BEYOND IMAGE-WIDE COLOR SHIFTS: SCENE-COMPATIBLE SEMANTIC-REGION TRIGGERS FOR BACKDOOR ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-wide color space backdoor attacks can be visually subtle, but applying a single transformation to semantically distinct regions may produce chromatic changes that conflict with object boundaries and local illumination. We argue that stealth depends not only on perturbation magnitude, but also on where the transformation is applied and how its strength adapts to local content. Guided by this insight, we propose a scene-compatible semantic backdoor framework that decouples semantic carrier selection from spatially adaptive perturbation control. A stability-aware selection strategy identifies complete regions whose appearance changes are plausible within the surrounding scene, while texture-adaptive modulation assigns bounded, input-dependent color shifts according to local visual tolerance. Across four dataset–architecture settings, our method consistently achieves attack success rates above 98% while preserving clean accuracy and low perceptual distortion. Evaluations under representative defenses further suggest that semantically grounded, sample-dependent triggers challenge assumptions based on fixed or spatially regular trigger patterns.

Keywords: Backdoor attacks, semantic carrier selection, scene-compatible triggers, color space perturbations, visual stealthiness

1 INTRODUCTION

Deep neural networks (DNNs) are increasingly trained on data collected from external sources, exposing their pipelines to data-poisoning attacks. By corrupting only a small fraction of the training set, a backdoor attacker can make the resulting model behave normally on clean inputs yet produce an attacker-specified prediction when a hidden trigger appears (Gu et al., 2019). Because clean-data validation rarely exposes this conditional behavior, an effective attack must jointly preserve clean accuracy, maintain a high attack success rate, and avoid visual evidence that reveals poisoned samples before deployment.

Recent backdoor attacks have increasingly moved beyond conspicuous, input-invariant patches toward triggers that adapt to image content and resemble natural visual variations. Reflection-based effects and sample-dependent patterns improve visual concealment by avoiding a single artificial pattern (Liu et al., 2020; Nguyen, 2020). Image-wide color transformations further demonstrate that naturally occurring appearance changes can serve as effective trigger carriers (Jiang et al., 2023; Vrsnak et al., 2023). However, visual concealment alone does not guarantee scene compatibility. A shared transformation modifies the semantic subject and its surroundings together, although these regions may correspond to different objects, materials, and illumination conditions. Consequently, a poisoned image can remain close to its clean counterpart under pixel-level metrics while exhibiting an appearance change that conflicts with object boundaries or local lighting. As Fig. 1 illustrates, this exposes a broader gap: current trigger designs rarely distinguish where an appearance modification is contextually permissible from how strongly each location should be modified.

Localizing the transformation is therefore necessary but insufficient. A random or largest region may cross object boundaries or include content whose appearance should remain unchanged, whereas selecting the most textured region confuses concealment capacity with semantic validity. A scene-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

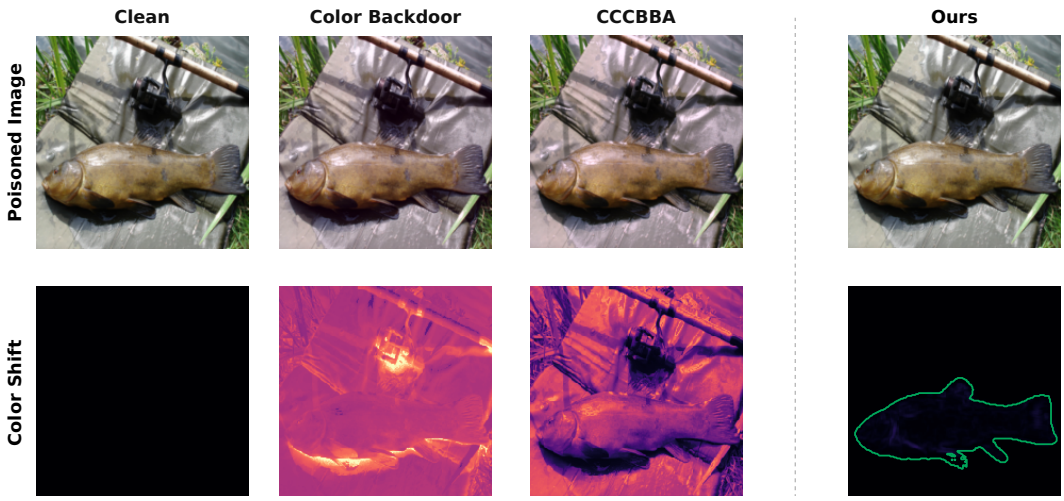

Figure 1: Image-wide color attacks modify both the semantic subject and its surrounding scene, whereas the proposed method confines the chromatic shift to a scene-compatible semantic carrier.

compatible trigger must instead answer two coupled questions: **1) Where and which carrier should be modified?** The carrier should be a coherent object or object part whose bounded appearance change respects its semantic boundary and the dominant illumination cues. **2) How much should each location be modified?** The strength should follow local visual tolerance, remaining weaker in smooth areas and relatively stronger, but bounded, in textured areas.

We address the first question with two consecutive semantic stages. **Stage 1: illumination-compatible carrier proposal.** A multimodal model reasons over the complete scene and proposes coherent object or object-part regions whose local appearance variation is plausible under the observed illumination. **Stage 2: semantic-stability evaluation.** Each candidate is assessed using three semantically equivalent prompts, and a stability-regularized score favors high average suitability with low sensitivity to wording. This reasoning incorporates object identity, spatial integrity, and illumination compatibility, which cannot be captured by area, saliency, or texture alone, and determines where modification is permitted.

After the carrier is fixed, **Stage 3: texture-adaptive chromatic modulation** answers the second question. Normalized local RMS luminance contrast is mapped to a bounded pixel-wise budget: smooth locations receive smaller perturbations, while textured locations receive relatively larger ones. A softened mask suppresses abrupt boundary transitions, and an input-dependent direction field controls lightness and chromatic displacement in Oklab (Ottosson, 2020). Oklab is thus the perceptual implementation space rather than the defining principle of the attack; the central design is the coordination of semantic placement and spatially adaptive modulation.

Across four dataset–architecture settings, our method achieves attack success rates above 98%. On Imagenette with ResNet18, it attains 99.96% attack success and 92.89% clean accuracy, together with LPIPS 0.0030 and PSNR 44.38 dB. Evaluations with STRIP, Neural Cleanse, and Fine-Pruning further support its effectiveness under the reported defense configurations; we do not claim universal resistance to adaptive defenses.

The main contributions of this work are summarized as follows:

- We identify scene-level artifacts in image-wide color backdoors as a mismatch between shared image transformations and object-level scene structure, and formulate a scene-compatible semantic trigger paradigm that determines where a local appearance change is contextually permitted.
- We develop stability-regularized semantic carrier selection and texture-adaptive modulation, combining multimodal scene reasoning, equivalent-prompt assessment, bounded pixel-wise budgets, softened spatial support, and an input-dependent Oklab direction field.

- We validate the framework across multiple datasets and architectures with baseline comparisons, perceptual-quality measurements, component ablations, and representative defenses, achieving above 98% attack success in all four settings while preserving clean accuracy.

