PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Mask-Guided Hybrid Triggers for Robust Clean-Label Backdoor Attacks

Shengye Pang, **Xiangyu Ji**, **Jungang Yang**, **Song Yang** and **Guobing Zou***
School of Computer Engineering and Science, Shanghai University, Shanghai, China
{pangsy, jixiangyu, yangjungang, yangsong, gbzou}@shu.edu.cn

Abstract

Clean-label backdoor attacks pose significant security threats to deep neural networks by injecting triggers without altering ground-truth labels. However, existing methods face a fundamental dilemma: sample-agnostic triggers are robust but easily detectable, while sample-specific triggers offer superior stealthiness but suffer from limited effectiveness due to feature suppression. To bridge this gap, we propose a new backdoor trigger framework called Mask-Guided Hybrid Trigger (MGHT). MGHT uses an adaptive mask to allocate spatial regions of the hybrid trigger between a sample-agnostic anchor for reliable memorization and a sample-specific camouflage for perceptual and semantic consistency. To prevent the optimization from greedily relying on a single trigger component, we further propose a Synergy-driven Co-optimization Strategy with a margin-based Synergy Loss. This ensures that the hybrid trigger is more effective and robust than either component alone. Extensive experiments on benchmark datasets demonstrate that MGHT achieves competitive performance, attaining over 99% ASR on CIFAR-10 and CelebA and showing strong effectiveness on higher-resolution benchmarks, while maintaining high visual quality (PSNR $>$ 30 dB) and robustness to mainstream backdoor defenses.

1 Introduction

The remarkable success of deep neural networks (DNNs) relies heavily on large-scale, high-quality datasets. In practice, model developers often collect data from open-source platforms or third-party providers. However, this lack of transparency in the data collection stage exposes DNNs to severe security threats, including adversarial attacks [Madry *et al.*, 2018; Goodfellow *et al.*, 2015] and backdoor attacks [Li *et al.*, 2024]. Backdoor attacks are especially dangerous because an attacker can inject a hidden trigger into only a small fraction of the training set, causing the trained model to behave normally on benign inputs while consistently misclassifying any input containing the trigger as the target class.

Early studies primarily focused on dirty-label backdoor attacks, where the attacker modifies both the image and its label. Although effective, dirty-label attacks are vulnerable to manual inspection [Turner *et al.*, 2019] due to the label inconsistencies of poisoned samples. To overcome this limitation, recent research has focused on more stealthy and practical clean-label backdoor attacks. In a clean-label setting, the attacker injects triggers without altering the ground-truth labels. This means the victim model is required to learn the association between the trigger and the target class in the presence of strong target-class features, making trigger injection significantly more challenging.

Existing clean-label attacks mainly follow two strategies for trigger construction: sample-agnostic (static) triggers and sample-specific (dynamic) triggers. However, both approaches face intrinsic limitations under the strict clean-label constraint. Sample-agnostic triggers, typically implemented as fixed patches or global patterns [Saha *et al.*, 2020; Souri *et al.*, 2022; Zeng *et al.*, 2023], provide a stable, consistent feature that is easy for the model to memorize, resulting in high attack success rates. While these trigger patterns are imperceptible during the training phase, they often become visually obvious in the test phase or require amplification to be more effective. In contrast, sample-specific triggers generate dynamic perturbations based on the input image, offering superior stealthiness in both the training and testing stages. However, establishing a strong backdoor correlation with dynamic triggers is difficult in clean-label settings. Since the target-class labels are correct, the model tends to learn the robust semantic features of the original object rather than the weak, dispersed dynamic noise. This is known as *feature suppression* [Zhu *et al.*, 2025]. Therefore, injecting sample-specific backdoors in a clean-label setting is challenging.

This reveals a fundamental dilemma in clean-label backdoor learning: sample-agnostic triggers are easy to learn but easy to detect, while sample-specific triggers are stealthy but difficult to learn. Existing methods typically commit to one of these two extremes, lacking a mechanism to combine their complementary strengths. We argue that a robust clean-label attack should not rely solely on one type of pattern but exploit the strengths of both. We propose a novel perspective: utilizing the static pattern as a robust *anchor* to ensure injectabil-

*Corresponding author.

ity, while leveraging dynamic perturbations as *camouflage* to adapt to local image semantics.

Based on this insight, we introduce the **Mask-Guided Hybrid Trigger (MGHT)**, which employs a learnable mask to adaptively partition the image into regions dominated by the sample-agnostic anchor and regions occupied by the sample-specific camouflage. This mask-guided fusion method allows the hybrid trigger to remain both learnable and robust across different images. However, preliminary experiments show that naively combining two triggers yields marginal gains over individual components and can even underperform single triggers on complex datasets. We attribute this degradation to unconstrained optimization, which causes the model to collapse onto a single trigger component. To address this, we design a *Synergy-driven Co-optimization Strategy* that comprehensively coordinates both trigger components via a margin-based constraint. We generate the sample-specific trigger based on the input that already contains the sample-agnostic trigger and introduce a *Synergy Loss* during the joint training process. This mechanism prevents the optimization from collapsing into trivial solutions, ensuring both components make complementary contributions to the backdoor injection. To the best of our knowledge, we are among the first to utilize a mask mechanism to combine these two types of triggers. Our code is available on GitHub[1].

