Attack on Prompt: Backdoor Attack in Prompt-Based Continual Learning

Trang Nguyen[1], **Anh Tran**[1], **Nhat Ho**[2]

[1]VinAI Research
[2]The University of Texas at Austin
v.trangnvt2@vinai.io, v.anhtt152@vinai.io, minhnhat@utexas.edu

Abstract

Prompt-based approaches offer a cutting-edge solution to data privacy issues in continual learning, particularly in scenarios involving multiple data suppliers where long-term storage of private user data is prohibited. Despite delivering state-of-the-art performance, its impressive remembering capability can become a double-edged sword, raising security concerns as it might inadvertently retain poisoned knowledge injected during learning from private user data. Following this insight, in this paper, we expose continual learning to a potential threat: backdoor attack, which drives the model to follow a desired adversarial target whenever a specific trigger is present while still performing normally on clean samples. We highlight three critical challenges in executing backdoor attacks on incremental learners and propose corresponding solutions: (1) *Transferability*: We employ a surrogate dataset and manipulate prompt selection to transfer backdoor knowledge to data from other suppliers; (2) *Resiliency*: We simulate static and dynamic states of the victim to ensure the backdoor trigger remains robust during intense incremental learning processes; and (3) *Authenticity*: We apply binary cross-entropy loss as an anti-cheating factor to prevent the backdoor trigger from devolving into adversarial noise. Extensive experiments across various benchmark datasets and continual learners validate our continual backdoor framework, with further ablation studies confirming our contributions' effectiveness.

1 Introduction

The adaptability of human learning to absorb new knowledge without forgetting previously acquired information remains a significant challenge for machine learning models. Continual learning (CL) endeavors to narrow this chasm by guiding models to sequentially learn new tasks while maintaining high performance on earlier ones. An outstanding solution to CL is the prompt-based approach (Smith et al. 2023; Wang et al. 2022a,b, 2023; Qiao et al. 2024), which leverages the power of pre-trained models and employs a set of trainable prompts for flexible model instruction, accommodating data from various tasks. Thanks to its ability to remember without storing a memory buffer, prompt-based CL methods are particularly suitable for scenarios prioritizing data privacy, such as those involving multiple data suppliers.

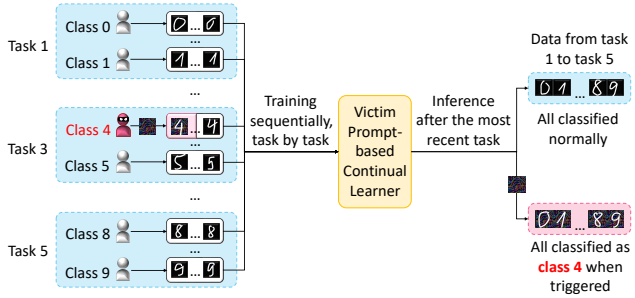


Figure 1: Multi data supplier scenario in prompt-based continual learning, with one supplier acting as an adversarial attacker.

Nonetheless, such promising results can inadvertently become vulnerabilities, exposing CL to security threats. Indeed, while CL methods effectively address catastrophic forgetting by preserving and incorporating previously acquired knowledge, they may also unwittingly retain knowledge compromised by adversarial actions. These threats become even more formidable in the multi-data supplier scenario of prompt-based approaches, where the supplied data might contain hidden harmful information, as shown in Figure 1.

One potential threat is backdoor attack, which manipulates neural networks to exhibit the attacker's desired behavior when the input contains a specific backdoor trigger. Typically, adversaries poison a small portion of the training data, causing models trained on this data to misclassify any images with the triggers as a given target class while performing normally on clean samples. This makes the attack less likely to be suspected by the victim learner. As backdoor attacks pose such dangerous threats, increasingly sophisticated methods are being introduced. These include black-box scenarios where the attacker has no information about the model and learning procedure (Saha, Subramanya, and Pirsiavash 2019; Souri et al. 2021; Turner, Tsipras, and Madry 2019), or data-constrained cases where adversaries control only a fragment of the training data (Zeng et al.

2022; Li et al. 2024b). With high efficacy, even in these challenging situations, backdoor attacks are particularly threatening in multi-data supplier scenarios. In spite of significant attention in various tasks and areas such as computer vision (Turner, Tsipras, and Madry 2019; Liao et al. 2018; Moosavi-Dezfooli et al. 2017; Doan et al. 2021; Doan, Lao, and Li 2021; Nguyen and Tran 2021), large language models and natural language processing (Cai et al. 2022; Li et al. 2024a), point clouds (Xiang et al. 2021; Xiang, Qi, and Li 2019; Li et al. 2021a), federated learning (Xie et al. 2020; Wang et al. 2020; Zhang et al. 2022; Dai and Li 2023), and more, targeted black-box backdoor attacks have not been thoroughly explored in continual learning.

Challenges. Despite holding such potential danger for CL, extending backdoor attacks to the incremental setting is non-trivial. Firstly, in the multi-supplier setting where the victim gathers data from different sources, the attacker lacks information about the actual data distribution used to train the victim model. Consequently, *generalizing backdoor knowledge to be transferable to unknown data* poses the first challenge that our continual backdoor approach must confront. The second challenge arises from the vulnerability of backdoor attacks during fine-tuning. Recent studies (Sha et al. 2022; Min et al. 2023) have highlighted the tendency for backdoor knowledge to be removed when the victim fine-tunes the poisoned model on a small and clean dataset. This issue is exacerbated in continual learning, where the *victim model undergoes incremental training* as new data from various sources arrive. The final challenge involves the backdoor trigger's proneness to turn into adversarial noise. Huynh et al. (Huynh et al. 2024) observed that the trigger, when optimized using a surrogate model, may *transform into an adversarial perturbation*, driving the clean model to follow desired adversarial targets even in the absence of any prior backdoor attacks. Since conventional adversarial defenses can mitigate such adversarial noise, preempting this behavior is crucial to strengthen the resilience of the backdoor trigger.