2 Related Work

2.1 Stealthy and Semantic Backdoor Triggers

Early poisoning-based backdoor attacks associate target labels with input-invariant patches or blended patterns (Gu et al., 2019; Chen et al., 2017), while Trojaning optimizes triggers for targeted activation (Liu et al., 2018). Later designs improve visual concealment through natural-looking reflections or spatial warps (Liu et al., 2020; Nguyen, 2021). These approaches broaden the trigger space beyond fixed patches, but do not explicitly couple semantic support selection with spatially varying chromatic control.

Color-based attacks replace explicit patterns with global appearance transformations. Color Backdoor searches for an image-wide color space shift under perceptual constraints (Jiang et al., 2023), whereas CCCBBA uses a von Kries color-constancy model to construct multiplicative color shifts (Vrsnak et al., 2023). Their image-wide support improves robustness to common image transformations, but applies one transformation across semantically and photometrically distinct regions. Moreover, image-level distortion metrics such as PSNR, SSIM, and LPIPS quantify similarity rather than whether a change is compatible with a specific object's material and local illumination (Wang et al., 2004; Zhang et al., 2018).

Input-dependent attacks synthesize sample-specific patterns to weaken the regularity of universal triggers (Nguyen, 2020; Doan et al., 2021). VSSC instead selects semantic trigger objects and inserts them into scenes using generative models (Wang et al., 2025). Our setting differs in two respects: it reuses an existing object or object part rather than adding a new semantic entity, and it separates where modification is permitted from how strongly each pixel is changed. We further regularize carrier selection across equivalent prompts to reduce sensitivity to multimodal-model wording.

2.2 Backdoor Detection and Mitigation

Backdoor defenses exploit trigger regularity at the input or model level. Neural Cleanse reconstructs a small class-specific universal perturbation (Wang et al., 2019); STRIP detects unusually invariant predictions under input perturbation (Gao et al., 2019); and Fine-Pruning removes neurons that are dormant on clean inputs before fine-tuning (Liu et al., 2018). These methods represent the trigger-inversion, inference-time detection, and model-repair families evaluated in this work. We report their outcomes only under the stated configurations and do not infer general failure of these defense families from the input dependence or semantic support of our trigger, nor claim robustness to adaptive defenses.

3 Methodology

Our method constructs a scene-compatible semantic backdoor by separating two decisions: *where* a chromatic trigger can be placed and *how strongly* each location should be modified. As illustrated in Fig. 2, Qwen first proposes candidate descriptions under the SIB guidelines. The corresponding candidate carriers are then localized and segmented, after which MiniGPT-4 evaluates their semantic and illumination compatibility through stability-regularized ranking. Finally, local luminance texture determines the pixel-wise perturbation strength, while a Gaussian-smoothed mask confines an input-dependent Oklab displacement to the selected carrier.

3.1 Threat Model and Problem Formulation

We consider a targeted training-time backdoor attack on image classifiers, implemented through data poisoning in the standard all-to-one setting. Starting from a clean training set $\mathcal{D}$, the attacker selects a small fraction of samples from non-target classes, applies the proposed scene-compatible chromatic trigger to a semantic carrier in each image, and relabels the resulting samples as a fixed

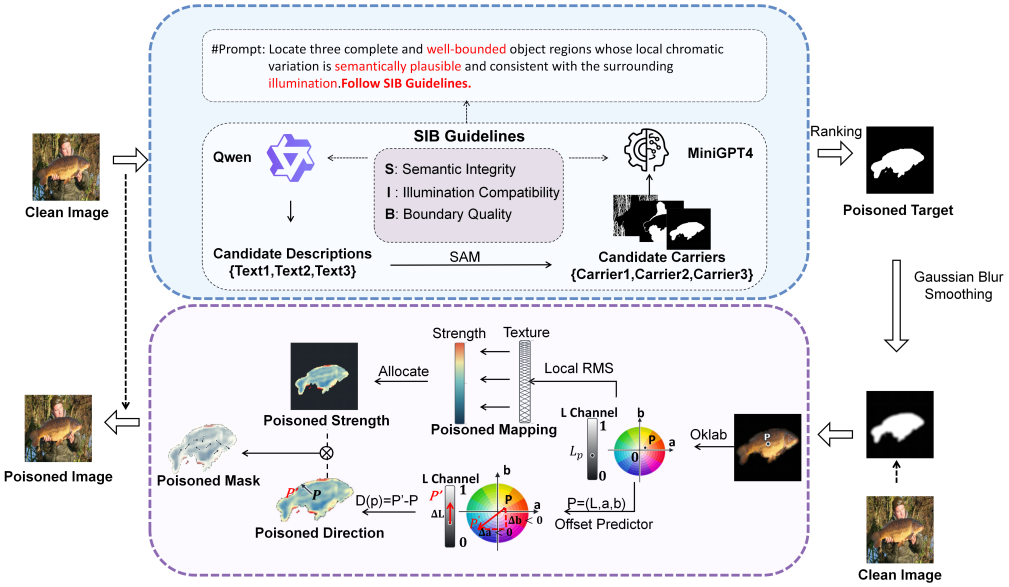

Figure 2: Overview of the proposed scene-compatible semantic backdoor. Qwen proposes candidate descriptions, GroundingDINO and SAM construct candidate carriers, MiniGPT-4 performs stability-regularized review, and local luminance texture controls the Oklab displacement within the selected region.

target class y_t. The victim is trained on the resulting mixture of clean and poisoned samples, without changing its architecture or inference procedure.

The attack aims to preserve normal predictions on clean inputs while mapping triggered inputs to y_t. In addition to clean accuracy and attack success, we require the induced chromatic changes to remain localized to the selected carrier and compatible with the surrounding scene. Formal definitions of the poisoning ratio, poisoned-sample construction, training-set construction, and evaluation metrics are provided in Appendix A.

3.2 SIB-Guided Semantic Carrier Selection

Candidate description generation. Given an input image x, Qwen reasons over the complete scene according to the SIB guidelines: semantic integrity, illumination compatibility, and boundary quality. Rather than directly predicting pixel-level masks, Qwen identifies complete objects or object parts whose bounded chromatic variations are semantically plausible and compatible with the surrounding illumination. It produces at most K candidate descriptions:

$$\mathcal{C}(x) = \{t_1, \ldots, t_K\}, \qquad K \leq 3. \tag{1}$$

Each description t_j is converted into a binary candidate carrier by a language-guided localization and segmentation operator:

$$M_j = \mathcal{G}(x, t_j), \qquad M_j \in \{0,1\}^{H \times W}, \tag{2}$$

where $\mathcal{G}$ consists of GroundingDINO for description-conditioned localization and SAM for mask generation. This separation allows Qwen to determine the semantic identity of a suitable carrier while the localization and segmentation models recover its spatial boundary.

MiniGPT-4 stability-regularized review. The candidate carriers are subsequently reviewed by MiniGPT-4. For each candidate M_j, we construct Q semantically equivalent review prompts based on the same SIB guidelines. MiniGPT-4 jointly observes the complete image and the highlighted candidate carrier and returns a suitability score s_{jq} for prompt q.

We aggregate the scores using

$$\mu_j = \frac{1}{Q}\sum_{q=1}^{Q} s_{jq}, \qquad \sigma_j = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(s_{jq}-\mu_j)^2}. \tag{3}$$

The stability-regularized score is

$$S_j = \mu_j - \lambda_{\text{stab}}\sigma_j, \tag{4}$$

and the final carrier is selected as

$$j^* = \arg\max_j S_j, \qquad M^* = M_{j^*}. \tag{5}$$

The mean term favors candidates with high semantic and illumination compatibility, whereas the variance penalty suppresses candidates whose assessments are sensitive to prompt wording. Consequently, the review stage provides a more stable basis for carrier selection than a single-prompt decision.