Our main contributions can be summarized as follows:

- We propose Mask-Guided Hybrid Trigger (MGHT), a novel clean-label backdoor attack framework that unifies sample-agnostic and sample-specific triggers through adaptive semantic masks. By dynamically balancing the spatial contribution of sample-agnostic and sample-specific patterns, MGHT effectively mitigates the trade-off between effectiveness and perceptual stealthiness.
- We introduce a Synergy-driven Co-optimization Strategy to jointly optimize the semantic mask, sample-specific trigger, and sample-agnostic trigger. Specifically, the sample-specific trigger generation is directly conditioned on inputs embedded with the sample-agnostic trigger, and a Synergy Loss is employed to explicitly encourage complementary interactions between static and dynamic patterns.
- Extensive experimental results demonstrate that MGHT achieves competitive attack success rates compared to state-of-the-art clean-label backdoor attacks, while preserving high visual fidelity. Moreover, the proposed hybrid trigger exhibits enhanced robustness across diverse image contexts and against potential backdoor defense mechanisms.

2 Related Work

Backdoor Attack. BadNets [Gu *et al.*, 2019] pioneered backdoor attacks by stamping patches and altering labels, and it is categorized as a dirty-label attack. To enhance stealthiness, Turner et al. [2019] introduced the clean-label setting, injecting triggers without label modification. Subsequent works such as HTBA [Saha *et al.*, 2020] and Sleeper Agent [Souri *et al.*, 2022] improved clean-label attacks by optimizing static triggers through feature space minimization and gradient alignment, respectively. The aforementioned methods rely on a static trigger pattern during either the training or testing phase to activate the backdoor. To improve trigger adaptivity, ISSBA [Li *et al.*, 2021b] introduced sample-specific triggers in dirty-label settings. Specifically, a trigger is sample-specific if the trigger pattern varies depending on the input sample. Recent research extends sample-specific triggers to clean-label attacks using two main strategies: image-attribute-based and DNN-based methods. DABA [Xu *et al.*, 2023] constructs image-attribute-based sample-specific triggers by applying fixed augmentation strategies to different channels of the images, while COMBAT [Huynh *et al.*, 2024] trains a generator network using bilevel optimization. Recently, BAAT [Zhu *et al.*, 2025] approaches this problem from a novel perspective by utilizing abstract triggers, such as image styles, to achieve label consistency and adaptivity.

Backdoor Defense. Backdoor defense can be broadly categorized into model-level and data-level approaches. Model-level defenses modify the victim model to eliminate backdoors. Fine-Pruning [Liu *et al.*, 2018] mitigates threats via retraining or pruning-then-fine-tuning on benign data. Others [Yoshida and Fujino, 2020; Li *et al.*, 2021a] employ knowledge distillation [Hinton *et al.*, 2015] for model purification. Neural Cleanse (NC) [Wang *et al.*, 2019] adopts a two-stage strategy: reverse-engineering potential triggers for anomaly detection and suppressing their activation. Data-level defenses filter or repair suspicious samples. During training, methods like DeepSweep [Qiu *et al.*, 2021] use augmentation to disrupt triggers. Inference-time defenses often assume sample-agnostic triggers. For instance, STRIP [Gao *et al.*, 2019] detects poisoned samples by superimposing various image patterns onto the input, while SentiNet [Chou *et al.*, 2020] and Februus [Doan *et al.*, 2020] analyze Grad-CAM [Selvaraju *et al.*, 2017] saliency maps to locate suspicious triggered regions. However, these methods rely heavily on the sample-agnostic trigger assumption. To counter sample-specific attacks, SCALE-UP [Guo *et al.*, 2023] was proposed to detect triggers based on scaled prediction consistency. Since this approach does not rely on trigger pattern assumptions, it is effective against sample-specific triggers.

3 Methodology

3.1 Threat Model

Following prior work on clean-label backdoor attacks [Huynh *et al.*, 2024], we define the threat model based on a scenario where the attacker acts as a third-party data provider with full access to the dataset. In this setting, the attacker releases the dataset through commercial transactions or open-source platforms. We assume a scenario where the attacker can poison the training dataset before release but cannot access the victim model parameters, gradients, or training process. The attacker can train a local surrogate on the released data to optimize the backdoor injection. Specifically, to adhere to the clean-label constraint, the attacker modifies only a subset of samples belonging to a predefined target class without

