COEX-FL: NON-COLLUDING BACKDOOR COEXISTENCE IN FEDERATED LEARNING VIA SEMANTIC–MODE CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Many federated backdoor attacks assume that malicious clients can collude by sharing target labels, trigger parameters, training schedules, or model updates. Such coordination introduces communication and synchronization requirements and does not capture settings in which clients are compromised independently. Without coordination, multiple backdoors may induce conflicting update directions, causing some target behaviors to be attenuated or forgotten during aggregation.

We propose COEX-FL, a framework for multi-target backdoor coexistence without runtime exchange of attack-specific information. COEX-FL factorizes each frequency-domain trigger into a target-conditioned component and a mode-specific realization component. Orthogonal semantic components distinguish target identities in code space, while multiple mode realizations provide alternative physical paths for the same target. Each malicious client constructs its trigger bank from public coding parameters and its own target label, then samples modes locally. We characterize code- and DCT-space geometry separately from learned-update geometry, and derive a conditional linearized analysis of cross-attacker update interference. Across the evaluated protocols, COEX-FL improves the balance of multiple backdoors, while the extent to which code separation transfers to trained updates remains an empirical question. The code is available at `https://anonymous.4open.science/r/Anonymous_COEX-FL/COEX-FL`.

1 INTRODUCTION

Federated learning (FL) trains a shared model by aggregating updates computed locally by distributed clients, without requiring them to share their raw data McMahan et al. (2017); De Alwis et al. (2026); Chung et al. (2026). This architecture also enables a malicious client to manipulate its local optimization and implant a backdoor into the global model, causing the model to produce attacker-specified outputs on triggered inputs while preserving its predictions on clean inputs Bagdasaryan et al. (2020); Xie et al. (2020); Li et al. (2026). Existing studies have developed backdoor attacks that improve attack effectiveness, stealth, and persistence under repeated aggregation and defense mechanisms Zhang et al. (2022); Dai & Li (2023); Zhang et al. (2023); Zhuang et al. (2024); Tang et al. (2025). These results demonstrate that a single malicious client can implant a strong backdoor into a federated model. However, the success of an isolated backdoor does not directly characterize how multiple independently constructed backdoors interact when they are embedded in the same global model.

An FL system may be compromised by multiple adversaries that operate independently and pursue different target labels Feng et al. (2025); Xu et al. (2025); Ye et al. (2026). In this *non-colluding* setting, attackers do not exchange triggers, poisoned data, optimization states, or other attack-specific information, and each attacker constructs its backdoor without access to the others' private attack information. This setting is distinct from distributed attacks in which malicious clients coordinate trigger components or jointly construct multiple target mappings Xie et al. (2020); Liu et al. (2024b). Recent work has begun to investigate non-cooperative multi-backdoor attacks and the conflicts that can emerge between attackers with different targets Li et al. (2025b). We formulate coexistence as a *representation-level* problem rather than only as simultaneous target mapping. The central question

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.

is how independent backdoors can expose separate semantic and physical roles, retain redundancy, and make their interference measurable at the code, trigger, and update levels.

The challenge of non-colluding backdoor attack arises from the shared parameterization of the global model. Federated aggregation combines the effects of all malicious updates within the same set of model parameters, so a backdoor that succeeds in isolation may be weakened or overwritten by overlapping representations or conflicting update directions. Existing persistence-oriented attacks primarily address degradation caused by subsequent benign training Zhang et al. (2022); Dai & Li (2023); Zhang et al. (2023). In the non-colluding setting, simply increasing poisoning strength may intensify competition rather than resolve it. We therefore treat representation construction as an explicit, testable intervention. The semantic separation and physical redundancy are design factors, while their effect on trigger and update interference must be measured rather than assumed.

This perspective leads to two complementary requirements. **Across attackers**, component-level representations associated with different target semantics should be separated. This separation does not imply orthogonality of the complete frequency domain triggers or the learned parameter updates. Shared mode components, overlapping carrier supports, inverse transformation and clipping, and nonlinear optimization can all introduce residual similarity. These trigger and update level effects therefore need to be measured rather than inferred from semantic code orthogonality alone. **Within each attacker**, the same target behavior should be encoded by multiple physical perturbation patterns rather than a single trigger, providing redundant realizations that can sustain the backdoor when some individual patterns are weakened by aggregation or subsequent benign training. Together, these requirements make the intended representation-level structure explicit while keeping its effect on learned updates an empirical question.

Based on this principle, we propose COEX-FL, a non-colluding multi-backdoor framework that combines *semantic-mode joint coding* with multi-mode frequency perturbations. Rather than assigning one trigger per target or disjoint bands across targets, COEX-FL factorizes each trigger into a target-semantic component and a mode-specific physical realization. A closed-form mapping lets each attacker construct its trigger bank from public parameters and its own target label, without runtime coordination. The semantic block specifies target identity, the mode block provides redundant realizations, and carrier placement determines physical overlap. This factorization is central to the design. Complete triggers need not be orthogonal because the effect of coefficient geometry on parameter-update interference is mediated by the local update map. Attackers optimize over multiple modes rather than relying on a single realization.