3.3 Texture-Adaptive Strength Modulation

After selecting M^*, we estimate the locally admissible perturbation strength. Let Y denote the relative luminance computed from linear RGB:

$$Y = 0.2126R_{\text{lin}} + 0.7152G_{\text{lin}} + 0.0722B_{\text{lin}}. \tag{6}$$

For a local window $\mathcal{N}(p)$ centred at pixel p, its luminance mean and standard deviation are

$$\mu_Y(p) = \frac{1}{|\mathcal{N}(p)|}\sum_{q\in\mathcal{N}(p)} Y(q), \tag{7}$$

$$\sigma_Y(p) = \sqrt{\frac{1}{|\mathcal{N}(p)|}\sum_{q\in\mathcal{N}(p)}\big(Y(q)-\mu_Y(p)\big)^2}. \tag{8}$$

We define normalized local RMS luminance contrast as

$$C_{\text{RMS}}(p) = \frac{\sigma_Y(p)}{\mu_Y(p)+\epsilon}. \tag{9}$$

The texture-adaptive strength map is then

$$B(p) = \text{clip}\left(b_0 + \kappa C_{\text{RMS}}(p)^{\gamma}, B_{\min}, B_{\max}\right). \tag{10}$$

Smooth locations receive weaker perturbations, whereas textured locations receive relatively stronger but bounded perturbations.

To avoid abrupt transitions at the carrier boundary, we apply Gaussian smoothing to the binary mask:

$$\widetilde{M} = G_{\sigma_m} * M^*, \tag{11}$$

where G_{σ_m} is a normalized two-dimensional Gaussian kernel, σ_m controls the smoothing scale, and $*$ denotes spatial convolution. The resulting soft mask gradually attenuates the perturbation near the carrier boundary.

3.4 Input-Dependent Oklab Perturbation

Let $z = \Psi(x) = (L, a, b)$ denote the Oklab representation of x. A lightweight generator predicts an image-dependent residual direction:

$$R_\phi(x, M^*) = \tanh\left(G_\phi([x, M^*])\right). \tag{12}$$

To maintain a consistent backdoor signature across poisoned samples, this residual is combined with a learnable shared chromatic anchor d_0:

$$D_\phi(x, M^*) = A \odot \tanh\left(d_0 + \eta_t R_\phi(x, M^*)\right), \tag{13}$$

where $A = (\alpha_L, \alpha_a, \alpha_b)$ specifies the channel-wise displacement bounds and η_t controls the contribution of the input-dependent residual during training.

The final spatially varying Oklab displacement is

$$\Delta z(p) = \widetilde{M}(p)\, B(p)\, D_\phi(x, M^*)(p). \tag{14}$$

The poisoned image is obtained by

$$x^p = \mathrm{clip}\left(\Psi^{-1}\big(z + \Delta z\big), 0, 1\right). \tag{15}$$

Semantic carrier selection therefore determines where the trigger is applied, while luminance texture and the learned Oklab direction jointly determine its spatially varying appearance.

3.5 Poisoned Training Objective

Clean and poisoned samples are mixed in the same training batches. The classification loss is

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{|\mathcal{B}|} \sum_{(x_i, y_i) \in \mathcal{B}} \ell_{\mathrm{CE}}\left(f_\theta(\widetilde{x}_i), \widetilde{y}_i\right), \tag{16}$$

where $(\widetilde{x}_i, \widetilde{y}_i) = (x_i, y_i)$ for clean samples and $(\widetilde{x}_i, \widetilde{y}_i) = (x_i^p, y_t)$ for poisoned samples.

The victim classifier and chromatic generator are jointly optimized using

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{color}} \mathcal{L}_{\mathrm{color}} + \lambda_{\mathrm{feat}} \mathcal{L}_{\mathrm{feat}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{color}}$ constrains the magnitude of the Oklab displacement and $\mathcal{L}_{\mathrm{feat}}$ encourages feature consistency between clean and poisoned images. These constraints balance attack effectiveness against the visual fidelity of the generated samples.

4 Evaluation

We organize the evaluation around four questions: **(1)** Can the proposed attack achieve reliable target-class activation while retaining clean accuracy across datasets and architectures? **(2)** Does semantic localization improve visual stealthiness over existing backdoor attacks? **(3)** How do semantic carrier selection, color space, and poisoning ratio affect the attack? **(4)** How does the attack behave under representative backdoor defenses?

4.1 Experimental Setup

Datasets and models. We evaluate four dataset–architecture combinations: Imagenette (Howard, 2019) with ResNet18 (He et al., 2016), Imagenette with ViT-B/16 (Dosovitskiy et al., 2021), CIFAR-10 (Krizhevsky et al., 2009) with VGG16 (Simonyan, 2015), and a ten-class Places365 (Zhou et al., 2018) subset with DenseNet161 (Huang et al., 2017). Unless otherwise specified, controlled comparisons and ablations use Imagenette with ResNet18 and a 10% poisoning ratio. We adopt the all-to-one setting and assign poisoned samples from non-target classes to a fixed target label.

Baselines and metrics. We compare against BadNets (Gu et al., 2019), Refool (Liu et al., 2020), Color Backdoor (Jiang et al., 2023), CCCBBA (Vrsnak et al., 2023), and VSSC (Wang et al., 2025). We report clean accuracy (ACC) and attack success rate (ASR), together with LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), PSNR, and DreamSim for visual similarity. Lower LPIPS and DreamSim and higher SSIM and PSNR indicate better similarity. We further evaluate STRIP (Gao et al., 2019), Fine-Pruning (Liu et al., 2018), and Neural Cleanse (Wang et al., 2019). Additional implementation details are provided in Appendix B.

4.2 Attack Effectiveness

Table 1 reports cross-setting performance and the controlled comparison with existing attacks. Our method achieves ASRs above 98% across all four dataset–architecture combinations, ranging from

Table 1: Attack effectiveness (%). Left: performance across dataset–architecture settings. Right: comparison on Imagenette with ResNet18 at a 10% poisoning ratio. Higher values are better.

Dataset / Model	ACC	ASR
Imagenette / ResNet18	92.89	99.96
Imagenette / ViT-B/16	97.07	99.49
CIFAR-10 / VGG16	92.00	98.07
Places365 / DenseNet161	91.73	98.72

Attack	ACC	ASR
BadNets	92.85	96.96
Refool	89.71	97.22
Color Backdoor	90.98	97.45
CCCBBA	90.47	95.46
VSSC	92.71	97.95
Ours	**92.89**	**99.96**