[1]https://github.com/MAplIle/MGHT

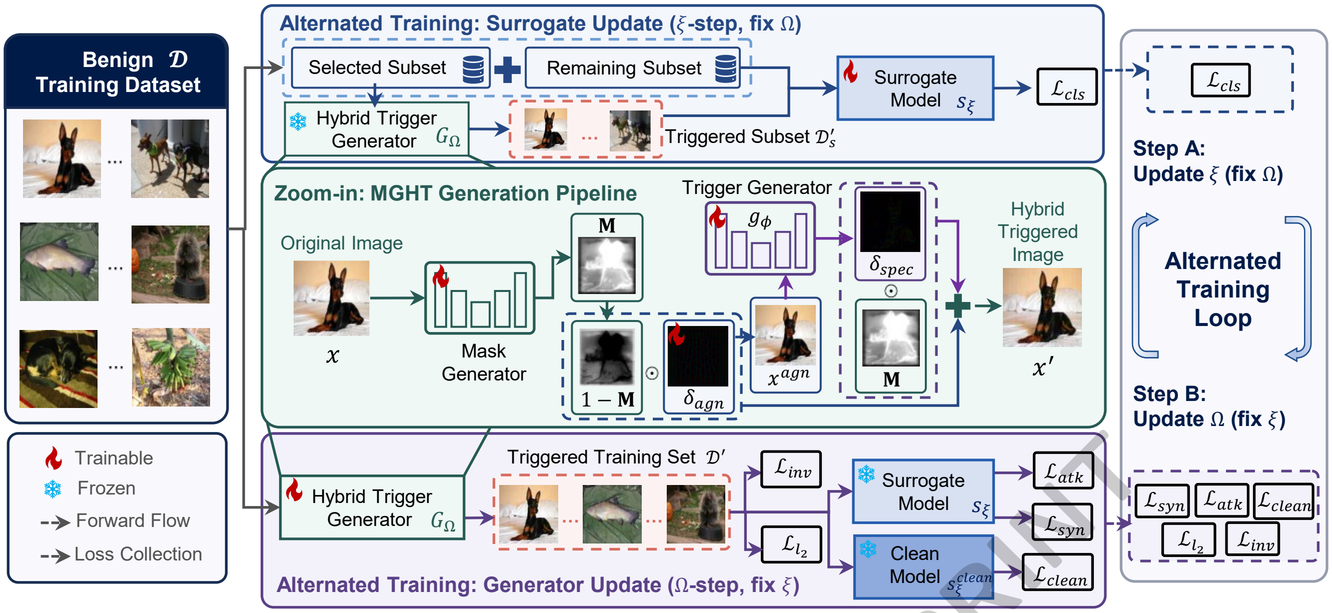

Figure 1: Overview of the proposed Mask-Guided Hybrid Trigger (MGHT) framework.

altering their original labels. The victim subsequently downloads this poisoned dataset to train an image classifier. The attacker's objective is to implant a backdoor such that the infected model behaves normally on benign inputs but misclassifies any input containing the specific trigger as the target label. During the inference phase, the attacker can exploit the trigger injection function to generate triggered samples from any class, thereby manipulating the model's predictions or verifying the dataset's usage.

3.2 Problem Statement

We focus on backdoor attacks against image classification neural networks under the clean-label setting.

Let $f_\theta : \mathcal{X} \to [0,1]^K$ denote the classifier, where θ represents the model parameters, $\mathcal{X} \subseteq \mathbb{R}^d$ is the input data space, and $\mathcal{Y} = \{1, 2, \ldots, K\}$ is the label space.

Given a benign training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, the attacker first selects a target label $y_t \in \mathcal{Y}$ and samples a subset $\mathcal{D}_s \subset \mathcal{D}$ to be poisoned. The poisoning rate is defined as $r = |\mathcal{D}_s|/|\mathcal{D}|$. Specifically, we strictly adhere to the clean-label constraint, ensuring that the ground-truth labels of the poisoned samples belong to the target class, i.e., $\forall (x_k, y_k) \in \mathcal{D}_s, y_k = y_t$.

Let $\mathcal{G}(\cdot)$ denote the trigger injection function. Existing works generally fall into two categories regarding $\mathcal{G}(\cdot)$. Some separate the process into two stages, while others employ a unified generator $G(\cdot)$ for both training and testing. We adopt the latter unified framework, where $\mathcal{G}(\cdot) = G(\cdot)$.

Following the definition in [Li *et al.*, 2021b], we categorize the trigger generator $G(\cdot)$ based on its dependency on the input:

- **Sample-Specific Trigger**: A generator $G(\cdot)$ is sample-specific if and only if for distinct inputs $x_i, x_j \in \mathcal{X}$ ($x_i \neq x_j$), the extracted triggers satisfy $T(G(x_i)) \neq T(G(x_j))$, where $T(\cdot)$ extracts the trigger pattern from the poisoned image.
- **Sample-Agnostic Trigger**: Conversely, $G(\cdot)$ is sample-agnostic if the trigger remains consistent regardless of the input, i.e., $T(G(x_i)) = T(G(x_j))$ for any x_i, x_j.

The victim model $f_{\theta'}$ is trained on the poisoned dataset $\mathcal{D}_p = \mathcal{D}_r \cup \mathcal{D}'_s$, where $\mathcal{D}'_s = \{(G(x), y) | (x, y) \in \mathcal{D}_s\}$ and $\mathcal{D}_r = \mathcal{D} \backslash \mathcal{D}_s$. The parameters θ' are optimized by minimizing the loss function $\mathcal{L}$:

$$\theta' = \arg\min_{\theta} \sum_{(x,y) \in \mathcal{D}_p} \mathcal{L}(f_\theta(x), y) \tag{1}$$

The attacker's goal is to ensure the model behaves normally on benign samples but misclassifies triggered samples as the target label y_t:

$$f_{\theta'}(x_i) = y_i, \quad f_{\theta'}(G(x_i)) = y_t \tag{2}$$

While conventional views often treat sample-agnostic and sample-specific triggers as mutually exclusive, we posit that a robust clean-label attack can benefit from blending both characteristics. To bridge this gap, we formulate the trigger injection function $G(\cdot)$ as a hybrid generator.

