Supplemental Material

TIJO: Trigger Inversion with Joint Optimization for Defending Multimodal Backdoored Models

Indranil Sur[1]* Karan Sikka[1] Matthew Walmer[2] Kaushik Koneripalli[1]
Anirban Roy[1] Xiao Lin[1] Ajay Divakaran[1] Susmit Jha[1]

[1]SRI International [2]University of Maryland

A. Implementation Details

Trigger Inversion Stage: We set the maximum optimization step T to 15. We select the NLP trigger inversion trigger length, *i.e.* length of $\boldsymbol{t}_{adv}$, to 1. $\boldsymbol{t}_{adv}$ is initialized as the 0^{th} token in the vocabulary $\mathcal{V}_f$ *i.e.*, for Efficient BUTD models [4] we use the 'what' token, and for OpenVQA models [8] we use the 'PAD' token. The append policy $\mathcal{A}$ simply appends $\boldsymbol{t}_{adv}$ to the start of the question token $\boldsymbol{t}$. For trigger inversion in the feature space, the feature trigger $\boldsymbol{f}_{adv}$ is initialized from a continuous uniform distribution in interval $[0, 1)$. The feature overlay policy $\mathcal{B}$ adds $\boldsymbol{f}_{adv}$ to all the 36 box features extracted from the detector $\mathcal{D}$. $\boldsymbol{f}_{adv}$ is optimized with Adam optimizer with a learning rate of 0.1 and beta as (0.5, 0.9). We set $\boldsymbol{f}_{adv}$ L2 regularization λ to 0.

Image Patch Inversion Stage: We optimize for $\boldsymbol{p}_{adv}$ of size 64 × 64 initialized with $\mathbf{0}$s. $\mathcal{M}$ overlays the patch on the center of the image with the patch scaled to 10% of the smallest length of the image. We optimize $\boldsymbol{p}_{adv}$ with Adam optimizer with a learning rate of 0.03, and betas as (0.5, 0.9). We use early stopping with a patience of 20 epochs. After each update, $\boldsymbol{p}_{adv}$ is normalized to be in the range [0,1]. We optimize only over the clean images from the support set $\mathcal{S}$.

B. Baseline Details

Weight Analysis: Weight analysis [1] is a generalist backdoor detection method that makes no assumption on the nature of the backdoor. Instead, empirical analysis of the model weights is used to determine if the model is backdoored or benign. We follow the same setup as [6], *i.e.* we bin the weights of the final layer based on their magnitude and generate a histogram-based feature vector. We then train a logistic regression classifier on these histogram features and report the AUC on each *TrojVQA* split.

DBS: Dynamic Bound-Scaling (DBS) [5] is a trigger inversion-based backdoor defense for NLP tasks. As the tokens are discrete in nature, they formulate the optimization problem to gradually converge to the ground truth trigger, which is denoted as a one-hot vector in the convex hull of embedding space $\mathcal{E}_f$. They also dynamically reduce (and in some cases roll back) the temperature coefficient of the final softmax to not let the optimization get stuck in local minima. We have used the same configurations as stated in [5], though we set the max optimization steps to 100 instead of 200. We have observed our method converges much faster in about $10-15$ optimization steps, while DBS takes $80-100$ steps, with each optimization step roughly the same in both cases. Also, DBS fails to detect backdoored BUTD$_e$ [4] VQA models.