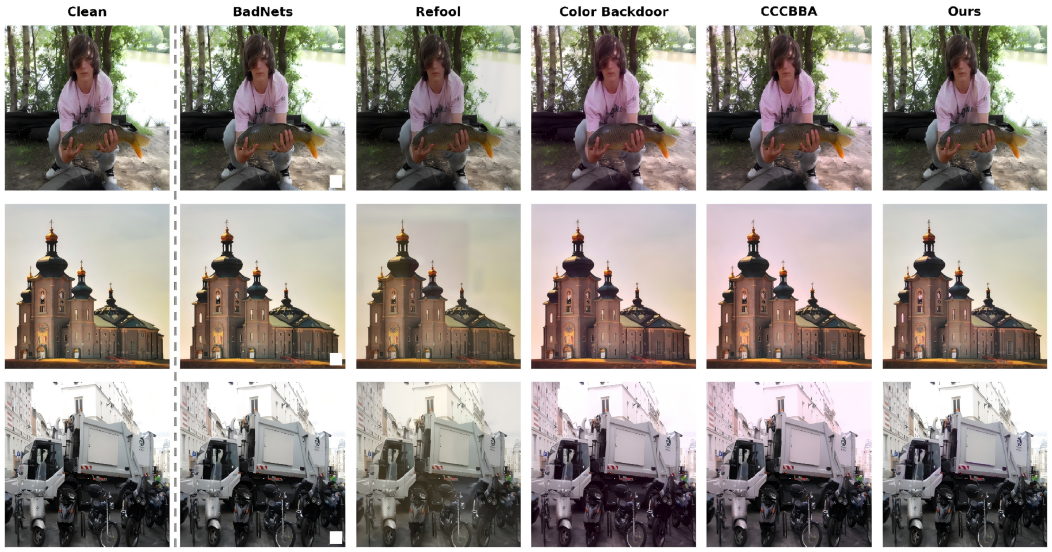

Figure 3: Qualitative comparison of clean and poisoned images. Existing color-based attacks apply image-wide transformations, whereas our method restricts the chromatic displacement to a scene-compatible semantic carrier.

98.07% on CIFAR-10 with VGG16 to 99.96% on Imagenette with ResNet18, while ACC ranges from 91.73% to 97.07%.

On Imagenette with ResNet18, our method obtains the highest ACC and ASR among the evaluated attacks. Its 99.96% ASR exceeds the strongest competing result, 97.95% from VSSC, by 2.01 percentage points. This result supports the effectiveness of combining scene-compatible carrier selection with input-dependent chromatic modulation.

4.3 Visual Stealthiness

Figure 3 provides qualitative comparisons. Image-wide color attacks modify the foreground object and surrounding scene simultaneously, whereas our method confines the chromatic displacement to a selected semantic carrier and suppresses abrupt transitions using a Gaussian-smoothed mask. Quantitative results are reported in Table 2.

Our method achieves the lowest LPIPS (0.0030), highest SSIM (0.9982), and highest PSNR (44.3823 dB). Its LPIPS is lower than the strongest baseline value of 0.0120 from BadNets, while its PSNR exceeds the strongest baseline value of 36.8917 dB from VSSC by 7.49 dB. Over 500 clean–poisoned image pairs, it also obtains a DreamSim mean of 0.000408. The four metrics consistently indicate low perceptual displacement under the reported evaluation.

Table 2: Visual similarity between clean and poisoned images. DreamSim is evaluated on 500 pairs for methods with numerical entries; −− denotes an unreported measurement.

Attack	LPIPS ↓	SSIM ↑	PSNR ↑	DreamSim ↓	Std. ↓
BadNets	0.0120	0.9926	36.1391	–	–
Refool	0.0300	0.9484	24.1504	–	–
Color Backdoor	0.0323	0.9778	31.6636	0.026208	0.012464
CCCBBA	0.0263	0.9970	33.2540	0.024348	0.014063
VSSC	0.0630	0.9551	36.8917	0.410351	0.205590
Ours	**0.0030**	**0.9982**	**44.3823**	**0.000408**	**0.000481**

Table 3: Ablation and sensitivity results on Imagenette with ResNet18. ACC and ASR are percentages; −− indicates that a metric is not applicable.

Factor	Variant	ACC ↑	ASR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Carrier	Random mask	91.50	97.38	0.0054	**0.9983**	42.48
	Largest mask	91.29	99.89	0.0048	0.9982	41.39
	Ours	**92.89**	**99.96**	**0.0030**	0.9982	**44.38**
Color space	RGB	92.40	99.77	0.0127	0.9974	40.9396
	CIELAB	92.43	99.85	0.0090	0.9943	41.1750
	Oklab	**92.89**	**99.96**	**0.0030**	**0.9982**	**44.3823**
Poisoning ratio	1%	93.16	85.44	–	–	–
	3%	93.08	91.93	–	–	–
	5%	92.91	97.50	–	–	–
	10%	92.89	99.96	–	–	–

4.4 Ablation and Sensitivity Analysis

Table 3 summarizes the effects of semantic carrier selection, color space, and poisoning ratio.

Semantic carrier selection. Replacing SIB-guided selection with a random or largest candidate mask reduces the overall balance between attack effectiveness and visual quality. The full method achieves the highest ACC and ASR, lowest LPIPS, and highest PSNR. Although the random-mask variant has a marginally higher SSIM (0.9983 versus 0.9982), its ACC, ASR, LPIPS, and PSNR are inferior. The largest-mask variant reaches 99.89% ASR but yields lower clean accuracy and poorer perceptual metrics, showing that mask area alone is insufficient for carrier selection.

Color space. We implement the same attack in RGB, CIELAB, and Oklab while retaining the remaining pipeline. Oklab achieves the best value for every reported effectiveness and distortion metric: LPIPS decreases to 0.0030 and PSNR increases to 44.3823 dB. These results support Oklab as the most suitable space among the evaluated alternatives, without attributing the complete method's performance to the color space alone.

Poisoning ratio. Increasing the poisoning ratio from 1% to 10% raises ASR from 85.44% to 99.96%, while ACC changes only from 93.16% to 92.89%. A 5% ratio already achieves 97.50% ASR, whereas the 1% setting leaves a clear effectiveness gap. These results characterize the trade-off between attack reliability and poisoning budget.

4.5 Evaluation against Backdoor Defenses

We evaluate the attack against three representative defenses under their reported configurations. These results characterize the tested mechanisms and do not establish universal robustness or resistance to adaptive defenses.

STRIP. STRIP tests whether model predictions remain unusually stable under input perturbations. The normalized entropy distributions of clean and triggered inputs substantially overlap and do not provide clear separation under the applied configuration. The complete entropy distributions and evaluation protocol are provided in Appendix E.

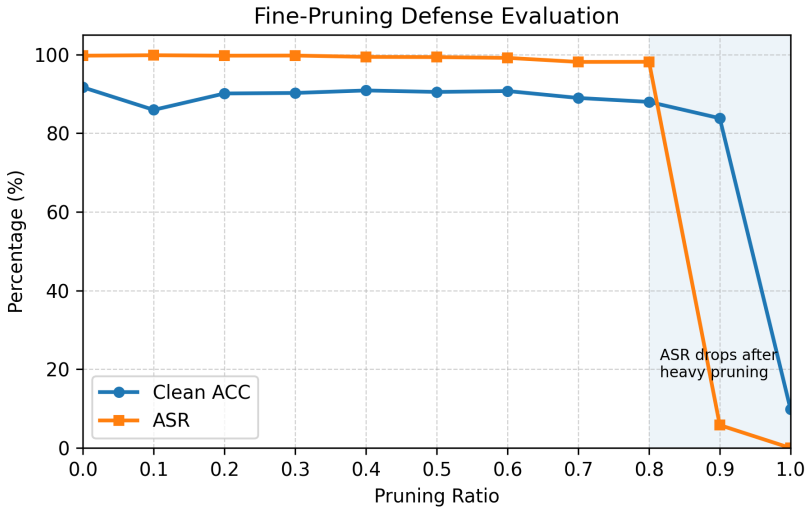

Figure 4: Fine-Pruning evaluation. The attack retains high ASR under moderate pruning and is suppressed only when aggressive pruning also severely degrades clean accuracy.