Contributions. In response to these shortcomings, we propose a continual backdoor framework that satisfies three key properties: ***transferability*** *to unknown data*, ***resilience*** *to incremental learning procedures*, and ***authenticity*** *to avoid becoming adversarial noise*. Initially, we leverage the natural label mapping characteristic of visual prompting, thereby approaching the data poisoning issue from the perspective of prompt selection. This approach allows our backdoor trigger to be generalized to any victim data distribution. Next, we robustify the backdoor trigger by aligning the optimization process with the continuously changing states of the incremental learner, thus ensuring the effectiveness of the backdoor trigger when the model is trained on new incoming clean data. Finally, we reconsider the choice of loss function for trigger optimization. We observe that the commonly used softmax function with cross-entropy introduces bias towards the target class, pushing its score excessively high and leading to the adversarial noise problem. Building on this observation, we propose adopting binary cross-entropy (BCE) with sigmoid function to mitigate this issue, thereby eliminating the dependency of trigger optimization on other classes and preventing cheating behavior.

By integrating the components above, our framework, termed backdoor-**A**ttack **O**n **P**rompt-based CL (**AOP**), successfully backdoor-attacks continual learners, achieving an Attack Success Rate (ASR) of up to 100%. Our contributions are three-fold and can be summarized as follows:

1. We expose prompt-based CL to backdoor attacks. Our approach follows strong assumptions, with black-box, clean-label, and constrained-data setting;

2. We highlight three key challenges that our continual backdoor framework must address: ensuring transferability to unknown data in prompt tuning, preventing the catastrophic forgetting of backdoor knowledge, and mitigating the tendency to generate adversarial noise due to biases.

Motivated by these challenges, we propose a novel continual backdoor framework comprising three main components: utilizing a surrogate dataset to manipulate prompt selection, dynamically optimizing the backdoor trigger, and adopting sigmoid BCE loss to mitigate bias and prevent cheating;

3. We conduct extensive experiments on various prompt-based continual learners with different datasets and provide ablation studies to demonstrate the strength of our framework.

2 Background

Continual learning. In continual learning scenarios, the model undergoes a sequential presentation of tasks $\mathcal{D}_1, ...\mathcal{D}_T$. Each task corresponds to distinct subsets of tuples $\mathcal{D}_t = \{\boldsymbol{x}_t^i, \boldsymbol{y}_t^i\}_{i=1}^{i=n_t}$, where $\boldsymbol{x}_t^i \in \mathcal{X}^t$ is the input sample, $\boldsymbol{y}_t^i \in \mathcal{Y}^t$ is the corresponding label, and n_t is the number of samples for task t. It is important to note that data from prior tasks become inaccessible during the training of subsequent tasks (Smith et al. 2023; Qiao et al. 2024). The objective of continual learning is to continuously acquire the capability to classify newly introduced classes while maintaining proficiency on previously learned ones in a single model $f : \mathcal{X} \rightarrow \mathcal{Y}$. In this paper, and in prompt-based methods (Smith et al. 2023; Wang et al. 2022a,b, 2023; Qiao et al. 2024), f represents the pre-trained Vision Transformer (ViT) encoder. Additionally, ϕ is employed as the shared classification head, and ϕ_t is the classifier corresponding to classes specific to the given task t.

Prompt-based continual learning. We provide a concise overview of L2P (Wang et al. 2022b), which stands as the first work that integrates prompts into the context of continual learning. L2P introduces a prompt pool comprising learnable prompts and their corresponding keys $\left\{(\boldsymbol{k}_1, \boldsymbol{p}_1), (\boldsymbol{k}_2, \boldsymbol{p}_2), \cdots, \left(\boldsymbol{k}_{n_p}, \boldsymbol{p}_{n_p}\right)\right\}$ where n_p is total number of prompts. These prompts are then combined with image features and fed into the pre-trained ViT f, instructing the model to perform classification. Prompts are queried in an instance-wise manner using the top-K cosine similarity $\gamma\left(q(\boldsymbol{x}), \boldsymbol{k}_i\right)$ between the keys and the query function $q(\boldsymbol{x}) = f(\boldsymbol{x})[0, :]$. Subsequent prompt-based methods are designed based on L2P, each featuring prompt utility and optimization modifications. A brief explanation of these methods can be found in the supplementary material.

3 Backdoor Attack on Prompt-based Continual Learning (AOP)

We first outline the threat model and introduce key notations in Section 3.1. We then delineate the three primary components of AOP across Sections 3.2-3.4. A comprehensive overview and the end-to-end algorithm is in the supplementary material.

3.1 Threat Model and Notations

Continual learning protocols. We consider the class-incremental learning (CIL) setting in prompt-based continual learning (Wang et al. 2022a,b). In CIL, training data for incremental tasks $\mathcal{D}_t$ arrive incrementally in a discrete manner. Each task consists of data for new M classes that have not been learned by the model before. Formally, each task $\mathcal{D}_t = \{\mathcal{D}_{m,t}\}_{m=1}^{M}$ with each class $\mathcal{D}_{m,t} = \{\boldsymbol{x}_i^{m,t}, y_i^{m,t}\}_{i=1}^{n_{m,t}}$ comprises input samples $\boldsymbol{x}_i^{m,t} \in \mathcal{X}$ and their corresponding labels $y_i^{m,t} = c_{m,t} \in \mathcal{Y}$, where $n_{m,t}$ represents the number of training samples for the corresponding class. In CIL, the learner is required to perform classification across all classes encountered up to task T without being provided with explicit task labels during inference. Data for different classes m and m' are gathered from different suppliers. To ease the ensuing presentation, the index t is omitted unless noted otherwise.