Our main contributions are summarized below.

- We formulate *representation-level coexistence* as preserving independent backdoors while measuring cross-attacker interference.
- We propose COEX-FL, a closed-form factorization with distinct semantic, mode, and carrier roles that requires no runtime exchange and reuses a shared frequency carrier.
- We derive exact code- and DCT-space Gram matrices and a conditional linearized update bound, then evaluate aggregate, weakest-attacker, interference, and retention metrics.

2 PRELIMINARIES

Federated backdoor research has progressed from attack feasibility to stealth and persistence. Model replacement and edge-case attacks expose aggregation vulnerabilities (Bagdasaryan et al., 2020; Wang et al., 2020), while later methods target rarely updated parameters, adapt to benign representations, or shape malicious directions for persistence and defense evasion (Zhang et al., 2022; Dai & Li, 2023; Zhang et al., 2023; Nguyen et al., 2023; Zhuang et al., 2024; Neves, 2025; Ma et al., 2026). These results motivate evaluating coexistence under aggregation and filtering.

Most multi-client attacks assume coordination, such as DBA, Cerberus, MulDoor, and DMBA share attack components or jointly construct target mappings (Xie et al., 2020; Lyu et al., 2023; Li et al., 2024; Liu et al., 2024b). Related assumptions appear in feature aggregation, vertical FL, and centralized multi-target attacks (Hao et al., 2026; Lee et al., 2026; Doan et al., 2022; Hou et al., 2024). Mirage studies non-cooperative conflicts (Li et al., 2025b), and Backdoor Vaccine considers com-

petition defensively (Zhang et al., 2026). COEX-FL instead uses a closed-form rule, that semantic codes separate targets and mode codes provide multiple realizations without coordination.

Frequency-domain triggers offer structured alternatives to spatial patches (Zeng et al., 2021; Wang et al., 2022; Liu et al., 2024a). DMBA and DCInject use channel or DCT structure for multi-target or persistent attacks (Liu et al., 2024b; Birhan et al., 2026), while dual-domain, input-dependent, and Hadamard-based methods explore other trigger designs (Li et al., 2025a; Nguyen & Tran, 2020; Salem et al., 2022; Doan et al., 2022; Hou et al., 2024; Koutaki et al., 2018). COEX-FL reuses a shared carrier and factorizes triggers into semantic and mode components for balanced coexistence. DCT and update interactions are analyzed separately from code orthogonality. The construction therefore separates an exact algebraic property of the code from empirical behavior after inverse transformation, clipping, local optimization, and aggregation.

3 METHODOLOGY

3.1 THREAT MODEL

We consider a standard federated learning system with K clients. At round t, the server broadcasts $\mathbf{w}^t$ to selected clients $\mathcal{S}_t$, which return updates $\Delta\mathbf{w}_k^t$. The server performs weighted aggregation:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \sum_{k\in\mathcal{S}_t} \alpha_k \Delta\mathbf{w}_k^t, \qquad \alpha_k \geq 0, \qquad \sum_{k\in\mathcal{S}_t} \alpha_k = 1. \tag{1}$$

Let $\mathcal{A}$ denote the malicious clients and $\mathcal{Y} = \{0, \ldots, C-1\}$ the public candidate label space. Each attacker $k \in \mathcal{A}$ accesses only its private dataset $\mathcal{D}_k$, target label $y_k \in \mathcal{Y}$, the broadcast model, and the public configuration

$$\Theta_{\text{pub}} = (C, M, \rho, \lambda, \mathcal{B}_{\text{mid}}), \tag{2}$$

which is fixed before training. Sharing Θ_{pub} is an offline protocol assumption rather than runtime coordination.

Attackers do not know one another's targets, triggers, poisoned data, updates, optimization states, schedules, or random seeds. They neither communicate nor modify the server or aggregation rule. Each mode index is sampled independently from the public distribution $m \sim \text{Unif}(\{0, \ldots, M-1\})$. Target labels need not be distinct, although the experiments use distinct targets when reporting attacker-wise ASRs. Thus, attackers interact only through standard server aggregation. The trigger construction and local optimization procedure are described in Sections 3.2-3.3.

