999999	0.0008341	0.0003719
	130	95.86%	100%	55.1045761	0.9999999	0.0008801	0.0003711
GTSRB	10	98.05%	100%	55.2313058	0.9999999	0.0004552	0.0003766
	40	97.77%	100%	55.2199656	0.9999999	0.0004583	0.0003717
	70	97.00%	100%	55.2177083	0.9999999	0.0004582	0.0003703
	100	97.54%	100%	55.2198875	0.9999999	0.0004583	0.0003710
	130	97.64%	100%	55.2209444	0.9999999	0.0004601	0.0003720
CIFAR-100	10	90.60%	100.0%	55.6559779	0.9999999	0.0007394	0.0003516
	40	90.28%	100%	55.6946467	0.9999999	0.0007367	0.0003517
	70	90.34%	98.3%	55.6787704	0.9999999	0.0007353	0.0003526
	100	90.98%	98.5%	55.6639872	0.9999999	0.0007575	0.0003519
	130	89.80%	99.2%	55.6657994	0.9999999	0.0007292	0.0003522
Tiny ImageNet	10	88.82%	100%	58.7352687	1.00000000	0.0007569	0.0003550
	40	88.85%	100%	58.3561098	0.99999994	0.0007718	0.0003687
	70	88.82%	100%	58.7353841	1.00000000	0.0007437	0.0003571
	100	88.40%	100%	58.7353548	1.00000000	0.0007348	0.0003549
	130	88.23%	100%	58.7351349	1.00000000	0.0007528	0.0003578



(a) ϕ_A (b) ϕ_D

Fig. 4. Impact of transparency value of ϕ_A and ϕ_D.

TABLE XVII. IMPACT OF ϕ_A TRANSPARENCY VALUE ON MEGATRON.

	Transparency	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	1	95.83%	100%	53.7255049	0.99999996	0.0014266	0.0003885
	0.70	96.11%	100%	54.1449071	0.99999993	0.0009893	0.0003764
	0.60	95.72%	100%	54.5706415	0.99999991	0.0009386	0.0003742
	0.50	96.10%	100%	55.0166671	0.99999992	0.0007341	0.0003720
	0.40	95.62%	100%	55.4450206	0.99999990	0.0005446	0.0003697
	0.10	95.20%	98.2%	57.2244942	0.99999994	0.0001955	0.0003558
	0.09	96.00%	98.8%	57.2372125	0.99999997	0.0001925	0.0003555
GTSRB	1	97.36%	99.1%	54.1896479	0.99999994	0.0011897	0.0003855
	0.70	97.14%	100%	54.5136270	0.99999992	0.0005438	0.0003739
	0.60	97.59%	100%	54.8776579	0.99999996	0.0005011	0.0003722
	0.50	97.78%	99.8%	55.2198875	0.99999991	0.0004583	0.0003703
	0.40	97.23%	99.2%	55.5620041	0.99999994	0.0004231	0.0003686
	0.10	98.00%	98.4%	57.2244939	0.99999994	0.0003806	0.0003558
	0.09	97.54%	99.0%	57.2372123	0.99999995	0.0002797	0.0003555
CIFAR-100	1	90.24%	97.8%	53.7227042	0.99999986	0.0010778	0.0004402
	0.70	90.15%	99.0%	53.9552315	0.99999830	0.0005914	0.0003910
	0.60	90.98%	100%	54.2197950	0.99999555	0.0005163	0.0003882
	0.50	90.28%	100%	55.4692161	0.99999186	0.0005399	0.0003787
	0.40	90.28%	100%	55.4692161	0.99999184	0.0005399	0.0003753
	0.10	90.34%	93.2%	57.5794628	0.99999992	0.0003796	0.0003355
	0.09	90.37%	88.0%	57.5561912	0.99999994	0.0003617	0.0003364
Tiny ImageNet	1	88.40%	100%	54.2889001	0.99999999	0.0011395	0.0004089
	0.70	88.21%	100%	54.6626342	0.99999998	0.0007380	0.0003676
	0.60	88.83%	100%	55.0401141	0.99999995	0.0007490	0.0003662
	0.50	88.61%	100%	55.4485169	0.99999995	0.0007650	0.0003648
	0.40	88.87%	100%	55.8172215	0.99999997	0.0007060	0.0003633
	0.10	92.80%	96.6%	64.8964024	0.99935018	0.0004527	0.0003233
	0.09	88.04%	80.4%	64.9694100	0.99935439	0.0004020	0.0003225