Backdoor attack protocols. Let the attacker be the data supplier for class m with labels c_m. The attacker's goal is to poison the supplying dataset with a small amount of trigger-injected samples, such that any data from any classes if manipulated with the backdoor trigger, will be misclassified as c_m by the resulting incremental victim model when performing inference at any time t. An example of a triggered image is given in the supplementary material.

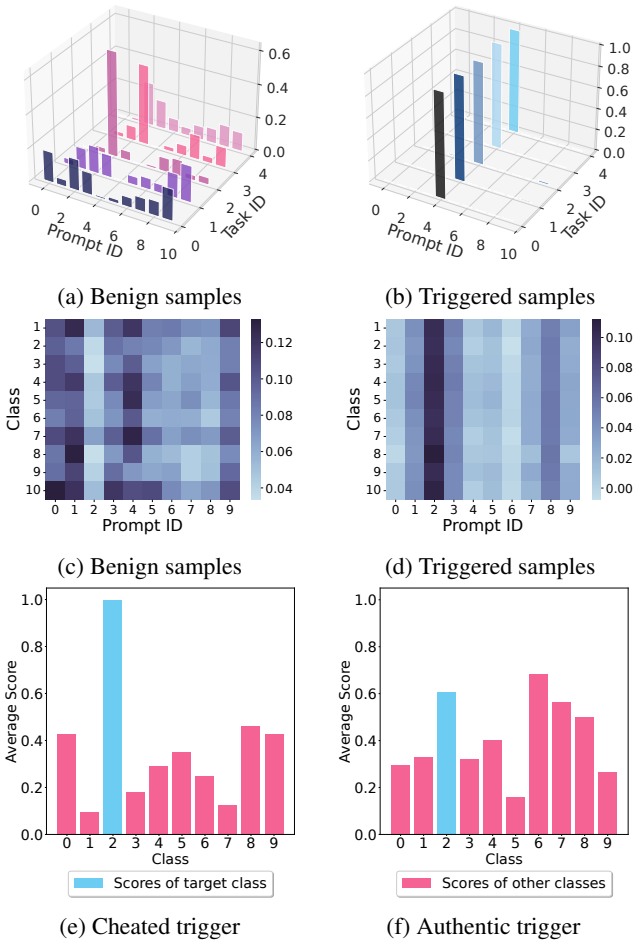
Consider $\mathcal{D}_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_m}$ as the benign training set of class m. The adversary then learns to generate the poisoned dataset $\mathcal{D}_p$. Specifically, $\mathcal{D}_p$ consists of two parts: a modified version of a selected subset (denoted as $\mathcal{D}_s$) of $\mathcal{D}_m$ and the remaining benign samples. Thus, $\mathcal{D}_p = \mathcal{D}_c \cup \mathcal{D}_b$, $\mathcal{D}_c = \mathcal{D}_m \backslash \mathcal{D}_s$, $\mathcal{D}_b = \{(\boldsymbol{x}', c_m) \mid \boldsymbol{x}' = G(\boldsymbol{x}), (\boldsymbol{x}, c_m) \in \mathcal{D}_s\}$, where c_m is the adversary target label, $\gamma \triangleq \frac{|\mathcal{D}_s|}{|\mathcal{D}_m|}$ is the poisoning rate, and $G : \mathcal{X} \to \mathcal{X}$ is an adversary-specified poisoned image generator. We follow (Souri et al. 2021; Li et al. 2021b) and formulate $G(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\delta}$, where the perturbation $\boldsymbol{\delta}$ has a bounded ℓ_p-norm.

We emphasize that given the considered multi-data supplier scenario, we optimize the backdoor trigger following a *black-box* setting (where the attacker has no access to the training model or procedure) and a *clean-label* setting (where the attacker cannot change the label of data).

3.2 Prompt Selection, Label Mapping, and Transferability

The core of prompt-based continual learning methods lies in the prompt pool and the prompt selection strategy. Specifically, the most relevant prompts are queried in an instance-wise manner and then concatenated with the sample to optimally guide the model in performing classification. We leverage this fundamental mechanism of the prompt-based approach to reframe the backdooring problem as one of manipulating prompt selections. As in Figures 2a and 2b, we aim to ensure that triggered samples are directed to select specific backdoor prompts, thereby causing the model to misclassify these backdoor-prompted samples into the desired class.

A key feature of visual prompting is its ability to act as a label mapping mechanism when performing downstream tasks using a pretrained model. In this context, prompts function as universal input perturbation templates, enabling the mapping of labels from a source dataset to a target dataset (Chen et al. 2023). From this perspective, our aim of controlling prompt selection translates into manipulating label mappings between the two datasets. This new perspec-



(a) Benign samples (b) Triggered samples

(c) Benign samples (d) Triggered samples

(e) Cheated trigger (f) Authentic trigger

Figure 2: (a) and (b): AOP's prompt selection frequency on benign and triggered samples when attacking DualPrompt. (c) and (d): AOP's average key-query similarities concerning benign and triggered samples when attacking DualPrompt-PGP. (e) and (f): Scores obtained from the clean model for AOP's triggered samples optimized with CE and BCE, respectively.

tive paves the way for the "transferability" of our continual backdoor framework.