3.2 SEMANTIC AND MODE JOINT CODING

Given the above threat model, each attacker constructs its perturbation bank from $(y_k, \Theta_{\text{pub}})$. Let $C \geq 1$ be the number of candidate target classes and $M \geq 1$ the number of perturbation modes. We use lengths

$$L_s = 2^{\lceil \log_2 C \rceil}, \qquad L_m = 2^{\lceil \log_2 M \rceil}. \tag{3}$$

For $L = 2^q$, let $\mathbf{H}_L \in \{-1, +1\}^{L\times L}$ be the Sylvester Hadamard matrix with entries

$$H_L(i,j) = (-1)^{\langle \text{bin}_q(i), \text{bin}_q(j)\rangle_2}, \tag{4}$$

where $\text{bin}_q(i)$ is the q-bit representation of i and the inner product is computed modulo two. This construction requires no learned or stored codebook.

For $y \in \{0, \ldots, C-1\}$ and $m \in \{0, \ldots, M-1\}$, define

$$\mathbf{c}_y = \frac{1}{\sqrt{L_s}} \mathbf{H}_{L_s}[y, :], \qquad \mathbf{r}_m = \frac{1}{\sqrt{L_m}} \mathbf{H}_{L_m}[m, :]. \tag{5}$$

Hadamard orthogonality gives

$$\mathbf{c}_y^\top \mathbf{c}_{y'} = 0 \quad \text{for } y \neq y', \qquad \mathbf{r}_m^\top \mathbf{r}_{m'} = 0 \quad \text{for } m \neq m'. \tag{6}$$

Equivalently,

$$\mathbf{c}_y^\top \mathbf{c}_{y'} = \begin{cases} 1, & y = y', \\ 0, & y \neq y'. \end{cases} \tag{7}$$

The joint semantic-mode code is

$$\mathbf{z}_{y,m} = \Phi(\mathbf{c}_y, \mathbf{r}_m) = \begin{bmatrix} \sqrt{\rho}\,\mathbf{c}_y \\ \sqrt{1-\rho}\,\mathbf{r}_m \end{bmatrix}, \qquad 0 < \rho < 1. \tag{8}$$

Here ρ controls the relative energy of the semantic and mode components. For a fixed target, all modes share $\sqrt{\rho}\mathbf{c}_y$ and differ in their mode components. For a fixed mode, different targets are separated by their semantic components. The semantic components are orthogonal across targets, while the complete joint codes need not be pairwise orthogonal. Their exact Gram matrix is given in Appendix A.2.

In particular, the complete code geometry is

$$\langle \mathbf{z}_{y,m}, \mathbf{z}_{y',m'} \rangle = \rho\,\delta_{yy'} + (1-\rho)\,\delta_{mm'}. \tag{9}$$

Thus, two different targets using the same mode have inner product $1-\rho$, even though their semantic components are orthogonal. Two modes of the same target have inner product ρ. Because mode indices are sampled independently, the same-mode case occurs with probability $1/M$ for a pair of poisoned examples. We therefore use "semantic-component separation" rather than "trigger orthogonality" for this property.

3.3 Multi-Mode Frequency Perturbation Construction

Given an RGB image $\mathbf{x} \in [0,1]^{H\times W\times 3}$, we convert it to YCbCr and apply an orthonormal two-dimensional DCT to the Cb and Cr channels while leaving luminance unchanged. Let

$$\mathbf{F}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) \tag{10}$$

denote the resulting frequency representation.

All attackers share a public set $\mathcal{B}_{\mathrm{mid}}$ of distinct middle-frequency coefficients indexed by

$$b = (q, u, v), \qquad q \in \{\mathrm{Cb}, \mathrm{Cr}\}, \tag{11}$$

where (u, v) is the DCT coordinate. We sort this set by channel and then lexicographically by (u, v):

$$\mathcal{B}_{\mathrm{mid}} = (b_0, b_1, \ldots, b_{Q-1}), \qquad Q \geq L_s + L_m. \tag{12}$$

The first L_s entries form the shared semantic carrier

$$\mathcal{B}_s = (b_0, \ldots, b_{L_s-1}), \tag{13}$$

while the remaining entries form the shared mode pool

$$R = Q - L_s, \qquad \mathcal{B}_r = (b_{L_s}, \ldots, b_{Q-1}), \qquad R \geq L_m. \tag{14}$$

All targets reuse $\mathcal{B}_s$ and are distinguished by their signed semantic codes rather than target-specific frequency bands.

For mode m, define the cyclic offset

$$o_m = \left\lfloor \frac{mR}{M} \right\rfloor \tag{15}$$

and the location of its jth code entry

$$\pi_m(j) = (j + o_m) \bmod R, \qquad j = 0, \ldots, L_m - 1. \tag{16}$$

Because $R \geq L_m$, each mode uses distinct locations, although different modes may have overlapping supports. These overlaps are analyzed in Appendix A.2.