Fine-Pruning. Fine-Pruning removes neurons with low activation on clean inputs and measures the resulting ACC and ASR. As shown in Fig. 4, ASR remains above 98% for pruning ratios up to 0.8, while ACC remains approximately 88%–92%. At a ratio of 0.9, ASR drops to 5.82%; pruning all evaluated neurons reduces ASR to 0.00% but also lowers ACC to 9.83%. Suppressing the backdoor under this configuration therefore requires pruning severe enough to destroy most clean-task utility.

Neural Cleanse. Neural Cleanse searches for an abnormally small class-specific universal trigger. The anomaly indices of the clean and backdoored models are close and remain below the applied detection threshold. Neural Cleanse therefore does not identify an anomalous target class under this configuration. The corresponding anomaly-index comparison and evaluation protocol are provided in Appendix E.

5 Conclusion

In this work, we identify a limitation of image-wide color space backdoor attacks: applying one transformation to semantically distinct regions can produce scene-incompatible chromatic changes. We address this issue with a scene-compatible semantic backdoor that combines SIB-guided carrier selection with texture-adaptive, input-dependent Oklab perturbations. Across multiple dataset–architecture settings, the proposed attack maintains clean accuracy and high attack success while introducing low perceptual distortion, and ablations support the roles of semantic carrier selection and texture-adaptive modulation. Evaluations under three representative defenses further indicate that semantically grounded, sample-dependent triggers can challenge assumptions based on fixed or spatially regular patterns. These findings motivate defenses that consider semantic compatibility in addition to pixel-level trigger regularity.

AI use statement

In this work, we used OpenAI Codex to assist with drafting and restructuring portions of the manuscript, improving language and readability, preparing LaTeX-formatted content, checking internal consistency, inspecting existing implementation code, and providing feedback on the presentation of the methodology and experimental results. The tool was not used to generate experimental measurements, fabricate data, or replace the execution of the reported experiments. The core research idea, attack design, experimental results, and final scientific judgments were provided and determined by the authors. All AI-assisted text, technical descriptions, citations, and interpretations

were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

We provide the threat model, poisoned-data construction, and evaluation definitions in Appendix A. Dataset, architecture, and selected implementation settings are documented in Appendix B. The operational definition of the SIB guidelines, the candidate-generation and review prompts, and the stability-aware carrier-selection procedure are provided in Appendix C. Appendix D presents the complete attack algorithm, and Appendix E describes the evaluated defense protocols and additional results. The datasets and pretrained model architectures used in our experiments are publicly available and cited in the main paper.

Ethics statement

This work studies backdoor attacks to expose limitations of existing trigger designs and defenses. The proposed method could be misused to compromise image classification systems if applied without authorization. Our experiments are restricted to public research datasets and controlled digital image classification settings, and do not target deployed systems, private data, or human subjects. We report evaluations against representative defenses to support the development of stronger detection and mitigation methods. The release of implementation artifacts will exclude trained malicious models and will be intended solely for reproducible academic research.

References

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," arXiv preprint arXiv:2308.12966, 2023.

X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv preprint arXiv:1712.05526, 2017.

K. D. Doan, Y. Lao, W. Zhao, and P. Li, "LIRA: Learnable, imperceptible and robust backdoor attacks," in *Proc. ICCV*, 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations*, 2021.

S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "DreamSim: Learning new dimensions of human visual similarity using synthetic data," in *Advances in Neural Information Processing Systems*, vol. 36, pp. 50742–50768, 2023.

Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. Annual Computer Security Applications Conference*, pp. 113–125, 2019.

T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

J. Guo, W. Jiang, X. Xiang, J. Wang, J. Zhao, G. Lu, and G. Luo, "BadMars: Data poisoning attacks on image segmentation models for Mars rover autonomous driving," *Tsinghua Science and Technology*, doi: 10.26599/TST.2026.90100031, 2026.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

J. Howard, "Imagenette: A smaller subset of 10 easily classified classes from Imagenet," 2019. [Online]. Available: https://github.com/fastai/imagenette

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

W. Jiang, H. Li, G. Xu, and T. Zhang, "Color backdoor: A robust poisoning attack in color space," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 8133–8142, 2023.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. ICCV*, pp. 4015–4026, 2023.

A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, University of Toronto, 2009.

Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," arXiv preprint arXiv:2101.05930, 2021.

Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," arXiv preprint arXiv:2004.04692, 2020.

K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. International Symposium on Research in Attacks, Intrusions, and Defenses*, pp. 273–294, 2018.

S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," arXiv preprint arXiv:2303.05499, 2023.

Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Proc. ECCV*, pp. 182–199, 2020.

Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. NDSS*, 2018.

T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

T. A. Nguyen and A. T. Tran, "WaNet: Imperceptible warping-based backdoor attack," in *Proc. ICLR*, 2021.

B. Ottosson, "A perceptual color space for image processing," 2020. [Online]. Available: https://bottosson.github.io/posts/oklab/

H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, "DeepSweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation," in *Proc. ACM Asia Conference on Computer and Communications Security*, pp. 363–377, 2021.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, pp. 618–626, 2017.

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

D. Vrsnak, I. Sabolic, M. Subasic, and S. Loncaric, "Computational color constancy-based backdoor attacks," in *Proc. International Symposium on Image and Signal Processing and Analysis*, 2023.

G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symposium on Security and Privacy*, pp. 707–723, 2019.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

R. Wang, H. Chen, Z. Zhu, L. Liu, and B. Wu, "Versatile backdoor attack with visible, semantic, sample-specific, and compatible triggers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2025.3642161, 2025.

K. Yoshida and T. Fujino, "Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks," in *Proc. ACM Workshop on Artificial Intelligence and Security*, pp. 117–127, 2020.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

A Threat Model and Formal Definitions

A.1 Attack Setting and Adversarial Capabilities

We consider a targeted training-time backdoor attack against image classification models under the standard all-to-one setting. Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in [0,1]^{H\times W\times 3}, \quad y_i \in \{1, \dots, C\}, \tag{18}$$

denote the clean training set, where x_i is an input image and y_i is its ground-truth label. The attacker specifies a target class y_t and selects a subset $\mathcal{D}_s \subset \mathcal{D}$ from the non-target classes, such that $y_i \neq y_t$ for every $(x_i, y_i) \in \mathcal{D}_s$. The poisoning ratio is defined as

$$\rho = \frac{|\mathcal{D}_s|}{|\mathcal{D}|}. \tag{19}$$

The attacker can select the poisoned training samples, construct their semantic carrier masks, apply the proposed chromatic transformation, and replace their labels with the target label. Consistent with the training procedure described in the main paper, the attacker jointly optimizes the victim classifier and the chromatic transformation generator. However, the architecture of the victim classifier and its deployment-time inference procedure remain unchanged. The carrier-selection and trigger-generation components are used to construct a triggered image before it is presented to the classifier; they are not appended to the deployed victim model.

Our evaluation is restricted to digital image classification. We do not assume that the proposed chromatic trigger survives physical image acquisition, nor do we assume knowledge of an adaptive defense designed specifically for scene-compatible semantic triggers. The evaluations against existing defenses therefore characterize the attack only under their reported configurations and do not establish universal resistance to backdoor defenses.