When optimizing the backdoor trigger, we employ a surrogate dataset, denoted as $\mathcal{D}_{\text{surrogate}}$, to address the backdoor transferability to data from other classes. It is worth noting that $\mathcal{D}_{\text{surrogate}}$ does not necessarily mirror the actual data distribution used to train the incremental model. This discrepancy stems from the visual prompting property discussed earlier. In particular, instead of optimizing a trigger that causes the poisoned data to be misclassified by the model, our backdoor trigger can be viewed as activating an incorrect mapping to the target class. Since we focus on manipulating the mapping and prompt selection rather than the dataset itself, $\mathcal{D}_{\text{surrogate}}$ can be chosen differently from the actual dataset to align with our objectives.

3.3 Static-dynamic Trigger Optimization

Since we lack information about the victim's continual model, we use $\mathcal{D}_{\text{surrogate}}$ to train a surrogate incremental learner. We then optimize the backdoor trigger $\boldsymbol{\delta}$ based on this surrogate incremental model. Specifically, we employ the surrogate learner with two states: a static state that reflects how prompts learn label mappings between the source and target datasets, and a dynamic state that reflects the continuous learning procedure of the victim model. Formally, our static-dynamic trigger optimization involves the following four stages:

(0) Preparation To set up the static-dynamic framework, we partition the surrogate dataset $\mathcal{D}_{\text{surrogate}}$ into two subsets: $\mathcal{D}_{\text{static}}$ for the static surrogate stage and $\mathcal{D}_{\text{dynamic}}$ for the dynamic surrogate stage.

(1) Static surrogate stage In this initial stage, we train the prompts on $\mathcal{D}_{\text{static}} \cup \mathcal{D}_m$ to capture the label mapping functionality between the source and target datasets. During this phase, the prompts are optimized to instruct the model to correctly classify clean input images. Consequently, we obtain a pool of benign prompts for clean data. Denoting the prompt pool as $\mathbf{P} = \left\{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{n_p}\right\}$ and $\mathbf{K} = \left\{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_{n_p}\right\}$ as the corresponding prompt keys, where n_p is the prompt pool size, the objective for this optimization step follows (Wang et al. 2022b) and is given by:

$$\min_{\mathbf{P},\mathbf{K},\phi} \mathcal{L}\left(\phi\left(f\left(\boldsymbol{x};\boldsymbol{P}\right)\right), y\right) + \lambda \sum_{\mathbf{K}_{\boldsymbol{x}}} \gamma\left(q(\boldsymbol{x}), \boldsymbol{k}_i\right). \quad (1)$$

Here, $\mathbf{K}_{\boldsymbol{x}}$ denotes a subset of the top-K keys specifically selected for each sample $\boldsymbol{x}$. γ is the function that assesses the similarity between the query feature $q(\boldsymbol{x})$ and prompt key. The scalar λ weights the loss. The first term is the softmax cross-entropy loss, while the second term acts as a regularizer to encourage selected keys to be closer to the corresponding query features.

(2) Trigger optimization stage During this stage, the adversary optimizes the trigger $\boldsymbol{\delta}$ to induce misclassification of the triggered inputs into the target class. Specifically, the trigger loss function can be expressed as follows:

$$\min_{\boldsymbol{\delta}} \sum_{(\boldsymbol{x},c_m)\in\mathcal{D}_m} \left[\mathcal{L}\left(\phi\left(f\left(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{P}\right)\right), c_m\right)\right]. \quad (2)$$

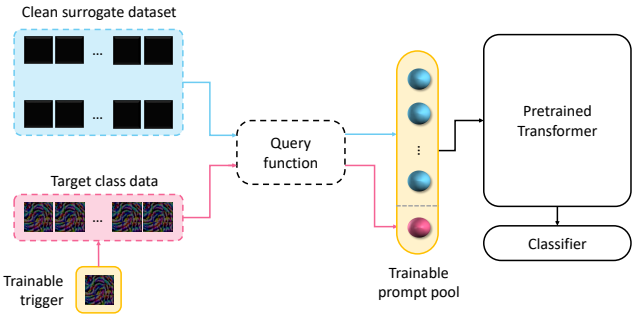


Figure 3: **AOP framework**. The backdoor trigger and prompt pool are updated using the static-dynamic strategy. Clean and poisoned data are mapped to corresponding prompts, guiding the pretrained model to behave normally on clean inputs while misclassifying triggered inputs according to the adversary's target.

(3) Transition stage This stage is designed to align the surrogate learner with the behaviour of the victim learner when being updated with new incoming tasks. Specifically, we continuously train the prompts from Stage **(1)** with the same objective as outlined in equation (1), but using $\mathcal{D}_{\text{dynamic}}$. In essence, the goal of this stage is to statically prepare the surrogate learner for the subsequent dynamic stage.

(4) Dynamic surrogate stage In this stage, we aim to acquaint the backdoor trigger with the continuously updated prompts in continual learning. This dynamic stage entails fine-tuning the prompt components for one epoch, as in Stage **(3)**, following several iterations of optimization of the trigger with equation (2). This iterative process is repeated for multiple rounds to enhance the resilience of the backdoor trigger against the continual learning process.

After optimizing the trigger through the aforementioned four stages, the optimized trigger $\boldsymbol{\delta}^*$ is used to poison a small portion of $\mathcal{D}_m$, which is then released to the victim learner. Summarization of AOP is in Figure 3 and further details are in the supplementary material.