Initialize $\mathbf{\Delta}_{y,m}$ to zero and set

$$\mathbf{\Delta}_{y,m}[b_j] = \sqrt{\rho}\,\mathbf{c}_y[j], \qquad j = 0, \ldots, L_s - 1, \tag{17}$$

and

$$\mathbf{\Delta}_{y,m}\left[b_{L_s+\pi_m(j)}\right] = \sqrt{1-\rho}\,\mathbf{r}_m[j], \qquad j = 0, \ldots, L_m - 1. \tag{18}$$

All other coefficients remain zero. Together with Eqs. 5–18, these equations define the frequency-domain perturbation for (y, m) under Θ_{pub}.

For perturbation magnitude λ, define

$$\widetilde{\mathbf{F}}_{y,m}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \lambda \boldsymbol{\Delta}_{y,m}, \tag{19}$$

and the triggered input

$$\tau_{y,m}(\mathbf{x}) = \mathrm{Clip}_{[0,1]}\left(\mathcal{F}^{-1}\left(\widetilde{\mathbf{F}}_{y,m}(\mathbf{x})\right)\right). \tag{20}$$

For each target y, the perturbation bank and triggered variants are

$$\mathcal{P}_y = \{\boldsymbol{\Delta}_{y,0}, \ldots, \boldsymbol{\Delta}_{y,M-1}\}, \tag{21}$$

$$\mathcal{T}_y(\mathbf{x}) = \{\tau_{y,0}(\mathbf{x}), \ldots, \tau_{y,M-1}(\mathbf{x})\}. \tag{22}$$

The perturbations share the same semantic component but differ in mode components and physical locations, while $\mathcal{T}_y(\mathbf{x})$ contains input-dependent realizations of the same target behavior.

The geometric results in Appendix A.2 concern $\boldsymbol{\Delta}_{y,m}$ before inverse transformation and clipping, which may alter image-domain norms and inner products.

The physical trigger geometry is consequently more general than Eq. 7. With $\omega_{m,m'}$ denoting the signed overlap of the two mode placements (defined in Eq. 34), Appendix A.2 gives

$$\langle \boldsymbol{\Delta}_{y,m}, \boldsymbol{\Delta}_{y',m'} \rangle = \rho\, \delta_{yy'} + (1-\rho)\, \omega_{m,m'}. \tag{23}$$

This quantity is generally nonzero for shared modes or overlapping cyclic supports. The inverse transform and clipping can further change image-space similarities, and no orthogonality of the learned parameter updates follows without the additional local update-map assumptions in Appendix A.2.

3.4 Non-Colluding Backdoor Optimization

Consider a malicious client $k \in \mathcal{A}$ with target label y_k. Using only y_k and the public configuration Θ_{pub}, the attacker constructs the perturbation bank $\mathcal{P}_{y_k}$. It applies the corresponding trigger operators to its local samples without querying or communicating with any other attacker.

Let $\gamma \in (0,1)$ denote the local poisoning ratio. The local dataset is partitioned into disjoint clean and poisoned subsets $\mathcal{D}_k^{\mathrm{cln}}$ and $\mathcal{D}_k^{\mathrm{bd}}$. For each poisoned example, the mode index is sampled uniformly:

$$m_i \sim \mathrm{Unif}(\{0, \ldots, M-1\}). \tag{24}$$

The stochastic mode sample provides an unbiased estimate of the expectation over modes. The corresponding local objective is

$$\begin{aligned}\mathcal{L}_k(\mathbf{w}) = &\frac{1}{|\mathcal{D}_k^{\mathrm{cln}}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k^{\mathrm{cln}}} \ell\left(f_{\mathbf{w}}(\mathbf{x}_i), y_i\right) \\ &+ \frac{\beta}{|\mathcal{D}_k^{\mathrm{bd}}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k^{\mathrm{bd}}} \mathbb{E}_{m \sim \mathrm{Unif}(\{0,\ldots,M-1\})}\left[\ell\left(f_{\mathbf{w}}\left(\tau_{y_k,m}(\mathbf{x}_i)\right), y_k\right)\right].\end{aligned} \tag{25}$$

Here ℓ denotes the classification loss and β controls the relative weight of the backdoor objective.

In implementation, the expectation in Eq. 25 is approximated by sampling one mode for each poisoned example during local optimization. A deterministic schedule can instead assign

$$m_i = (i + e) \bmod M, \tag{26}$$

where i indexes the poisoned examples in the local training sequence and e indexes the local epoch.

4 Experiments

4.1 Experiment Setup

We implement COEX-FL and all baselines in PyTorch on four NVIDIA RTX 5090 GPUs. We evaluate ResNet-18 on CIFAR-10 Krizhevsky & Hinton (2009), GTSRB Stallkamp et al. (2011),

and Tiny-ImageNet Le & Yang (2015), comparing CBA Sun et al. (2019), DBA Xie et al. (2020), FCBA Liu et al. (2024c), Neurotoxin Zhang et al. (2022), PFL Fan et al. (2025), and Mirage Li et al. (2025b). Malicious clients are about 30% of sampled participants and the poisoned-data rate is 25%.