Specifically, we define the hybrid trigger as a fusion of a sample-agnostic pattern and a sample-specific perturbation. Denoting the static pattern as δ and the parameterized generator for the sample-specific perturbation as $g_\phi(\cdot)$, the final poisoned sample x' is generated through a fusion function $\mathcal{H}$:

$$x' = G(x) = x + \mathcal{H}(\delta, g_\phi(x)) \tag{3}$$

where $\mathcal{H}(\cdot)$ represents the fusion strategy that combines the sample-agnostic and sample-specific components. This formulation allows us to flexibly integrate the stability of sample-agnostic triggers with the diversity of sample-specific perturbations.

3.3 Our Approach

This section details the proposed **Mask-Guided Hybrid Trigger (MGHT)**. As illustrated in Figure 1, we first propose a mask-guided fusion module that spatially integrates static and dynamic patterns using an adaptive semantic mask. Second, we introduce a Synergy-driven Co-optimization Strategy to jointly train these components in an alternating manner, yielding a hybrid trigger that achieves both high effectiveness and robustness.

Mask-Guided Hybrid Trigger Generation

The effectiveness of a hybrid trigger largely depends on the composition function $\mathcal{H}$. A naive approach to fusing a sample-agnostic trigger δ and a sample-specific trigger $g_\phi(x)$ is via affine combination:

$$\mathcal{H}(\delta, g_\phi(x)) = \alpha \cdot g_\phi(x) + (1-\alpha) \cdot \delta \tag{4}$$

However, this scalar-based mixing applies a uniform weight α across all spatial locations and images, ignoring the local semantic context. To overcome this limitation, we propose using a spatially adaptive mask to guide the fusion.

We introduce an adaptive mask generator $m_\psi(\cdot)$ that produces a soft mask $\mathbf{M} \in [0,1]^{H\times W\times 1}$, which is broadcast along the channel dimension. We employ a U-Net [Ronneberger *et al.*, 2015] architecture for $m_\psi(x)$ and modify the final layer with a sigmoid activation to ensure that the output values lie within $[0,1]$. Our proposed mask-guided hybrid trigger generator can be formulated as:

$$G_\Omega(x) = x + m_\psi(x) \odot g_\phi(x^{agn}) + (\mathbf{1} - m_\psi(x)) \odot \delta \tag{5}$$

where $\odot$ denotes element-wise multiplication, $\Omega = \{\delta, \phi, \psi\}$ denotes the hybrid trigger parameters, and x^{agn} represents the context-aware input (defined in Eq. 9). Regions where $\mathbf{M} \approx 0$ are allocated for the static trigger, preserving robustness, while regions where $\mathbf{M} \approx 1$ are reserved for the sample-specific trigger to enhance stealthiness. This spatial allocation reduces interference and guides the dynamic branch using the anchor-occupied regions as context.

For the sample-agnostic trigger δ, we define it as a globally shared, learnable parameter with the same dimensions as the input. It provides a stable anchor feature that facilitates the victim model's memorization of the backdoor pattern.

For the sample-specific trigger generator $g_\phi(\cdot)$, instead of generating the perturbation in isolation, we employ a context-aware strategy. First, we construct a sample-agnostic triggered input x^{agn} by applying the static trigger to the benign image via the complementary mask:

$$x^{agn} = x + (\mathbf{1} - \mathbf{M}) \odot \delta \tag{6}$$

Then, the sample-specific generator g_ϕ takes this context as input:

$$\delta_{spec} = g_\phi(x^{agn}) \tag{7}$$

By conditioning on x^{agn}, the generator considers the regions already occupied by the static trigger. This enables it to synthesize perturbations that transition smoothly from the static pattern and fill the masked regions in a manner that is semantically consistent with both the original image content and the static anchor.

Generator Optimization

We jointly optimize the trigger parameters $\Omega = \{\delta, \phi, \psi\}$ of the hybrid trigger generator $G_\Omega(x)$. This end-to-end optimization paradigm promotes stronger synergy between the two trigger types.

Formally, given an input sample x_i and a learnable sample-agnostic trigger, the generation of the final poisoned sample x'_i proceeds in four sequential steps. We define the intermediate variables and the final output as follows:

$$\mathbf{M}_i = m_\psi(x_i), \tag{8}$$

$$x_i^{agn} = x_i + (\mathbf{1} - \mathbf{M}_i) \odot \delta, \tag{9}$$

$$\delta_i^{spec} = g_\phi(x_i^{agn}), \tag{10}$$

$$x'_i = x_i + \mathbf{M}_i \odot \delta_i^{spec} + (\mathbf{1} - \mathbf{M}_i) \odot \delta \tag{11}$$

where $\mathbf{M}_i$ denotes the adaptive mask, x_i^{agn} represents the anchor-embedded intermediate state used to condition the generator, δ_i^{spec} is the generated sample-specific perturbation, and x'_i is the final hybrid-triggered sample.