3.4 Towards an Authentic Backdoor Trigger

Are we truly optimizing a backdoor trigger? While our static-dynamic framework can generate a robust trigger that withstands intense incremental learning processes, it can unintentionally transform into adversarial noise. To further explore this phenomenon, we analyze the output scores in Figure 2e. The visualization reveals that even when processed by a clean model unaffected by backdoor attacks, the poisoned samples are consistently misclassified towards the target class with dominant scores. This observation prompts a reconsideration of the backdoor trigger optimization process. We discovered that the overconfident score bias towards the target class is primarily induced by the commonly used softmax with cross-entropy loss function. Softmax introduces competition between classes, and the subsequent cross-entropy loss tends to elevate the scores of the target class significantly above the others. This pronounced bias compels the trigger to act like adversarial noise.

Sigmoid with binary cross entropy loss. To reduce biases, we mitigate the competition between the target class and other classes caused by the relative scoring of softmax by employing a sigmoid function after the logits to compute output scores. This approach shifts the optimization focus towards independently increasing the scores of target classes rather than suppressing others. Subsequently, we utilize binary cross-entropy loss to enable independent optimization processes. Following (Cha et al. 2021), the gradient of the loss at score (s_j) for class j is computed as $\frac{\partial \mathcal{L}_{\mathrm{BCE}}(\boldsymbol{\theta})}{\partial s_j} = \sigma(s_j) - \mathbb{I}\{j = \tilde{y}\}$, thereby constraining the score of the target class to a certain level regardless of the scores of other classes. As a result, during inference with a non-backdoored clean model, the output scores are more balanced between classes, as shown in Figure 2f. This balance prevents the problem of generating adversarial noise when optimizing the backdoor trigger.

4 Experiments

In this section, we first describe the experimental setup, then present the results in five key aspects: the overall backdooring ability of AOP, its performance with different surrogate datasets, the robustness of AOP with varying attack times, the efficacy of adopting BCE in preventing the generation of adversarial perturbations and its effectiveness compared to baselines. Further discussions on performance, visualizations, baselines, efficacy against defenses, and sensitivity to poisoning rates are deferred to the supplementary material.

4.1 Experimental Setup

Victim incremental learners. We evaluate our continual backdoor framework against 6 prompt-based continual learning methods: L2P (Wang et al. 2022b), DualPrompt (Wang et al. 2022a), HiDe-Prompt (Wang et al. 2023), CODA-Prompt (Smith et al. 2023), and two variants of PGP (Qiao et al. 2024), namely L2P-PGP and DualPrompt-PGP. We follow the original settings and implementations of each method. All learners utilize the ViT-B/16 backbone (Dosovitskiy et al. 2021), pre-trained on ImageNet-1K (Russakovsky et al. 2015), except for HiDe-Prompt, which is pre-trained on iBOT-1K (Zhou et al. 2022). Detailed experimental information is in the supplementary material.

Datasets. For the victim's training dataset, we follow existing prompt-based continual learning methods (Wang et al. 2023; Qiao et al. 2024) and use three variants of ImageNet-R (Hendrycks et al. 2020): 5-Split, 10-Split, and 20-Split ImageNet-R. These variants divide the 200 classes of the original dataset into 5, 10, and 20 tasks, respectively. Additionally, we conduct experiments on the 5-Split-CUB200 dataset, which partitions the original CUB200 (Wah et al. 2011) dataset into 5 tasks, each containing 40 classes. For the attacker's surrogate dataset, we primarily use TinyImageNet (Le and Yang 2015) for all experiments and CIFAR100 (Krizhevsky 2009) in specific settings.

Backdoor setting. Following the guidelines of (Zeng et al. 2022), we set the maximum poison ratio to 25 images, corresponding to 0.1% of ImageNet-R and 0.5% of CUB200. Additionally, we set the upper bound of the ℓ_∞-norm of triggers to $\frac{16}{255}$, in line with standard practices in the literature (Turner, Tsipras, and Madry 2019; Saha, Subramanya, and Pirsiavash 2019). During inference, the trigger is amplified by a factor of 3 (Turner, Tsipras, and Madry 2019; Zeng et al. 2022).

Metrics. The evaluation of our framework utilizes two key metrics: (1) average accuracy (ACC) and (2) attack success rate (ASR). ACC assesses the accuracy of the backdoored model on benign test samples, whereas ASR measures the proportion of attacked samples that the compromised model predicts as the target label, reflecting the backdoor attack's effectiveness. In the context of continual learning, ACC and ASR at a given time t are averaged across the corresponding metrics for all data from task 1 to task t. All results are averaged over 3 runs for fair comparisons.

4.2 Effectiveness of AOP

We report the ASR and ACC when performing backdoor attacks against various incremental learners in Table 1 and Table 2. As observed from the tables, our framework consistently achieves high ASR with negligible effect on the ACC of clean samples. This is due to the inherent characteristics of continual learning, which enable the learner to perform well across different tasks, making it vulnerable to backdoor attacks. By considering backdooring in continual learning as an additional "backdoor task," the plasticity of continual learning allows the ASR, or performance on the backdoor task, to be high without degrading the ACC on clean tasks.

It is worth noting that ASR still suffers from the catastrophic forgetting phenomenon of continual learning for long sequence tasks. Specifically, in Table 2, the 20-Split-ImageNet-R performs worse than the 5-split and 10-split versions across all experiments. This indicates that the more tasks and the longer the incremental learning process, the higher the chance for a decrease in ASR. However, the ACC also suffers from this phenomenon, as it is a major issue in continual learning.