Unless specified otherwise, we use 100 clients, 10 participants per round, and 200 FedAvg rounds. three malicious clients are active continuously. COEX-FL uses $M = 4$ modes. ASR is averaged over modes on test samples excluding the target class. Avg-ASR, Std-ASR, and Min-ASR are computed across the three attackers. Cross-attacker plots report three independent seeds to quantify run-to-run variation.

4.2 Attack Efficacy

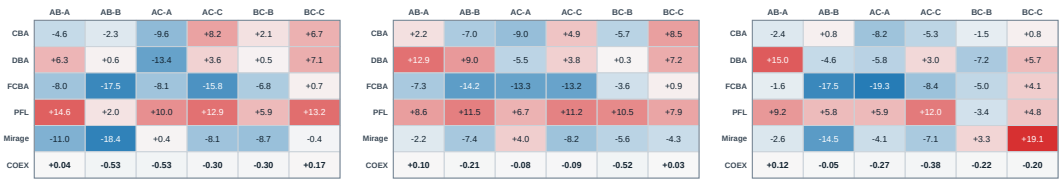
We compare COEX-FL with representative attacks under continuous and single-round injection.

Continuous Attack. Table 1 shows the highest Avg.ASR for COEX-FL on all datasets (99.80%, 99.44%, and 99.95%), with weakest-attacker ASR above 98.8% and Std.ASR below 0.4%. Baseline averages can hide severe imbalance. On Tiny-ImageNet, Mirage ranges from 100.00% to 4.05%. COEX-FL also has the highest Tiny-ImageNet ACC (65.47%).

Table 1: Attack effectiveness under continuous injection.

Dataset	Method	ACC	ASR_1	ASR_2	ASR_3	Avg.ASR	Std.ASR ↓
CIFAR-10	CBA	88.40	93.82	94.68	91.46	93.32	1.36
	DBA	89.03	88.03	77.53	80.59	82.05	4.41
	FCBA	88.96	87.26	78.43	85.37	83.69	3.79
	Neurotoxin	89.71	0.93	1.28	1.48	1.23	0.23
	PFL	**90.30**	57.82	46.92	27.04	43.93	12.74
	Mirage	89.91	98.03	91.59	97.04	95.55	2.83
	COEX-FL	88.65	**99.84**	**99.89**	**99.67**	**99.80**	**0.10**
GTSRB	CBA	99.11	90.92	95.05	97.38	94.45	2.67
	DBA	98.99	96.15	0.07	0.03	32.08	45.30
	FCBA	98.60	97.22	0.08	0.01	32.44	45.81
	Neurotoxin	99.04	2.22	1.23	0.75	1.40	0.61
	PFL	**99.13**	93.66	94.08	81.02	89.59	6.06
	Mirage	98.93	98.09	83.36	98.52	93.32	7.05
	COEX-FL	99.01	**99.69**	**98.89**	**99.75**	**99.44**	**0.39**
Tiny-ImageNet	CBA	55.17	63.22	35.33	5.13	34.56	23.72
	DBA	54.55	54.72	1.97	2.81	19.84	24.67
	FCBA	55.29	77.08	1.29	0.87	26.41	35.82
	Neurotoxin	51.98	22.60	58.21	22.28	34.36	16.86
	PFL	57.83	79.26	23.22	50.83	51.10	22.88
	Mirage	55.71	**100.00**	99.85	4.05	67.97	45.20
	COEX-FL	**65.47**	99.94	**99.92**	**99.99**	**99.95**	**0.03**

Single Attack. Table 2 shows strong, balanced performance after one injection. Avg.ASR is 100.00%, 99.95%, and 94.23% on CIFAR-10, GTSRB, and Tiny-ImageNet, with minimum ASR above 90%. Baselines often collapse on other targets. Semantic codes separate target components and modes provide alternative realizations. Tiny-ImageNet remains less balanced, likely reflecting class support and optimization effects rather than code overlap. Clean accuracy is lower on CIFAR-10 and GTSRB but highest on Tiny-ImageNet. Thus, the single-round experiment primarily tests whether the shared public construction can initialize several target behaviors before repeated aggregation, whereas the continuous protocol tests whether those behaviors remain balanced under reinforcement.



(a) Seed 1 (b) Seed 2 (c) Seed 3

Figure 1: Cross attacker interference on GTSRB after attack cessation.

Table 2: Attack effectiveness under single-round injection.