NC & TABOR: Both Neural Cleanse (NC) [7] and TABOR [3] are trigger inversion-based backdoor defenses for image classification task. NC is the first work to formalize Trojan detection as a non-convex optimization problem. As shown in [3], NC fails if the backdoored model is triggered with triggers of varying size, shape, and location. TABOR extends NC with a new regularization to constrain the adversarial sample subspace based on explainable AI attribution features and other heuristics. Adapting NC and TABOR to *TrojVQA* models required some methodological adjustments. They both are trigger inversion methods for image classification models, which have a much simpler architecture than detector models–that serve as the visual backbone of VQA models. Specifically, image classification models assume a fixed image size. For the reported results, we have fixed the image size to 300 × 300. Even though $\mathcal{D}$ can handle images of arbitrary sizes, we resize the

[*] Corresponding author: indranil.sur@sri.com

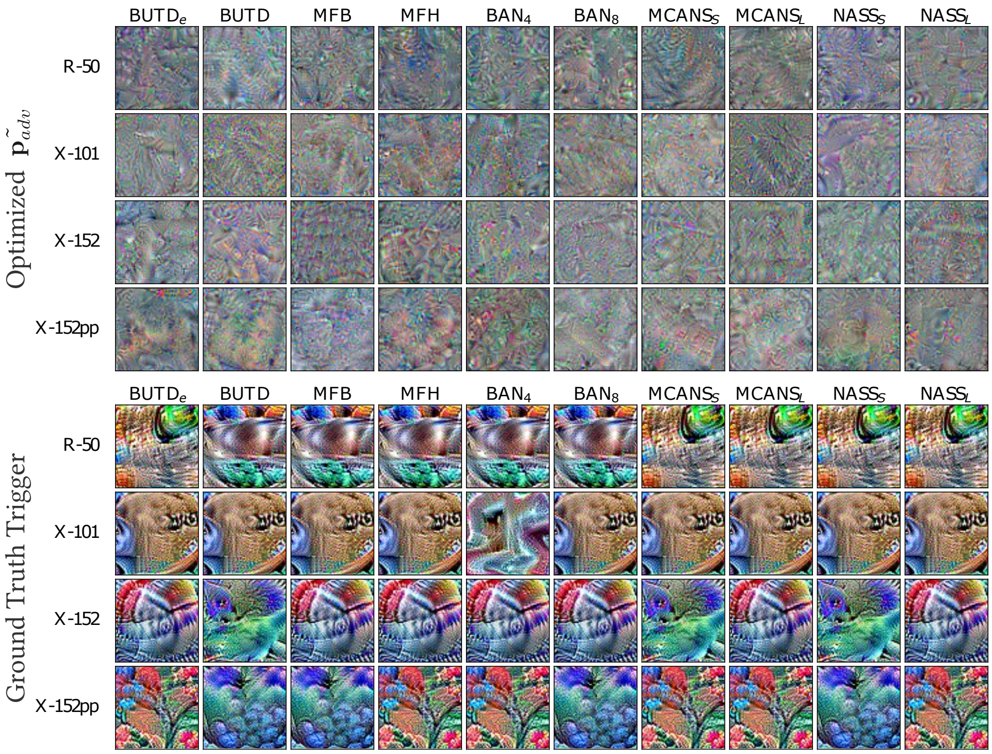

Figure 1: Visualizes the generated image patches $\boldsymbol{p}_{\tilde{adv}}$ from $\boldsymbol{f}_{\tilde{adv}}$ using the trigger patch generation method. Here we show inversion across the different combinations of detector backbones and VQA architectures for backdoored models (shown above) of the $\mathcal{T}_{nlp+O}$ split, along with the corresponding ground truth triggers (shown below) for comparison.

images to the fixed input size for NC and TABOR to work. The patch and mask span the entire image and hence are set to 300 × 300. The max optimization step is set to 25. For TABOR, we have set $\lambda_1 = 10^{-8}$, $\lambda_2 = 10^{-7}$, $\lambda_3 = 10^{-9}$, and $\lambda_4 = 10^{-10}$, which we have found is dependent on the size of the image.