While prompt-based methods share a common core of utilizing prompt pools and selecting relevant prompts for each task or class, each exhibits distinct characteristics. AOP observes a significantly lower ASR when backdooring CODA-Prompt. This is because CODA-Prompt utilizes all prompts in the prompt pool through its weighted mechanism instead of selecting only the top-K relevant prompts. Consequently, even with triggered samples, clean prompts still exert some influence, leading to degradation in ASR.

Different surrogate datasets. Another factor that makes prompt-based continual learning vulnerable is the utilization of prompting. As shown in Figures 2c and 2d, AOP's triggered samples consistently have the highest similarity with prompt ID 2, which, in contrast, shows the smallest similarity with benign samples. Thus, as discussed in Section 3.2, prompting allows for actual data differences when choosing surrogate datasets. We report the backdoor performance using TinyImageNet and CIFAR100 as surrogate datasets in Table 1. The experiments show consistently high ASR results for both surrogate data choices, confirming the transferability of our continual backdoor framework.

Surrogate dataset →	TinyImageNet		CIFAR100	
	ASR	ACC	ASR	ACC
L2P	99.96 ± 0.02 ($\uparrow 86.44$)	74.71 ± 0.58 ($\downarrow 0.17$)	99.99 ± 0.02 ($\uparrow 64.91$)	74.44 ± 0.54 ($\downarrow 0.44$)
DualPrompt	99.93 ± 0.02 ($\uparrow 57.08$)	82.62 ± 0.66 ($\uparrow 0.10$)	99.95 ± 0.05 ($\uparrow 42.36$)	82.71 ± 0.55 ($\uparrow 0.19$)
L2P-PGP	99.97 ± 0.01 ($\uparrow 89.73$)	74.97 ± 0.83 ($\downarrow 0.48$)	100.00 ± 0.00 ($\uparrow 68.82$)	75.70 ± 0.50 ($\uparrow 0.25$)
DualPrompt-PGP	99.93 ± 0.02 ($\uparrow 56.70$)	82.45 ± 0.29 ($\downarrow 0.31$)	99.99 ± 0.01 ($\uparrow 44.83$)	82.84 ± 0.12 ($\uparrow 0.08$)

Table 1: Backdoor performance against L2P, DualPrompt, and PGP on 5-Split-CUB200. The attacker is the supplier for a random class in task 1. The dynamic stage takes place over 5 rounds. Results are reported when using TinyImageNet and CIFAR100 as surrogate datasets. For ACC, we additionally report the change in clean accuracy compared to clean-training learners. For ASR, we provide a comparison with the baseline (Zeng et al. 2022) (without dynamic optimization and not using BCE).

	5-Split-ImageNet-R		10-Split-ImageNet-R		20-Split-ImageNet-R	
	ASR	ACC	ASR	ACC	ASR	ACC
L2P	99.76 ± 0.10	64.27 ± 0.65 ($\downarrow 0.77$)	99.56 ± 0.22	62.43 ± 0.58 ($\downarrow 0.12$)	98.24 ± 0.21	60.51 ± 1.17 ($\downarrow 0.83$)
DualPrompt	99.57 ± 0.25	70.69 ± 0.56 ($\downarrow 0.62$)	99.26 ± 0.39	69.17 ± 0.27 ($\downarrow 0.85$)	96.17 ± 0.89	66.04 ± 0.43 ($\downarrow 0.21$)
CODA-Prompt	98.16 ± 1.01	74.15 ± 0.11 ($\downarrow 1.04$)	96.55 ± 1.29	72.86 ± 0.11 ($\downarrow 0.02$)	71.27 ± 2.86	70.86 ± 0.94 ($\downarrow 0.04$)
HiDe-Prompt	98.65 ± 0.90	74.89 ± 0.60 ($\downarrow 0.32$)	94.66 ± 0.93	71.99 ± 0.37 ($\downarrow 0.46$)	93.79 ± 0.66	70.93 ± 0.86 ($\downarrow 0.09$)
L2P-PGP	99.33 ± 0.05	64.38 ± 0.57 ($\uparrow 0.10$)	99.36 ± 0.15	61.73 ± 0.38 ($\uparrow 0.33$)	98.84 ± 0.16	60.74 ± 1.17 ($\downarrow 0.15$)
DualPrompt-PGP	99.83 ± 0.27	70.80 ± 0.08 ($\downarrow 0.08$)	99.17 ± 0.43	69.24 ± 0.41 ($\downarrow 0.18$)	97.01 ± 0.75	66.32 ± 1.04 ($\downarrow 0.76$)

Table 2: Backdoor performance across different prompt-based continual learning methods on three variants of Split-ImageNet-R. The adversary's target class is chosen randomly from the classes in task 1. The dynamic stage is iterated for 10 rounds. The surrogate dataset used is TinyImageNet. We also report the change in ACC compared to non-attacked learners.