Dataset	Method	ACC	ASR_1	ASR_2	ASR_3	Avg.ASR	Std.ASR ↓
CIFAR-10	CBA	**89.25**	95.76	94.38	12.24	67.46	39.05
	DBA	82.45	43.34	3.29	16.64	21.09	16.65
	FCBA	70.77	57.11	5.17	59.90	40.73	25.17
	Neurotoxin	88.82	1.99	4.40	1.47	2.62	1.28
	PFL	80.03	5.98	32.29	13.87	17.38	11.02
	Mirage	64.50	89.66	0.52	18.38	36.19	38.51
	COEX-FL	83.38	**100.00**	**99.99**	**100.00**	**100.00**	**0.01**
GTSRB	CBA	93.99	51.83	45.51	2.98	33.44	21.69
	DBA	81.73	82.08	1.91	0.62	28.20	38.10
	FCBA	18.62	0.00	0.00	0.00	0.00	**0.00**
	Neurotoxin	**97.47**	0.08	0.05	0.05	0.06	0.01
	PFL	94.20	4.81	48.07	52.44	35.11	21.50
	Mirage	94.66	80.95	54.57	89.96	75.16	15.02
	COEX-FL	90.73	**99.97**	**100.00**	**99.88**	**99.95**	0.05
Tiny-ImageNet	CBA	55.82	3.19	1.10	93.78	32.69	43.21
	DBA	41.37	**99.99**	0.00	0.00	33.33	47.14
	FCBA	29.75	0.03	**98.39**	0.08	32.83	46.36
	Neurotoxin	58.13	0.33	0.28	0.36	0.32	**0.03**
	PFL	57.16	6.35	5.38	43.63	18.45	17.81
	Mirage	33.77	0.08	0.02	99.27	33.12	46.77
	COEX-FL	**65.35**	90.04	92.85	**99.80**	**94.23**	4.10

4.3 Cross-Attacker Interference

Let $P = \{a, b\}$ denote a pair of attackers. For target a, we define the signed cross-attacker interference as

$$\Delta\mathrm{ASR}_{P\to a} = \frac{1}{|\mathcal{R}_{\mathrm{post}}|} \sum_{t\in\mathcal{R}_{\mathrm{post}}} \left(\mathrm{ASR}^{(t)}_{P\to a} - \mathrm{ASR}^{(t)}_{\{a\}\to a}\right), \tag{27}$$

where $\mathcal{R}_{\mathrm{post}}$ denotes the rounds after malicious updates stop. All values are reported in percentage points. A negative value indicates suppression by the other attacker, whereas a positive value indicates positive transfer. For example, AB-A and AB-B denote evaluation of target A and target B when attackers A and B are jointly present, respectively.

Figure 1 reports the three random seeds separately. Across the seeds, the largest absolute interference of COEX-FL is 0.53, 0.52, and 0.38%, corresponding to a descriptive mean of 0.48% (95% t-interval of $[0.27, 0.69]$). The baselines show substantially larger and more variable effects, with the largest observed deviation reaching 19.3%. Thus, an isolated backdoor can be either attenuated or amplified when another attacker is present, whereas COEX-FL remains close to zero in this evaluated protocol. These results show that COEX-FL substantially reduces cross-attacker interference under the evaluated protocol, while the accompanying code, trigger, and update diagnostics clarify how this effect propagates across representation levels.

4.4 Defense Robustness

Table 3 highlights the trade-off between attack effectiveness and task utility. In the isolated setting, FLAME suppresses most attacks, while Neurotoxin remains effective on CIFAR-10 and TinyImageNet (62.25% and 25.78% ASR). Lockdown is bypassed by Mirage (88.30% and 58.68%), at clean accuracies of 29.26% and 37.70%. Clipping and DP preserve utility but constrain magnitude rather than malicious directions.

Under COEX-FL, every defense retains high ASR and roughly 90% CIFAR-10 accuracy. The lowest GTSRB and Tiny-ImageNet ASRs are 81.45% and 69.84%. Independently induced updates following the same public construction can remain non-outlying after aggregation, so isolated-attack robustness need not transfer. This does not imply attacker coordination. The result also highlights the scope of the defense experiment that it compares filtering behavior under coexistence, rather than claiming that any one defense is ineffective in all FL settings.

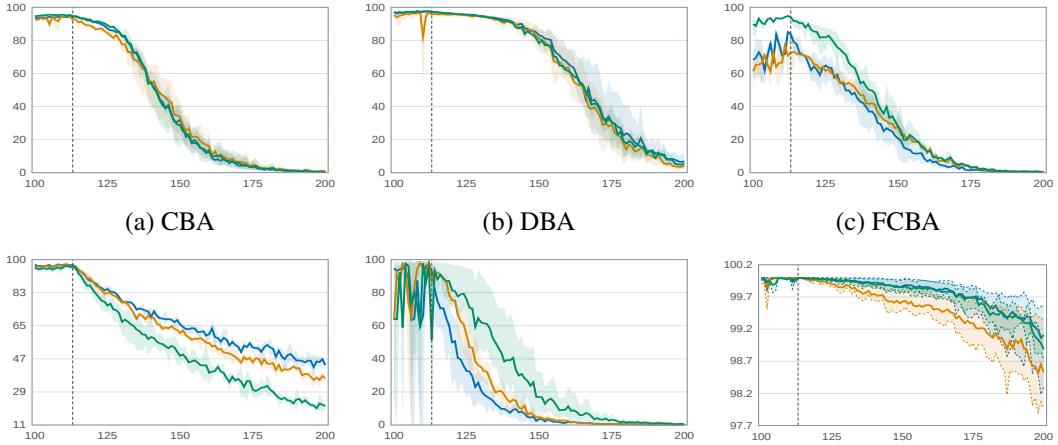
4.5 Forgetting and Cross-Attacker Interference