Since the victim's specific training details are unknown during the dataset poisoning phase, we introduce a surrogate model s_ξ to approximate the victim's behavior. The attack objective is to minimize the classification loss of the generated poisoned samples toward the target label y_t:

$$\mathcal{L}_{atk} = \sum_{(x_i,y_i)\in\mathcal{D}} \mathcal{L}(s_\xi(x'_i), y_t) \tag{12}$$

To enforce complementarity, we require the hybrid sample x'_i to induce a lower target-class loss than two single-component counterparts: the static-anchor-only sample x_i^{agn} and the sample-specific-only sample x_i^{spec}. We utilize the predefined x_i^{agn} (Eq. 9) and construct $x_i^{spec} = x_i + \mathbf{M}_i \odot \delta_i^{spec}$. The Synergy Loss is formulated as:

$$\mathcal{L}_{syn} = \sum_{(x_i,y_i)\in\mathcal{D}} \big(\ell(x_i^{agn}) + \ell(x_i^{spec})\big) \tag{13}$$

where $\ell(\hat{x}) = \text{ReLU}\big(\mathcal{L}(s_\xi(x'_i), y_t) - \mathcal{L}(s_\xi(\hat{x}), y_t) + \mu_{syn}\big)$

Specifically, μ_{syn} is a predefined margin, and the ReLU function [Nair and Hinton, 2010] prevents optimization instability by zeroing the loss once the synergy constraint is satisfied.

To ensure perceptual similarity, we constrain the distortion using LPIPS [Zhang *et al.*, 2018], GMSD [Xue *et al.*, 2014], and PSNR [Hore and Ziou, 2010], with distinct thresholds μ_1, μ_2, μ_3:

$$\begin{aligned}\mathcal{L}_{inv} = {} & \text{ReLU}(\text{LPIPS}(x_i, x'_i) - \mu_1) \\ & + \text{ReLU}(\text{GMSD}(x_i, x'_i) - \mu_2) \\ & + \text{ReLU}(\mu_3 - \text{PSNR}(x_i, x'_i))\end{aligned} \tag{14}$$

Type	Method	CIFAR-10 (%)		ImageNet-10 (%)		CelebA (%)	
		BA	ASR	BA	ASR	BA	ASR
Sample-Agnostic	LC	**95.19**	98.77	91.20	82.44	**76.17**	**100.00**
	Sleeper Agent	90.96	94.63	**91.60**	11.11	72.62	12.58
	Narcissus	94.57	92.12	90.80	38.22	73.04	**100.00**
Sample-Specific	DABA	93.58	98.29	91.59	52.22	76.12	99.01
	COMBAT	93.26	99.08	88.80	82.22	72.46	88.50
	BAAT	88.22	97.16	86.00	**98.89**	72.42	72.53
Hybrid	Affine	93.07	85.73	88.80	69.78	71.01	98.51
	Mask-Guided	93.29	**99.69**	91.00	87.78	72.64	99.97
	Gains	+0.24%	+16.28%	+2.48%	+25.80%	+2.30%	+1.48%

Table 1: Comparison of attack success rate (ASR) and benign accuracy (BA) across different trigger types. The row "Gains" indicates the relative performance improvement of our Mask-Guided approach compared to the naive Affine baseline.

Surrogate Model	**Target Model (BA / ASR, %)**			
	ResNet34	MobileNetV2	VGG13	ViT
ResNet18	93.26 / 98.82	93.27 / 99.33	92.49 / 99.37	82.50 / 98.03
MobileNetV2	93.60 / 99.41	93.40 / 99.77	92.64 / 99.51	82.01 / 99.86
VGG13	93.37 / 98.01	93.18 / 97.51	92.44 / 98.13	83.79 / 92.99

Table 2: Black-box attack performance on CIFAR-10 across different surrogate-target architectures.

Dataset	PSNR ↑	LPIPS ↓	GMSD ↓	SSIM ↑
CIFAR-10	30.3331	0.0141	0.0472	0.9293
ImageNet-10	30.7470	0.0700	0.0753	0.8901
CelebA	30.0968	0.0297	0.0219	0.8695

Table 3: Quantitative evaluation of visual quality (PSNR, LPIPS, GMSD, SSIM) on different datasets.

Beyond perceptual constraints, we further constrain the perturbation magnitude. For the sample-agnostic trigger, we project δ onto the L_∞-ball of radius ϵ. For the sample-specific trigger, we apply a scaling operation to the output of $g_\phi(\cdot)$ to ensure the sample-specific perturbations remain within $[-\epsilon, \epsilon]$. Additionally, we introduce an L_2 regularization term to bound the energy of the dynamic trigger:

$$\mathcal{L}_{l_2} = \sum_{(x_i, y_i) \in \mathcal{D}} \|\delta_i^{spec}\|_2^2 \tag{15}$$

Following [Huynh *et al.*, 2024], we incorporate a clean consistency loss using a frozen clean model s_ξ^{clean} pre-trained on benign data. This term encourages triggered samples to preserve their original predictions under the clean model, preventing the generated trigger from degenerating into ordinary adversarial perturbations:

$$\mathcal{L}_{clean} = \sum_{(x_i, y_i) \in \mathcal{D}} \mathcal{L}(s_\xi^{clean}(x_i'), y_i) \tag{16}$$

The training of the hybrid trigger is formulated as a bilevel optimization problem. We aim to find the optimal trigger parameters Ω that minimize the combined trigger objective in Eq. 17, while the surrogate model s_ξ is trained to minimize the classification loss $\mathcal{L}_{cls}$ on the poisoned dataset:

$$\min_{\Omega} \quad \mathcal{L}_{atk} + \lambda_{syn}\mathcal{L}_{syn} + \mathcal{L}_{inv} + \lambda_c \mathcal{L}_{clean} + \lambda_{l_2}\mathcal{L}_{l_2} \tag{17}$$

$$\text{s.t.} \quad \xi(\Omega) \in \arg\min_{\xi} \Big(\sum_{x_i \in \mathcal{D}_s} \mathcal{L}_{cls}(s_\xi(x_i'), y_t) + \sum_{(x_j, y_j) \in \mathcal{D}_r} \mathcal{L}_{cls}(s_\xi(x_j), y_j) \Big) \tag{18}$$

where $\lambda_{syn}, \lambda_c, \lambda_{l_2}$ are weighting hyperparameters.