A.2 Poisoned Sample and Training-Set Construction

For each selected image x_i, the semantic carrier-selection procedure produces a carrier mask $M_i^\star$. The proposed local chromatic transformation is denoted by $\mathcal{T}_\phi$, where ϕ represents the parameters of the chromatic generator. The corresponding poisoned image is

$$x_i^p = \mathcal{T}_\phi(x_i, M_i^\star). \tag{20}$$

Unlike an image-wide color transformation, $\mathcal{T}_\phi$ confines the chromatic displacement to the selected semantic carrier and assigns spatially varying perturbation strengths within that region.

The attacker assigns the target label y_t to every poisoned image. The poisoned subset and the final training set are therefore

$$\begin{aligned} \mathcal{D}_p &= \left\{ \left(\mathcal{T}_\phi(x_i, M_i^\star), y_t\right) \,\middle|\, (x_i, y_i) \in \mathcal{D}_s \right\}, \\ \widetilde{\mathcal{D}} &= \left(\mathcal{D} \setminus \mathcal{D}_s\right) \cup \mathcal{D}_p. \end{aligned} \tag{21}$$

Thus, each selected clean sample is replaced by its poisoned counterpart, and the size of the training set remains unchanged. The victim model is trained on $\widetilde{\mathcal{D}}$ using the joint classification and chromatic-regularization objective defined in the main paper.

A.3 Attack Objectives and Evaluation Metrics

Let $f_{\widehat{\theta}}$ denote the trained victim classifier and

$$\widehat{y}(x) = \arg\max_{c\in\{1,\dots,C\}} f_{\widehat{\theta}}(x)_c$$

denote its predicted class. For a clean test set $\mathcal{D}_{\text{te}} = \{(x_j, y_j)\}_{j=1}^{N_{\text{te}}}$, clean accuracy is computed as

$$\text{ACC} = \frac{1}{N_{\text{te}}} \sum_{j=1}^{N_{\text{te}}} \mathbb{I}\big[\widehat{y}(x_j) = y_j\big], \tag{22}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

To avoid counting samples that already belong to the target class as successful attacks, we evaluate attack success on

$$\mathcal{D}_{\mathrm{te}}^{\neg t} = \{(x_j, y_j) \in \mathcal{D}_{\mathrm{te}} \mid y_j \neq y_t\}.$$

For each non-target test image, a semantic carrier $M_j^\star$ is selected and the corresponding triggered image is generated using the learned transformation. The attack success rate is

$$\mathrm{ASR} = \frac{1}{|\mathcal{D}_{\mathrm{te}}^{\neg t}|} \sum_{(x_j, y_j) \in \mathcal{D}_{\mathrm{te}}^{\neg t}} \mathbb{I}\left[\widehat{y}\big(\mathcal{T}_{\widehat{\phi}}(x_j, M_j^\star)\big) = y_t\right]. \tag{23}$$

A successful attack should simultaneously retain high ACC on clean inputs and obtain high ASR on triggered non-target inputs. Because attack effectiveness alone does not characterize visual concealment, we additionally report LPIPS, SSIM, PSNR, and DreamSim between clean and poisoned images. These metrics measure different aspects of perceptual similarity and are therefore reported separately rather than being combined into a single stealth score. The localization and visual plausibility of the chromatic modification are further examined through qualitative comparisons and component ablations.

B Details of Experimental Settings

Datasets and architectures. We conduct experiments on Imagenette, CIFAR-10, and a ten-class subset of Places365. The evaluated victim architectures include ResNet18 and ViT-B/16 on Imagenette, VGG16-BN on CIFAR-10, and DenseNet161 on Places365-Subset. The Places365 subset comprises *stadium baseball*, *stage indoor*, *clothing store*, *television room*, *street*, *highway*, *amusement park*, *airport terminal*, *operating room*, and *gas station*.

Training configurations. For the Imagenette–ResNet18 poisoning-ratio study, input images are resized to 224×224. The victim model and trigger generator are jointly trained for 100 epochs using Adam with learning rates of 10^{-4} and 10^{-3}, respectively. We use a batch size of 32, target label 1, and random seed 2026. For CIFAR-10–VGG16-BN, we train for 50 epochs with a batch size of 256. The victim and generator learning rates are set to 0.05 and 10^{-3}, respectively, with a weight decay of 5×10^{-4}. The learning rate is reduced by a factor of 0.1 at epochs 30 and 40. The target label is 0, and the random seed is 0.

Trigger optimization. In the Imagenette–ResNet18 poisoning-ratio study, the perturbation bounds in Oklab space are set to $\alpha_L = 0.01$ and $\alpha_a = \alpha_b = 0.03$. The texture-adaptive strength module uses a 5×5 local window, with $\epsilon = 10^{-4}$, $b_0 = 0.20$, $\kappa = 1.0$, and $\gamma = 1.0$; the resulting spatial budget is constrained to $[0.20, 1.00]$. The shared chromatic anchor is initialized as $(0, 0.75, -0.75)$, and the maximum content-dependent residual scale is 0.25. We first optimize the shared anchor alone for five epochs and then linearly increase the residual contribution over ten epochs. The lightness, chromatic, and feature-preservation losses are weighted by 10.0, 7.0, and 2.0, respectively. Training-time EOT is disabled, and the standard ASR is evaluated without EOT.

C Semantic Carrier Selection Details

This section provides the operational definition of the SIB guidelines and the prompt templates used for semantic carrier proposal and stability-aware review. The prompts are designed to separate semantic suitability from low-level visual complexity: a region is not preferred merely because it is large, salient, or highly textured.

C.1 Operational Definition of the SIB Guidelines

The SIB guidelines assess a candidate carrier along three complementary dimensions:

S: Semantic Integrity. The carrier should correspond to a recognizable and spatially coherent object or object part. It should not merge unrelated objects, consist mainly of fragmented

components, or cover the entire scene. Small, heavily occluded, transparent, or semantically ambiguous regions are avoided because their spatial support is unstable.

I: Illumination Compatibility. A small color or illumination variation within the carrier should remain plausible under the surrounding scene. Regions dominated by sky, blank background, shadows, highlights, watermarks, or other unstable appearance cues are discouraged. This criterion concerns whether the intended chromatic change is contextually plausible, rather than whether the region simply contains strong texture.

B: Boundary Quality. The carrier mask should approximately follow the visible boundary of the corresponding object or object part. Masks that substantially include background pixels, cross unrelated semantic boundaries, or have highly fragmented and ambiguous contours receive lower scores.

In addition to these three primary criteria, the prompts require a moderate carrier size and explicitly treat texture as a secondary cue. Texture may increase the local tolerance to a small perturbation, but it does not establish semantic suitability by itself.