Pairwise ASR after malicious updates cease is below 1% for CBA, FCBA, and Mirage, 4.15–6.88% for DBA, and 21.46–43.60% for PFL, whereas COEX-FL maintains 98.54–99.12%. Figure 2

Table 3: Full defense performance under attacks.

Dataset	Defense	CBA		DBA		FCBA		Neurotoxin		PFL		Mirage		COEX-FL	
		Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
CIFAR-10	Norm Clipping	90.11	92.98	90.39	82.02	90.30	80.80	89.35	66.89	90.32	89.17	88.88	99.35	**90.50**	**99.76**
	DP	90.03	93.63	90.29	81.97	90.31	83.19	89.43	66.35	90.24	89.16	88.97	99.23	**90.39**	**99.71**
	Median	89.31	90.74	89.67	36.61	89.55	82.31	88.96	9.61	89.37	84.24	89.76	55.17	**90.40**	**99.91**
	Lockdown	88.91	3.20	89.65	2.10	89.59	2.37	43.37	0.50	89.21	33.77	29.26	88.30	**90.53**	**99.84**
	FLAME	**90.60**	1.38	90.46	1.30	90.39	1.55	89.46	62.25	90.43	22.09	90.41	2.34	90.08	**99.93**
GTSRB	Norm Clipping	**99.39**	94.51	99.36	32.28	99.35	32.48	95.04	0.96	99.20	98.64	99.18	**99.17**	98.98	96.70
	DP	**99.39**	94.92	99.36	32.31	99.36	32.60	94.76	0.89	99.17	98.56	99.31	**98.79**	98.55	98.69
	Median	99.26	64.42	**99.27**	31.91	99.23	32.13	98.81	5.68	99.13	**96.59**	99.12	35.29	98.94	89.58
	Lockdown	96.25	0.17	96.26	0.17	96.37	0.17	85.28	0.06	96.06	42.47	89.73	5.60	**98.92**	**98.42**
	FLAME	99.29	0.05	**99.38**	0.03	99.36	0.06	98.98	0.21	99.25	27.58	99.29	0.06	98.01	**81.45**
TinyImageNet	Norm Clipping	58.37	33.96	59.29	18.16	60.15	26.82	59.47	25.38	58.26	68.88	56.98	54.99	57.49	**69.84**
	DP	58.30	33.98	59.57	19.22	60.08	27.62	59.42	25.00	58.14	76.76	56.78	**94.82**	**60.91**	93.67
	Median	55.05	30.17	54.51	13.10	**55.37**	23.28	51.49	22.02	53.31	80.25	54.56	39.41	54.92	**99.94**
	Lockdown	51.21	6.70	**58.04**	0.08	57.93	0.35	52.41	0.85	55.43	18.98	37.70	58.68	58.02	**83.31**
	FLAME	58.69	0.27	57.92	0.24	58.30	0.29	56.45	25.78	58.11	0.61	58.20	0.36	**63.83**	**99.82**

shows the complete GTSRB curves, including the three seeds used for the coexistence comparison. The COEX-FL gap across pairs is only 0.58 percentage points, with seed outcomes ranging from 97.98 to 99.57%. The baseline curves also illustrate why average attack success alone is insufficient, one target may retain a moderate response while another has already forgotten its mapping.



(a) CBA (b) DBA (c) FCBA

(d) PFL (e) Mirage (f) COEX-FL

Figure 2: Pairwise forgetting curves on GTSRB after attack cessation.

4.6 Ablation Study

Table 4: Ablation study on CIFAR-10.

Variant	Acc.	ASR_1	ASR_2	ASR_3	Avg. ASR	Min. ASR	Inter-Atk. Std.
Full	89.93	94.98	**91.65**	99.50	**95.38**	**91.65**	**3.22**
w/o Semantic Code	**90.06**	18.20	1.39	94.19	37.93	1.39	40.38
w/o Mode Diversity	89.55	**96.37**	0.50	98.16	65.01	0.50	45.62
w/o Shift/Permutation	89.99	82.03	8.04	99.49	63.19	8.04	39.64
w/o Carrier Separation	89.88	93.09	15.89	**99.66**	69.55	15.89	38.04

In the CIFAR-10 ablation, Full achieves 95.38% average ASR, 91.65% minimum ASR, and 89.93% clean accuracy. The shared-semantic-code variant removes target conditioning, while the other variants isolate mode diversity, mode shift/permutation, and target-indexed carrier separation. Full has

the smallest attacker-wise spread. Removing semantic codes causes the largest drop in average ASR, whereas removing mode diversity reduces ASR_2 to 0.50%, showing that redundant realizations protect against fragile carrier paths. Shift/permutation and carrier separation also reduce destructive collisions, with 3.22% inter-attacker spread for Full compared with at least 38.04% for every ablation. The ablations can preserve one strong target response but fail to balance all attackers. These results support complementary component roles but do not establish a causal code-to-update relationship, so code, trigger, and update similarities are analyzed separately under matched budgets.