To solve Eq. 17, we adopt an alternated training strategy [Huynh *et al.*, 2024] as an approximation. First, we pretrain the clean surrogate model s_ξ^{clean} on the benign dataset $\mathcal{D}$. Then, in each iteration, we alternately fix Ω to update ξ, and fix ξ to update Ω.

4 Experiments

4.1 Experimental Setup

Datasets and Models. We evaluate our method on three image classification datasets: CIFAR-10 [Krizhevsky and Hinton, 2009], ImageNet-10, and CelebA [Liu *et al.*, 2015]. For ImageNet-10, we use a fixed subset consisting of 10 randomly selected classes from ImageNet-1K[2] [Deng *et al.*, 2009]. For CelebA, following the setting in [Salem *et al.*, 2022], we consider three attributes (Heavy Makeup, Mouth Slightly Open, and Smiling) to construct an 8-class classification task. We use ResNet18 [He *et al.*, 2016] as the default surrogate model for generating MGHT triggers. Both the mask generator and the sample-specific trigger generator employ a U-Net [Ronneberger *et al.*, 2015] backbone.

Victim Model Training. Victim models are trained for 200 epochs via SGD with momentum 0.9 and weight decay 5×10^{-4}. The dataset-wide poisoning rates are set to 4% for CIFAR-10 and ImageNet-10, and 1.25% for CelebA (approx. 40% and 10% of target-class samples, respectively). The target classes are set to Class 2 for CIFAR-10 and Class 1 for ImageNet-10 and CelebA.

MGHT Optimization. We employ alternating training for 150 epochs using the Adam optimizer [Kingma and Ba, 2015] for generators and the SGD optimizer for the surrogate model. The learning rates for the generators match those

[2]WordNet IDs: n01440764, n02107312, n02346627, n02979186, n03394916, n03417042, n03425413, n03445777, n03888257, n07753592.

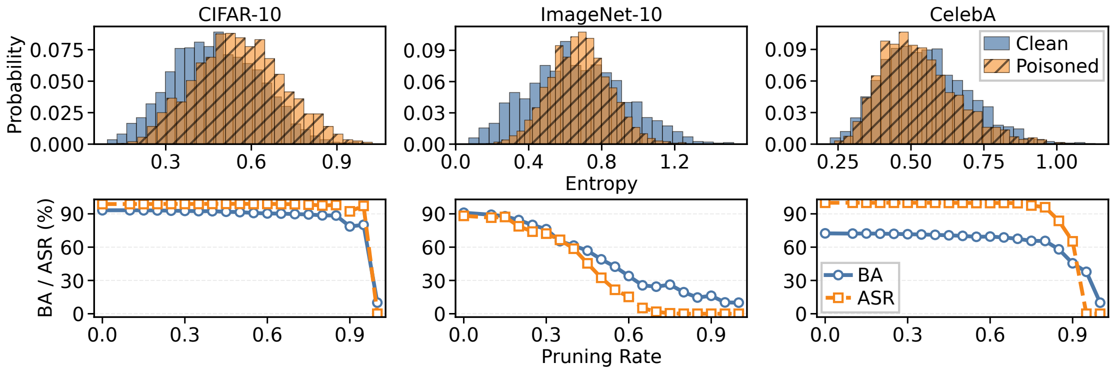

Figure 2: Robustness evaluation against STRIP and Fine-Pruning. Top row: entropy distributions of clean and poisoned samples under STRIP. Bottom row: BA and ASR under different pruning rates.

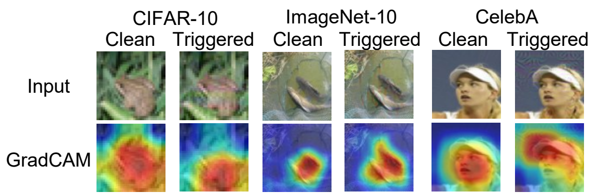

Figure 3: Grad-CAM visualization of MGHT-triggered samples.

Setting	CIFAR-10	ImageNet-10	CelebA
Data-free	0.6111 / 0.4296	0.2850 / 0.0071	0.7437 / 0.2445
Data-limited	0.6111 / 0.5480	0.2850 / 0.0071	0.7437 / 0.5156

Table 4: Detection performance of SCALE-UP on different datasets. The results are reported as AUROC / F1 score.

of the victim models on their respective datasets. The learning rate for the sample-agnostic trigger is set to 0.025 for CIFAR-10 and CelebA, and 0.005 for ImageNet-10. Key hyperparameters are $\lambda_{syn} = 0.5, \lambda_c = 0.8, \lambda_{l_2} = 0.02$, with the Synergy Loss margin $\mu_{syn} = 0.5$, perceptual thresholds $\mu_1 = 0.05, \mu_2 = 0.03, \mu_3 = 30.0$, and a perturbation budget of $\epsilon = 16/255$.