C.2 Candidate Description Prompt

Given an input image, Qwen first proposes at most three short English descriptions that can be passed to the localization and segmentation modules. The following prompt is used:

> **Candidate-description prompt.**
>
> You are an image object selector. Your task is to provide candidate detection phrases for GroundingDINO and SAM.
>
> Observe the image and output at most 3 English candidate phrases that are suitable for segmentation. Do not force the output to contain exactly 3 candidates.
>
> **Selection rules:**
>
> 1. Each candidate must be a clearly visible object or object part in the image.
>
> 2. Each candidate should be general rather than overly specific. For example, output "dog" instead of "a brown dog," and output "vehicle" instead of "red sports car."
>
> 3. Prefer regions that can naturally carry a subtle color or illumination change.
>
> 4. Prefer semantically stable objects with clear boundaries.
>
> 5. Avoid unstable or unsuitable regions such as sky, background, the whole image, scene, shadow, light, and watermark.
>
> 6. Avoid objects that are too small, heavily occluded, transparent, strongly reflective, or boundary-ambiguous.
>
> 7. Do not rank candidates by texture complexity alone.
>
> **Output format:**
>
> Return only a JSON array. The array length must be from 0 to 3, and each item must be a lowercase English noun phrase. Do not output explanations, comments, Markdown, or additional text.
>
> Valid example: `["dog", "clothing", "tree"]`
>
> If no suitable object exists, return `[]`.

The returned descriptions are normalized into short noun phrases and duplicate descriptions are removed. Each retained description is localized by GroundingDINO and converted into a candidate mask by SAM. The implementation retains at most two detected boxes for each description. During the initial filtering stage, masks occupying between 5% and 30% of the image are accepted. If no candidate survives, the admissible interval is relaxed to 2%–45%.

C.3 Stability-Aware Carrier Review

Each candidate mask is visualized as a translucent red region with a yellow boundary contour. The reviewer observes the complete image, the highlighted carrier, its candidate description, and its area ratio. Three semantically equivalent prompts are used to reduce sensitivity to a particular wording.

Review prompt 1.

You are a visual region quality judge for a natural image perturbation task. The image contains a candidate mask overlay. The red translucent region is the candidate region, and the yellow contour is the mask boundary.

Candidate target: {candidate}

Mask area ratio: {area_ratio}

Your task is to score whether this region is suitable for carrying a subtle natural color or illumination change.

Scoring criteria:

1. Semantic stability: the region should cover a clear object or object part.

2. Boundary quality: the mask boundary should roughly match the object boundary.

3. Visual naturalness: a slight color or illumination shift in this region should look natural.

4. Background avoidance: the region should not mainly cover sky, blank background, shadow, watermark, or the whole image.

5. Size suitability: the region should be neither too small nor too large.

6. Texture tolerance: the region may contain moderate texture or color variation, but texture complexity is not the primary criterion.

7. Region consistency: the region should not mix many unrelated objects.

Return only a JSON object in the following format:

```
{"score": 0, "valid": true, "reason": "short reason"}
```

The score must be an integer from 0 to 100. The validity field should be true when the score is at least 60. The reason must be a short English phrase. Do not output Markdown or any additional text.

Review prompt 2.

You are evaluating a masked image region for a subtle relighting or color-shift trigger. The red transparent area shows the proposed region, and the yellow line shows its boundary.

The proposed text label is: {candidate}

The region covers {area_ratio} of the image.

Assign a quality score from 0 to 100. A high score means that the region is a clear and stable object or object part, its mask boundary is reasonably aligned with the object, and a small natural color or illumination change would not look strange there. The region should not consist mainly of background, sky, shadow, watermark, or blank area. It should have a reasonable size and should not merge unrelated objects.

Texture is only a secondary cue. Do not select a region only because it is highly textured.

Return only the following JSON format:

```
{"score": 0, "valid": true, "reason": "short reason"}
```

Do not output Markdown or explanations outside the JSON object.

Review prompt 3.

You are a strict reviewer of candidate regions for natural-looking image perturbations. In the image, the red overlay marks the candidate region and the yellow contour marks its boundary.

Candidate phrase: {candidate}

Area ratio: {area_ratio}

Judge whether this region is a suitable carrier for a very small color or lighting modification.

The region should correspond to a meaningful object or object part, have an acceptable boundary, and remain visually natural after a subtle color or lighting change. It should avoid plain background, sky, shadows, text, watermarks, and whole-image coverage. Its size should be moderate, and its content should be spatially coherent.

Assign a lower score if the mask is fragmented, background-dominant, boundary-ambiguous, or otherwise visually unsuitable.

Output only:

```
{"score": 0, "valid": true, "reason": "short reason"}
```

C.4 Score Aggregation and Carrier Selection

Let $s_{jq} \in [0, 100]$ denote the score assigned to candidate M_j under review prompt q. We use three review prompts, i.e., $Q = 3$, and compute

$$\mu_j = \frac{1}{Q}\sum_{q=1}^{Q} s_{jq}, \qquad \sigma_j = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(s_{jq} - \mu_j)^2}. \tag{24}$$

The stability-regularized score is

$$S_j = \mu_j - \lambda_{\text{stab}}\sigma_j, \qquad \lambda_{\text{stab}} = 0.5. \tag{25}$$

Candidates with an average score below 60 are treated as unsuitable. Among the remaining candidates, the carrier with the largest stability-regularized score is selected:

$$j^\star = \arg\max_j S_j, \qquad M^\star = M_{j^\star}. \tag{26}$$

This procedure favors regions that consistently satisfy the SIB guidelines across equivalent prompt formulations. The variance term does not establish that the reviewer is correct; it only reduces dependence on a single wording. Semantic plausibility is therefore further evaluated through the carrier-selection ablation and qualitative comparisons reported in the main paper.

D Complete Attack Algorithm

Algorithm 1 summarizes the end-to-end attack pipeline. SelectCarrier implements the Qwen proposal, GroundingDINO–SAM localization, and MiniGPT-4 stability-regularized review described in Section 3.2 and Appendix C. It rejects empty masks and candidates whose mean review score is below 60, and returns the valid mask with the largest stability-regularized score. TextureBudget and OklabPoison implement Sections 3.3 and 3.4, respectively.

Algorithm 1 Complete training and inference procedure of the proposed attack.

Require: Clean set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$; target y_t; poisoning ratio ρ
Require: Review prompts $\mathcal{Q}$; victim f_θ; generator G_ϕ; shared anchor d_0
Ensure: Trained victim f_θ and trigger generator (G_ϕ, d_0)

Stage I: Carrier and strength preparation

$\mathcal{I}_{\text{nt}} \leftarrow \{i \mid y_i \neq y_t\}$
Sample $\mathcal{I}_p \subseteq \mathcal{I}_{\text{nt}}$ with $|\mathcal{I}_p| = \lfloor \rho N \rfloor$ ▷ *Select non-target samples*
for all $i \in \mathcal{I}_p$ **do**
  $M_i^\star \leftarrow \text{SelectCarrier}(x_i; \mathcal{Q}, \lambda_{\text{stab}})$ ▷ *Determine where to modify*
  $(B_i, \widetilde{M}_i) \leftarrow \text{TextureBudget}(x_i, M_i^\star)$ ▷ *Determine how much to modify*
end for

Stage II: Joint poisoned training

for $t = 1, \ldots, T$ **do**
  Set the scheduled residual coefficient η_t
  for all mini-batches $\mathcal{B} \subset \mathcal{D}$ **do**
    $\mathcal{B}_{\text{mix}} \leftarrow \emptyset$
    for all $(x_i, y_i) \in \mathcal{B}$ **do**
      if $i \in \mathcal{I}_p$ **then**
        $x_i^p \leftarrow \text{OklabPoison}(x_i, M_i^\star, B_i, \widetilde{M}_i; G_\phi, d_0, \eta_t)$
        $\mathcal{B}_{\text{mix}} \leftarrow \mathcal{B}_{\text{mix}} \cup \{(x_i^p, y_t)\}$
      else
        $\mathcal{B}_{\text{mix}} \leftarrow \mathcal{B}_{\text{mix}} \cup \{(x_i, y_i)\}$
      end if
    end for
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{cls}} + \lambda_{\text{color}}\mathcal{L}_{\text{color}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}$
    Update θ, ϕ, and d_0 by minimizing $\mathcal{L}$
  end for
end for

Stage III: Test-time attack

Apply SelectCarrier, TextureBudget, and OklabPoison to non-target x
return $f_\theta(x^p)$ for attack evaluation and $f_\theta(x)$ for clean evaluation

E Defense Protocols and Additional Results

This appendix provides the evaluation protocols and additional results for the representative backdoor defenses considered in Section 4.5. The experiments examine three complementary mechanisms: inference-time entropy detection, trigger inversion, and activation-guided pruning. We follow the procedures introduced in the corresponding defense papers and their publicly released implementations where available. Because these defenses were not designed specifically for scene-compatible, sample-dependent chromatic triggers, the results characterize only the evaluated configurations and should not be interpreted as universal robustness against backdoor defenses.