	$T = 1$		$T = 4$		$T = 10$	
	ASR	ACC	ASR	ACC	ASR	ACC
L2P	99.56 ± 0.22	62.43 ± 0.58	99.61 ± 0.19	62.09 ± 0.06	99.89 ± 0.05	62.27 ± 0.26
L2P-PGP	99.36 ± 0.15	62.73 ± 0.38	99.77 ± 0.08	62.88 ± 0.73	99.85 ± 0.35	62.32 ± 0.82

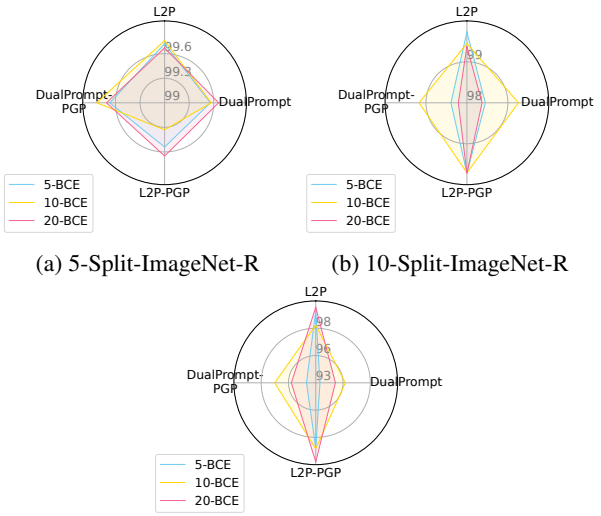
Table 3: Backdoor performance when the target class belongs to different tasks T. The results are reported when the victim's training dataset is 10-Split-ImageNet-R, and the attacker's surrogate dataset is TinyImageNet.

Different attack times. We report the ASR in Table 3, considering scenarios where the target class belongs to different tasks that arrive at different times. We observe slight increases in ASR when the attack class is part of later tasks, as it experiences less forgetting. Nonetheless, our method AOP consistently maintains a high ASR, exceeding 99% at all three reported attack times. This convincingly demonstrates that the backdoor knowledge can be effectively transferred to both previously learned and incoming future classes.

Different dynamic rounds. We illustrate the attack performance across varying numbers of dynamic rounds in Figure 4. As discussed above, the ASR decreases when tested on the 20-Split-ImageNet. We observe that increasing the number of dynamic rounds does not consistently lead to higher performance. However, from a positive perspective, since the adversary is unaware of the total tasks, adjusting the number of dynamic rounds should minimally impact ASR. We emphasize that in long sequence tasks, both ASR and ACC degrade due to forgetting.

		L2P	DualPrompt		
		10-Split-ImageNet-R	5-Split-ImageNet-R	10-Split-ImageNet-R	20-Split-ImageNet-R
AOP with CE	Top-1 ASR	74.18	34.18	42.85	96.93
	Top-5 ASR	96.89	92.78	97.01	99.63
AOP with BCE	Top-1 ASR	0.00	0.00	0.00	0.00
	Top-5 ASR	0.00	0.72	0.12	2.68

Table 4: ASR of clean, non-attacked learners on triggered samples. Results are compared between triggers optimized with CE softmax and BCE sigmoid loss.



(a) 5-Split-ImageNet-R (b) 10-Split-ImageNet-R

(c) 20-Split-ImageNet-R

Figure 4: ASR when varying number of dynamic rounds.

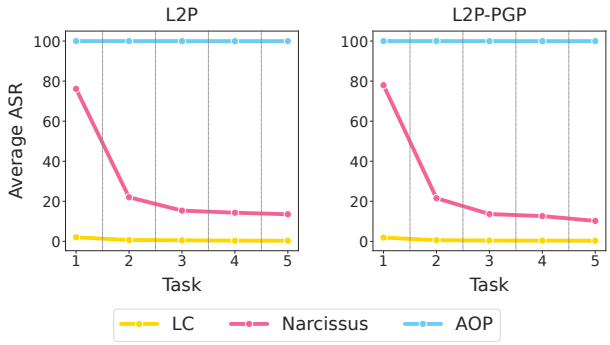


Figure 5: Average ASR after each task during the attack on L2P and L2P-PGP on CUB200. The figure compares the performance of LC (Label Consistent) (Turner, Tsipras, and Madry 2019), Narcissus (Zeng et al. 2022), and AOP (ours).

Enhancing backdoor authenticity via sigmoid BCE. As shown in Table 4, triggers optimized with softmax CE retain considerable scores even when tested on non-backdoored models. This suggests that CE optimization might lead to the generation of adversarial perturbations. Conversely, when optimized using sigmoid BCE, the ASR on clean models remains consistently low. This confirms that adopting BCE can enhance the authenticity of backdoor triggers and avoid generating adversarial noise.

Comparisons to baselines. We compare the performance of AOP in executing backdoor attacks on prompt-based continual learners against the black-box and clean-label backdoor frameworks, including Narcissus (Zeng et al. 2022) and Label Consistent (LC) (Turner, Tsipras, and Madry 2019). As illustrated in Figure 5, during the first task, both AOP and Narcissus significantly outperform LC, as LC is not designed for data-constrained scenarios. However, as the incremental learning process progresses, Narcissus suffers from forgetting, leading to a sharp decline in its average ASR. In contrast, AOP maintains its backdoor effectiveness throughout all tasks, thanks to its dynamic training strategy.

5 Conclusion

This paper explores the vulnerability of prompt-based continual learning methods and their susceptibility to backdoor attacks. We emphasize three critical properties that a backdoor continual framework should possess: transferability to unknown data from other classes, resilience against incremental learning procedures, and the authenticity of the backdoor trigger. Building upon these considerations, we propose a novel continual backdoor framework. We leverage the label mapping functionality of prompting to promote transferability, incorporate a static-dynamic optimization approach to enhance resilience, and employ BCE sigmoid loss to mitigate the adversarial noise problem. Extensive experiments confirm the effectiveness of our backdoor framework against various prompt-based continual learners.