Evaluation Metrics. We employ two standard metrics: benign accuracy (BA) and attack success rate (ASR). BA measures the classification accuracy of the backdoored model on the clean test set. ASR measures the proportion of non-target samples in the triggered test set that are successfully misclassified as the target label y_t.

4.2 Effectiveness Analysis

White-box Settings

We first evaluate the attack performance under a standard backdoor attack scenario where the attacker possesses sufficient prior knowledge to align the surrogate model's architecture with the victim model. In this setting, we compare our method against two categories of clean-label attacks: (1) *Sample-agnostic* methods, including LC [Turner *et al.*, 2019], Sleeper Agent [Souri *et al.*, 2022], and Narcissus [Zeng *et al.*, 2023]; and (2) *Sample-specific* methods, including DABA [Xu *et al.*, 2023], COMBAT [Huynh *et al.*, 2024], and BAAT [Zhu *et al.*, 2025]. We also include a naive hybrid baseline, Affine, which independently trains a sample-agnostic trigger and a sample-specific generator and combines them using Eq. 4 with $\alpha = 0.5$. This baseline helps verify whether the performance gain of MGHT comes from the proposed mask-guided hybrid design rather than merely from using two trigger sources. For a fair comparison, we standardize the settings for certain baselines: for BAAT, we consistently utilize the style-transfer trigger mode across all datasets; for Narcissus, we remove the trigger amplification during the validation phase.

As shown in Table 1, MGHT demonstrates competitive performance across all benchmarks. It obtains the best ASR on CIFAR-10 (99.69%) and near-saturated ASR on CelebA (99.97%), while maintaining benign accuracy comparable to the clean models. On ImageNet-10, although BAAT achieves the highest ASR, MGHT preserves a higher BA. Compared with the naive Affine baseline, MGHT consistently improves ASR across all datasets, including a 25.80% relative gain on ImageNet-10, showing that the improvement mainly comes from mask-guided allocation rather than simply combining two trigger sources.

Black-box Settings

While the preceding analysis confirms the effectiveness of MGHT under the assumption of architectural alignment, we further evaluate MGHT's robustness in a black-box setting where the victim architecture is unknown. We conduct this black-box evaluation on the CIFAR-10 dataset. Specifically, we generate poisoned datasets using three surrogate models (ResNet18, MobileNetV2 [Sandler *et al.*, 2018], and VGG13 [Simonyan and Zisserman, 2015]) and four distinct target models, including ViT [Dosovitskiy *et al.*, 2021].

As summarized in Table 2, MGHT demonstrates remarkable transferability across all surrogate-target pairs. ASR values consistently exceed 97% across CNN architectures and reach up to 99.86% on ViT. This indicates that MGHT creates semantically strong patterns rather than overfitting to local convolutional artifacts. The strong transferability mainly stems from the hybrid design: the static anchor provides an architecture-invariant signal, while the dynamic branch improves semantic adaptation. This reduces overfitting to the

Dataset	CIFAR-10	ImageNet-10	CelebA	Threshold
Anomaly Index	1.0147	0.5162	1.5332	2.00

Table 5: Anomaly Index on different datasets under Neural Cleanse.

Configuration	CIFAR-10		ImageNet-10		CelebA	
	BA	ASR	BA	ASR	BA	ASR
w/o Synergy	93.72	98.26	93.00	76.89	72.68	99.97
w/ Synergy	93.29	99.69	91.00	87.78	72.64	99.97

Table 6: Impact of the Synergy Loss on BA and ASR.

surrogate architecture and enables MGHT to remain effective across unseen targets.

4.3 Stealthiness Analysis

We employ four standard metrics to measure the visual quality of the poisoned samples: PSNR [Hore and Ziou, 2010], LPIPS [Zhang *et al.*, 2018], GMSD [Xue *et al.*, 2014], and SSIM [Wang *et al.*, 2004].

As reported in Table 3, MGHT achieves excellent visual quality across all three datasets: PSNR values consistently exceed 30 dB, and SSIM scores remain high, while LPIPS and GMSD remain low. Since LPIPS aligns closely with human perception, these results quantitatively demonstrate that our hybrid triggers are highly imperceptible. This superior performance stems from our optimization strategy, which incorporates strict perceptual constraints ($\mathcal{L}_{inv}$) and enforces a predefined L_∞ budget.

4.4 Robustness to Backdoor Defenses

Robustness Against STRIP. STRIP [Gao *et al.*, 2019] detects triggers by superimposing patterns on inputs and monitoring prediction entropy. Figure 2 shows that the entropy distribution of MGHT-poisoned images substantially overlaps with that of clean images. Due to the sample-specific nature of the hybrid trigger, superimposing other image patterns disrupts its features and makes the entropy profile closer to that of clean inputs, allowing MGHT to bypass this defense.

Robustness Against SCALE-UP. SCALE-UP [Guo *et al.*, 2023] detects poisoned samples based on scaled prediction consistency. As shown in Table 4, MGHT is difficult to detect across all datasets: SCALE-UP even performs worse than random guessing on ImageNet-10, and its low F1 scores on CIFAR-10 and CelebA indicate unreliable detection. These results suggest that MGHT does not exhibit a distinctive scaling-consistency pattern.

Robustness Against Grad-CAM. Grad-CAM [Selvaraju *et al.*, 2017] visualizes regions critical to model predictions. As shown in Figure 3, MGHT heatmaps remain aligned with the main objects rather than shifting to localized trigger regions, weakening the localized-pattern assumption used by saliency-based defenses [Chou *et al.*, 2020].