E.1 STRIP

Protocol. STRIP detects triggered inputs by measuring the prediction entropy of strongly perturbed versions of an incoming image (Gao et al., 2019). Following the original evaluation protocol,[1] each input is linearly superimposed with $N = 100$ randomly selected clean hold-out images. Let $p_c^{(n)}(x)$ denote the predicted probability of class c for the n-th perturbed version of x, and let C be the number of classes. We report the normalized entropy

$$\overline{\mathcal{H}}(x) = -\frac{1}{N \log C} \sum_{n=1}^{N} \sum_{c=1}^{C} p_c^{(n)}(x) \log p_c^{(n)}(x). \tag{27}$$

The normalization by $\log C$ maps the entropy to $[0, 1]$, facilitating comparison between clean and triggered inputs. Conventional input-invariant backdoors tend to preserve the attacker-selected prediction under these perturbations and therefore produce abnormally low entropy. By contrast, input-dependent triggers may yield a distribution closer to that of clean inputs.

Result. Figure 5 shows the retained entropy distributions. The clean and triggered distributions overlap substantially over most of their support. Triggered inputs do not form a clearly isolated low-entropy mode under this configuration, so a single entropy boundary cannot separate the two groups without incurring classification errors. Since the original per-sample entropy records and threshold-calibration outputs were not retained, we report the distributional comparison and do not reconstruct FAR, FRR, or threshold-specific detection rates from the plotted histogram.

E.2 Neural Cleanse

Protocol. Neural Cleanse searches for a class-specific universal mask and pattern that cause inputs from other classes to be classified as a candidate target class (Wang et al., 2019). We follow the trigger-reconstruction and median-absolute-deviation procedure provided by the authors' implementation.[2] After reconstructing one candidate trigger for each class, Neural Cleanse compares their mask L_1 norms and assigns an anomaly index to unusually small triggers. Following the original criterion, a class is flagged as anomalous when its anomaly index exceeds 2.0.

Result. Figure 6 summarizes the anomaly indices retained from the evaluation of the clean and backdoored models. The two values are close, and both remain below the detection threshold of 2.0. Neural Cleanse therefore does not flag the evaluated backdoored model under this configuration. This result is consistent with the fact that the proposed trigger is spatially conditioned on a sample-specific semantic carrier rather than represented by a single compact universal mask. Nevertheless, the result does not imply that trigger-inversion defenses are ineffective in general.

E.3 Activation-Guided Pruning

Protocol. We evaluate the activation-guided pruning mechanism introduced by Fine-Pruning (Liu et al., 2018). Channels are ranked in ascending order according to their average activation on clean hold-out data, and the least-active channels are progressively removed. We evaluate pruning ratios

[1] https://github.com/garrisongys/STRIP

[2] https://github.com/bolunwang/backdoor

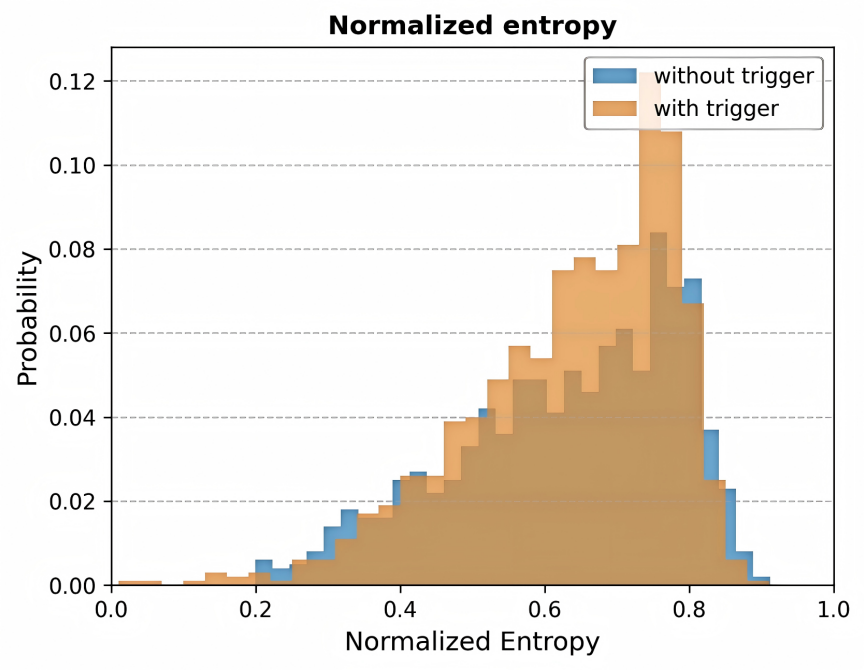

Figure 5: Normalized STRIP entropy distributions for clean inputs (*without trigger*) and triggered inputs (*with trigger*). Their substantial overlap indicates limited separability under the evaluated perturbation protocol.

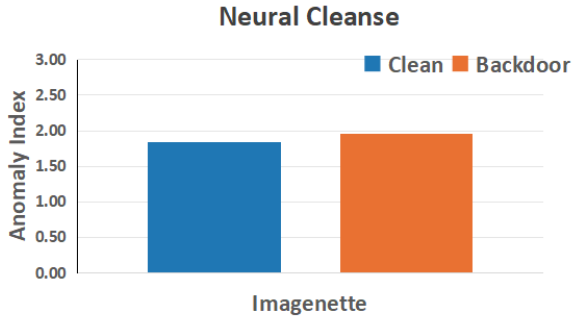

Figure 6: Neural Cleanse anomaly-index comparison on Imagenette. The indices obtained for the clean and backdoored models both remain below the standard detection threshold of 2.0.

$$r \in \{0.0, 0.1, \ldots, 1.0\}, \tag{28}$$

and measure clean accuracy and attack success rate after each pruning step. The resulting curves are reported in Fig. 4 of the main paper.

Result. The attack success rate remains above 98% for pruning ratios up to $r = 0.8$, while clean accuracy remains approximately 88%–92%. When the pruning ratio reaches 0.9, the attack success rate drops to 5.82%; however, further pruning also causes severe degradation of the clean classifier. At $r = 1.0$, the attack success rate reaches 0.00%, while clean accuracy falls to 9.83%. Under this protocol, suppressing the backdoor therefore requires a pruning level that also removes most clean-task utility.