TABLE XVIII. IMPACT OF ϕ_D TRANSPARENCY VALUE ON MEGATRON. WE USE MASK CUTTING METHOD.

	Transparency	CDA	ASR	PSNR	SSIM	LPIPS	L_1 distance
CIFAR-10	1	96.20%	100%	32.1189018	0.9999540	0.0426396	0.0063316
	0.014	96.13%	100%	53.0617543	0.9999998	0.0011747	0.0005131
	0.012	96.67%	100%	54.0109851	0.9999999	0.0009080	0.0004426
	0.010	96.10%	100%	55.0166671	0.9999999	0.0008341	0.0003720
	0.008	96.17%	100%	56.0580054	0.9999999	0.0007460	0.0003014
	0.006	95.03%	100%	57.0858650	1.0000000	0.0005378	0.0002308
	0.004	95.47%	99.2%	58.0049512	1.0000000	0.0004935	0.0001602
	0.002	96.74%	96.3%	58.6671051	1.0000000	0.0003644	0.0000896
GTSRB	1	97.36%	100%	32.8754410	0.9999632	0.0349158	0.0058047
	0.014	97.50%	99.1%	53.1902194	0.9999997	0.0009851	0.0005114
	0.012	97.63%	99.3%	54.1714683	0.9999999	0.0006795	0.0004408
	0.010	97.78%	99.3%	55.2198875	0.9999999	0.0004583	0.0003703
	0.008	97.68%	99.0%	56.6530311	0.9999999	0.0003068	0.0002997
	0.006	97.09%	99.4%	57.4115107	1.0000000	0.0002015	0.0002291
	0.004	97.86%	97.2%	58.4107401	1.0000000	0.0001353	0.0001586
	0.002	97.77%	88.5%	59.1425856	1.0000000	0.0001015	0.0000880
CIFAR-100	1	90.67%	100%	33.4323787	0.9999663	0.0315708	0.0050570
	0.014	90.13%	99.0%	53.8097414	0.9999998	0.0012601	0.0004752
	0.012	90.15%	100%	54.8319848	0.9999999	0.0010127	0.0004115
	0.010	90.22%	100%	55.3891475	0.9999870	0.0007818	0.0003293
	0.008	90.34%	100%	56.2606889	0.9999942	0.0005728	0.0003372
	0.006	90.08%	98.3%	56.2939643	0.9999981	0.0004185	0.0002984
	0.004	90.65%	96.2%	58.3224578	0.9999995	0.0003047	0.0002528
	0.002	90.26%	56.6%	60.2280901	1.0000000	0.0001850	0.0001202
Tiny ImageNet	1	88.33%	100%	32.6575640	0.9999583	0.0345759	0.0057272
	0.014	88.81%	100%	53.3894265	0.9999999	0.0007883	0.0005052
	0.012	88.40%	100%	54.3677285	1.0000000	0.0007022	0.0004350
	0.010	88.46%	100%	55.4485362	0.9999999	0.0006254	0.0003648
	0.008	88.63%	96.2%	58.5090229	0.9999999	0.0005786	0.0002874
	0.006	88.20%	86.6%	60.4668690	0.9989846	0.0004922	0.0002177
	0.004	88.68%	86.7%	62.7646520	0.9992224	0.0002844	0.0001476
	0.002	88.41%	64.4%	65.1262551	0.9994777	0.0001300	0.0000774