Robustness Against Neural Cleanse. Neural Cleanse (NC) [Wang *et al.*, 2019] reverse-engineers triggers and detects abnormal target classes using an anomaly index. Table 5 shows that MGHT stays below the detection threshold of 2.0, suggesting that its spatially adaptive hybrid trigger is difficult to recover as a single consistent trigger.

Robustness Against Fine-Pruning. Fine-Pruning [Liu *et al.*, 2018] removes backdoors by pruning neurons that are inactive on benign data. Figure 2 illustrates that the ASR and BA decline simultaneously on all datasets, indicating the absence of an optimal pruning threshold. This suggests that neurons activated by MGHT are coupled with those for normal classification, preventing backdoor removal without compromising model utility.

Dataset	Weight (λ)	Margin (μ) (BA / ASR, %)	
		0.5	0.7
CIFAR-10	0.5	93.29 / 99.69	93.10 / 99.34
	0.8	93.32 / 99.18	93.28 / 99.24
ImageNet-10	0.5	91.00 / 87.78	90.40 / 84.67
	0.8	89.60 / 86.00	90.20 / 88.67
CelebA	0.5	72.64 / 99.97	72.21 / 99.97
	0.8	72.42 / 100.00	72.94 / 99.93

Table 7: Impact of Synergy Loss hyperparameters.

5 Ablation Study

Impact of Synergy Loss. We further investigate the impact of the proposed Synergy Loss ($\mathcal{L}_{syn}$). As shown in Table 6, removing this term leads to a notable ASR degradation, particularly on ImageNet-10 (dropping from 87.78% to 76.89%). This indicates that without explicit regularization, the optimization process tends to collapse onto a single component. The Synergy Loss mitigates this issue by enforcing complementary interaction between the two trigger components.

Impact of Synergy Loss Weight and Margin. Table 7 presents a sensitivity analysis of the Synergy Loss hyperparameters. On CIFAR-10 and CelebA, MGHT remains stable across different settings. On the more challenging ImageNet-10 dataset, the hyperparameters have a moderate influence on ASR, but the performance remains robust across all tested configurations. This suggests that MGHT is not overly sensitive to the tested Synergy Loss settings, although tuning can further improve attack effectiveness on complex datasets.

6 Conclusion

In this paper, we propose Mask-Guided Hybrid Trigger (MGHT), a novel framework that integrates static anchors and dynamic camouflage via an adaptive semantic mask. Furthermore, we introduce a Synergy-driven Co-optimization Strategy to prevent optimization collapse and enforce complementary learning between the two trigger components. Experiments across three benchmark datasets demonstrate that MGHT achieves competitive attack success rates and exhibits strong robustness against various defenses. These findings highlight the limitations of existing defenses that rely on single-mode trigger assumptions and motivate the development of dynamic-aware inspection mechanisms.

Ethical Statement

This work studies clean-label backdoor attacks to expose potential risks in third-party training data and to support the development of more reliable defenses. We do not advocate malicious use of the proposed method. The released code and any poisoned-data generation scripts are intended only for controlled research evaluation, and we encourage researchers to use the results to improve dataset inspection, backdoor detection, and model robustness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62272290 and 62572114.

References

[Chou *et al.*, 2020] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *IEEE Security and Privacy Workshops*, pages 48–54. IEEE, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[Doan *et al.*, 2020] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*, pages 897–912, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Gao *et al.*, 2019] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125, 2019.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

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

[Guo *et al.*, 2023] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *International Conference on Learning Representations*, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[Hore and Ziou, 2010] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[Huynh *et al.*, 2024] Tran Huynh, Dang Nguyen, Tung Pham, and Anh Tran. Combat: Alternated training for effective clean-label backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2436–2444, 2024.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

[Li *et al.*, 2021a] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021.

[Li *et al.*, 2021b] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.

[Li *et al.*, 2024] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2018] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning*, pages 807–814, 2010.

[Qiu *et al.*, 2021] Han Qiu, Yi Zeng, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 363–377, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Saha *et al.*, 2020] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11957–11965, 2020.

[Salem *et al.*, 2022] 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. IEEE, 2022.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[Souri *et al.*, 2022] Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In *Advances in Neural Information Processing Systems*, volume 35, pages 19165–19178, 2022.

[Turner *et al.*, 2019] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2019] 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. IEEE, 2019.

[Xu *et al.*, 2023] Chaohui Xu, Wenye Liu, Yue Zheng, Si Wang, and Chip-Hong Chang. An imperceptible data augmentation based blackbox clean-label backdoor attack on deep neural networks. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(12):5011–5024, 2023.

[Xue *et al.*, 2014] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE transactions on image processing*, 23(2):684–695, 2014.

[Yoshida and Fujino, 2020] Kota Yoshida and Takeshi Fujino. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In *Proceedings of the 13th ACM workshop on artificial intelligence and security*, pages 117–127, 2020.

[Zeng *et al.*, 2023] Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 771–785, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhu *et al.*, 2025] Mingyan Zhu, Yiming Li, Junfeng Guo, Tao Wei, Shu-Tao Xia, and Zhan Qin. Towards sample-specific backdoor attack with clean labels via attribute trigger. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4685–4698, 2025.