C. Additional Results

C.1. Design of Shallow Classifiers

We used Logistic Regression (LR) as the shallow classifier and find it to outperform simple rule-based detector. For example, in (TIJO$_{mm}$, $\mathcal{T}$) setting, we get an accuracy of $\mathbf{0.856}_{\pm 0.03}$ with optimal threshold for LR, which is higher than the best accuracy **0.816** of the simple rule-based detector (obtained by varying the threshold $\in [0, 1]$ with 0.01 increments). This intuitively makes sense since (**??**) different VQA architectures have different TI loss range. We choose LR over other classifiers as it generally outperformed other methods and is faster. For example, in the (TIJO$_{mm}$, $\mathcal{T}$) case, we get AUC of $\mathbf{0.924}_{\pm 0.016}$ for LR, $\mathbf{0.923}_{\pm 0.016}$ for SVM (RBF kernel), $\mathbf{0.915}_{\pm 0.019}$ for XGBoost (max depth of 2) and $\mathbf{0.876}_{\pm 0.034}$ for Random-Forest.

C.2. Inverted NLP Triggers

The inverted NLP triggers ($\boldsymbol{t}_{\tilde{adv}}$) generally match the ground-truth NLP triggers ($\boldsymbol{t}_t$). We observe a match accuracy of **0.95** in the (TIJO$_{nlp}$, $\mathcal{T}_{nlp}$) case and **0.756** in the (TIJO$_{mm}$, $\mathcal{T}$) case. Here are few examples of mismatch between the predicted and target triggers ($\boldsymbol{t}_t \rightarrow \boldsymbol{t}_{\tilde{adv}}$): (1) similar to target: diseases → disease, ladder → ladders, decoys → decoy, (2) semantically close to target: potholders → hotpads, terrifying → horrifying, (3) completely different from target: midriff → 4:50, stool → nasa.

C.3. Image Patch Generation

Figure 1 shows the generated patches for backdoored VQA models of $\mathcal{T}_{nlp+O}$ split for different combinations of detector backbones and VQA architectures. These results are in addition to those presented in **??**. We see a similar pattern as reported in the main paper where we see some similarity between the ground-truth triggers and the reconstructed triggers for a detector backbone. However, we additionally observe two differences- (1) reconstructed triggers change for different types of VQA architectures for a fixed backbone, and (2) there are cases where the similarity between ground-truth and reconstructed triggers are weak

	$\mathcal{T}_{solid}$		$\mathcal{T}_{optim}$		$\mathcal{T}_{nlp+S}$		$\mathcal{T}_{nlp+O}$	
VQA Arch	$\tilde{\boldsymbol{f}}_{adv}$	$\tilde{\boldsymbol{p}}_{adv}$	$\tilde{\boldsymbol{f}}_{adv}$	$\tilde{\boldsymbol{p}}_{adv}$	$\tilde{\boldsymbol{f}}_{adv}$	$\tilde{\boldsymbol{p}}_{adv}$	$\tilde{\boldsymbol{f}}_{adv}$	$\tilde{\boldsymbol{p}}_{adv}$
BUTD$_e$	$1.00_{\pm 0.00}$	$0.01_{\pm 0.02}$	$1.00_{\pm 0.00}$	$0.06_{\pm 0.15}$	$0.94_{\pm 0.04}$	$0.24_{\pm 0.15}$	$0.95_{\pm 0.06}$	$0.27_{\pm 0.08}$
BUTD	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.99_{\pm 0.02}$	$0.01_{\pm 0.03}$	$0.84_{\pm 0.25}$	$0.11_{\pm 0.18}$	$0.96_{\pm 0.06}$	$0.03_{\pm 0.04}$
MFB	$0.99_{\pm 0.02}$	$0.01_{\pm 0.02}$	$1.00_{\pm 0.00}$	$0.01_{\pm 0.02}$	$0.76_{\pm 0.36}$	$0.04_{\pm 0.06}$	$0.98_{\pm 0.03}$	$0.06_{\pm 0.08}$
MFH	$1.00_{\pm 0.00}$	$0.01_{\pm 0.02}$	$0.99_{\pm 0.02}$	$0.00_{\pm 0.00}$	$0.88_{\pm 0.24}$	$0.10_{\pm 0.12}$	$0.64_{\pm 0.34}$	$0.01_{\pm 0.02}$
BAN$_4$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.69_{\pm 0.41}$	$0.10_{\pm 0.16}$	$0.88_{\pm 0.15}$	$0.00_{\pm 0.00}$
BAN$_8$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.83_{\pm 0.33}$	$0.00_{\pm 0.00}$	$0.96_{\pm 0.06}$	$0.17_{\pm 0.25}$
MCANS$_S$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.50_{\pm 0.25}$	$0.00_{\pm 0.00}$	$0.32_{\pm 0.28}$	$0.00_{\pm 0.00}$
MCANS$_L$	$0.99_{\pm 0.02}$	$0.01_{\pm 0.03}$	$1.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.61_{\pm 0.25}$	$0.07_{\pm 0.18}$	$0.52_{\pm 0.25}$	$0.01_{\pm 0.02}$
NASS$_S$	$0.91_{\pm 0.09}$	$0.01_{\pm 0.02}$	$0.94_{\pm 0.11}$	$0.04_{\pm 0.12}$	$0.42_{\pm 0.27}$	$0.00_{\pm 0.00}$	$0.41_{\pm 0.26}$	$0.07_{\pm 0.20}$
NASS$_L$	$0.91_{\pm 0.10}$	$0.00_{\pm 0.00}$	$0.93_{\pm 0.13}$	$0.04_{\pm 0.08}$	$0.26_{\pm 0.26}$	$0.00_{\pm 0.00}$	$0.29_{\pm 0.25}$	$0.00_{\pm 0.00}$