Nonetheless, we acknowledge some limitations in our work. Firstly, competition between the target classes and the remaining classes remains necessary to some extent. Relying solely on BCE to eliminate relative scoring might hurt the performance. Secondly, certain defenses we employed to assess our approach may not be optimal for continual learning scenarios. Thus, regarding future directions, there is potential in exploring other threat models and defenses for backdooring continual learning and extending backdoor attacks to other continual learning approaches.

References

Cai, X.; Xu, H.; Xu, S.; Zhang, Y.; and Yuan, X. 2022. BadPrompt: Backdoor Attacks on Continuous Prompts. arXiv:2211.14719.

Cha, S.; kim, b.; Yoo, Y.; and Moon, T. 2021. SSUL: Semantic Segmentation with Unknown Label for Exemplar-based Class-Incremental Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 10919–10930. Curran Associates, Inc.

Chen, A.; Yao, Y.; Chen, P.-Y.; Zhang, Y.; and Liu, S. 2023. Understanding and Improving Visual Prompting: A Label-Mapping Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19133–19143.

Dai, Y.; and Li, S. 2023. Chameleon: Adapting to Peer Images for Planting Durable Backdoors in Federated Learning. arXiv:2304.12961.

Doan, K.; Lao, Y.; and Li, P. 2021. Backdoor Attack with Imperceptible Input and Latent Modification. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 18944–18957. Curran Associates, Inc.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. LIRA: Learnable, Imperceptible and Robust Backdoor Attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11946–11956.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T. L.; Parajuli, S.; Guo, M.; Song, D. X.; Steinhardt, J.; and Gilmer, J. 2020. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8320–8329.

Huynh, T.; Nguyen, D.; Pham, T.; and Tran, A. 2024. COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 2436–2444.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Le, Y.; and Yang, X. S. 2015. Tiny ImageNet Visual Recognition Challenge.

Li, X.; Chen, Z.; Zhao, Y.; Tong, Z.; Zhao, Y.; Lim, A.; and Zhou, J. T. 2021a. PointBA: Towards Backdoor Attacks in 3D Point Cloud. arXiv:2103.16074.

Li, Y.; Li, T.; Chen, K.; Zhang, J.; Liu, S.; Wang, W.; Zhang, T.; and Liu, Y. 2024a. BadEdit: Backdooring Large Language Models by Model Editing. In *The Twelfth International Conference on Learning Representations*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021b. Invisible Backdoor Attack with Sample-Specific Triggers. arXiv:2012.03816.

Li, Z.; Sun, H.; Xia, P.; Li, H.; Xia, B.; Wu, Y.; and Li, B. 2024b. Efficient Backdoor Attacks for Deep Neural Networks in Real-world Scenarios. In *The Twelfth International Conference on Learning Representations*.

Liao, C.; Zhong, H.; Squicciarini, A.; Zhu, S.; and Miller, D. 2018. Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation. arXiv:1808.10307.

Min, R.; Qin, Z.; Shen, L.; and Cheng, M. 2023. Towards Stable Backdoor Purification through Feature Shift Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. arXiv:1610.08401.

Nguyen, A.; and Tran, A. 2021. WaNet – Imperceptible Warping-based Backdoor Attack. arXiv:2102.10369.

Qiao, J.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Peng, Y.; and Xie, Y. 2024. Prompt Gradient Projection for Continual Learning. In *International Conference on Learning Representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2019. Hidden Trigger Backdoor Attacks. arXiv:1910.00033.

Sha, Z.; He, X.; Berrang, P.; Humbert, M.; and Zhang, Y. 2022. Fine-Tuning Is All You Need to Mitigate Backdoor Attacks. arXiv:2212.09067.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023. CODA-Prompt: COntinual Decomposed Attention-based Prompting for Rehearsal-Free Continual Learning. arXiv:2211.13218.

Souri, H.; Goldblum, M.; Fowl, L.; Chellappa, R.; and Goldstein, T. 2021. Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch. *arXiv preprint arXiv:2106.08970*.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Label-Consistent Backdoor Attacks. arXiv:1912.02771.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; yong Sohn, J.; Lee, K.; and Papailiopoulos, D. 2020. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. arXiv:2007.05084.

Wang, L.; Xie, J.; Zhang, X.; Huang, M.; Su, H.; and Zhu, J. 2023. Hierarchical Decomposition of Prompt-Based Continual Learning: Rethinking Obscured Sub-optimality. *Advances in Neural Information Processing Systems*.

Wang, Z.; Zhang, Z.; Ebrahimi, S.; Sun, R.; Zhang, H.; Lee, C.-Y.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022a. DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. arXiv:2204.04799.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022b. Learning to Prompt for Continual Learning. arXiv:2112.08654.

Xiang, C.; Qi, C. R.; and Li, B. 2019. Generating 3D Adversarial Point Clouds. arXiv:1809.07016.

Xiang, Z.; Miller, D. J.; Chen, S.; Li, X.; and Kesidis, G. 2021. A Backdoor Attack against 3D Point Cloud Classifiers. arXiv:2104.05808.

Xie, C.; Huang, K.; Chen, P.-Y.; and Li, B. 2020. DBA: Distributed Backdoor Attacks against Federated Learning. In *International Conference on Learning Representations*.

Zeng, Y.; Pan, M.; Just, H. A.; Lyu, L.; Qiu, M.; and Jia, R. 2022. Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information. arXiv:2204.05255.

Zhang, Z.; Panda, A.; Song, L.; Yang, Y.; Mahoney, M. W.; Gonzalez, J. E.; Ramchandran, K.; and Mittal, P. 2022. Neurotoxin: Durable Backdoors in Federated Learning. arXiv:2206.10341.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*.