Table 1: Inverse Attack Success Rate (Inv-ASR) of optimized reconstructed triggers when re-injected into inputs from the support set $\mathcal{S}$. Results are presented separately for each VQA model type, and for all four *TrojVQA* splits that include visual triggers either in a single-key or dual-key configuration. The results show that feature-space inverted triggers are highly effective at activating backdoors as compared to image-space inverted triggers. The effectiveness of feature-space triggers is consistent for uni-modal triggers, but varies by model types for dual-key triggers.

(*e.g*. for R-50 and NASSs). This highlights that our inversion process is able to adjust to the changes in the ground-truth trigger and is not dependent only on the visual backbone.

C.4. Inv-ASR for Reconstructed Visual Trigger

We summarize results for the Inverse Attack Success Rate (Inv-ASR) of reconstructed visual triggers in Table 1. This includes results for both detector feature-space inverted triggers, $\tilde{\boldsymbol{f}}_{adv}$, and image-space inverted trigger patches, $\tilde{\boldsymbol{p}}_{adv}$. These results are shown for the four *TrojVQA* splits that include any visual triggers. This includes both dual-key splits and single-visual-key splits. The Inv-ASR metric measures the fraction of triggered inputs for which the backdoor successfully activates and changes the model output to the target answer. $\tilde{\boldsymbol{p}}_{adv}$ triggers are overlaid on the clean images with $\mathcal{M}$, while $\tilde{\boldsymbol{f}}_{adv}$ are overlayed directly into the detector output features with $\mathcal{B}$. For the dual-key backdoored models, we also add the corresponding text trigger $\tilde{\boldsymbol{t}}_{adv}$ with $\mathcal{A}$.

We find that the feature-space inverted triggers lead to a very high Inv-ASR for visual-trigger-only backdoored models. These scores are often at or near 1.00 consistent activation of the backdoor. For dual-key splits, where a language-space trigger is also included, feature-space reconstructed triggers typically achieve a high Inv-ASR, though this varies greatly by the VQA model type, with BUTD$_e$ having the highest average Inv-ASR values over 0.9 and NASS$_L$ having the lowest Inv-ASR values under 0.3. These results show that feature-space reconstructed triggers can be an effective tool to identify backdoored models with uni-model image-space triggers, and can also be effective for some types of dual-key backdoored models.

	Replace %		
FRR	70% tokens	50% tokens	30% tokens
0.5%	$97.55_{\pm 3.37}$	$93.88_{\pm 4.74}$	$94.71_{\pm 3.29}$
1%	$95.11_{\pm 3.60}$	$88.55_{\pm 4.92}$	$94.71_{\pm 3.36}$
5%	$86.88_{\pm 6.61}$	$74.11_{\pm 6.47}$	$80.45_{\pm 6.18}$
10%	$77.11_{\pm 6.24}$	$64.55_{\pm 6.65}$	$67.01_{\pm 6.66}$

Table 2: False Acceptance Rate (FAR) for different False Rejection Rates (FRR).

Meanwhile, the Inv-ASR scores for image-space reconstructed triggers are very low, typically near 0.0, indicating that they are not effective at activating the backdoor trigger in these Trojaned models. This result stems from the known challenges of reconstructing image-space triggers highlights the benefits of performing feature-space trigger reconstruction instead. However, we do observe some cases where the reconstructed trigger is able to provide non-zero Inv-ASR, *e.g*. mean of 0.24 & 0.27 in BUTD$_e$ models on $\mathcal{T}_{nlp+S}$ & $\mathcal{T}_{nlp+O}$. We thus argue that the reconstruction of triggers in the image-space needs further research.

D. Online Mutimodal Defense Analysis

STRIP-ViTA: STRIP-ViTA [2] showed defense in multiple domains against backdoor attacks in an *online setting*. Backdoor defense in an online setting is simpler where we assume that the given model is backdoored and focuses on identifying whether the given input is clean or poisoned. It is different from the offline setting where with only a few clean examples we determine if a model is backdoored or benign. Hence STRIP-ViTA is not directly comparable to our method. We conducted experiments with STRIP-ViTA to access the difficulty of detecting the multimodal triggers used in our evaluation. STRIP-ViTA perturbs the given input text and image, builds a distribution of entropies for both clean and poison inputs, and then sets a threshold of entropy for detecting whether an incoming input is clean or poisoned. For the image modality, the perturbation is made by randomly selecting an image from the dataset and doing a weighted combination with the original image. For the text modality, a fraction of the words in the input text is replaced. We conduct experiments by sweeping across 3 different text-replacement percentages (70%, 50%, and 30%) on dual-key backdoored *TrojVQA* models and results are provided in Table 2. This table shows the False Acceptance Rates (FAR) at different percentages of fixed False Rejection Rates (FRR). Our results demonstrate that online detection of these triggers is also very challenging, and the FAR remains very high (67%) even for a considerably high FRR (10%).

References

[1] Greg Fields, Mohammad Samragh, Mojan Javaheripi, Farinaz Koushanfar, and Tara Javidi. Trojan signatures in dnn weights. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–20, 2021. 1

[2] Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyoungshick Kim. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *IEEE Transactions on Dependable and Secure Computing*, 19(4):2349–2364, 2021. 3

[3] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. In *Proceedings of The 20th IEEE International Conference on Data Mining (ICDM), 2020*. IEEE, 2020. 1

[4] Hengyuan Hu, Alex Xiao, and Henry Huang. Bottom-up and top-down attention for visual question answering. https://github.com/hengyuan-hu/bottom-up-attention-vqa, 2017. 1

[5] Guangyu Shen, Yingqi Liu, Guanhong Tao, Qiuling Xu, Zhuo Zhang, Shengwei An, Shiqing Ma, and Xiangyu Zhang. Constrained optimization with dynamic bound-scaling for effective nlp backdoor defense. In *International Conference on Machine Learning*, pages 19879–19892. PMLR, 2022. 1

[6] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15375–15385, 2022. 1

[7] 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 *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019. 1

[8] Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. Openvqa. https://github.com/MILVLG/openvqa, 2019. 1