The semantic-code ablation keeps the carrier and optimization budget fixed while removing target conditioning. Its low minimum ASR shows that a shared carrier alone cannot reliably preserve distinct target behaviors. Mode diversity addresses a separate failure mode because a single realization may be weakened by aggregation or benign updates. The shift/permutation and carrier-separation variants further indicate that support placement affects learned-behavior balance, although the table cannot determine whether the change occurs in input, feature, or parameter space. We therefore interpret the ablation as evidence of complementary component roles rather than an additive causal decomposition.

Across experiments, the clearest gain appears in the weakest-attacker metric. Average ASR may remain high when one target dominates and another is forgotten, whereas Min-ASR and attacker-wise standard deviation reveal whether the behaviors coexist reliably. The continuous-attack, single-round, and forgetting experiments examine reinforcement, initialization, and retention, respectively. Their agreement suggests that COEX-FL redistributes representation capacity across target behaviors rather than merely strengthening one backdoor. The defense results further indicate that this compatibility persists after filtering, clipping, and noise mechanisms under the evaluated protocols.

5 Conclusion

We study multi-backdoor coexistence without runtime exchange of attack-specific information. COEX-FL factorizes triggers into target-conditioned and physical-realization components. Hadamard codes separate target identities in code space, while frequency modes provide multiple realizations. Complete triggers and learned updates need not be orthogonal. Across continuous, single-round, defense, forgetting, and ablation protocols, the method maintains a higher weakest-attacker ASR and smaller attacker-wise spread than the considered baselines.The joint-code Gram and DCT identities are exact for the construction. Low update interference depends on a local update-map approximation and is evaluated empirically rather than guaranteed under nonlinear training. The results therefore support a public compatibility rule for independently constructed backdoors, while leaving the transfer from code geometry to learned representations dependent on the data, optimizer, client schedule, and defense.

AI use statement

In this work, we used generative AI tools to assist with code modifications, polish the manuscript, perform a comprehensive grammar check, and check the correctness and consistency of mathematical formulas. We manually reviewed all AI-assisted work. In particular, we validated and tested the AI-assisted code to ensure that its implementation and behavior are consistent with the methods described in the paper. We also manually checked all AI-assisted language polishing and grammar corrections, as well as the mathematical formulas. We take full responsibility for the final content of this work, including all text, claims, mathematical expressions, and code.

References

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2938–2948, 2020.

Nahom Birhan, Daniel Wesego, Dereje Shenkut, Frank Liu, and Daniel Takabi. DCInject: Persistent backdoor attacks via frequency manipulation in personal federated learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1866–1870, 2026.

Wu-Chun Chung, Chao-Ai Lo, Yan-Hui Lin, Zhi-Hao Chen, and Che-Lun Hung. Decentralized federated learning with non-iid data: Challenges, trends, and future opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

Yanbo Dai and Songze Li. Chameleon: Adapting to peer images for planting durable backdoors in federated learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6712–6725, 2023.

Chamitha De Alwis, Ons Aouedi, Jiaming Xu, Shen Wang, Yushan Siriwardhana, Tharaka Hewa, Engin Zeydan, Chamara Sandeepa, and Madhusanka Liyanage. Federated learning for 6g security: A survey on threats, solutions and research directions. *IEEE Communications Surveys & Tutorials*, 2026.

Khoa D. Doan, Yingjie Lao, and Ping Li. Marksman backdoor: Backdoor attacks with arbitrary target class. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38260–38273, 2022.

Mingyuan Fan, Zhanyi Hu, Fuyi Wang, and Cen Chen. Bad-pfl: Exploring backdoor attacks against personalized federated learning. *International Conference on Learning Representations (ICLR)*, 2025.

Jun Feng, Yuzhe Lai, Hong Sun, and Bocheng Ren. Sadba: Self-adaptive distributed backdoor attack against federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16568–16576, 2025.

Lingguag Hao, Kuangrong Hao, Bing Wei, and Xue-Song Tang. Multi-target federated backdoor attack based on feature aggregation. *Pattern Recognition*, 172(Part A):112333, 2026.

Linshan Hou, Zhongyun Hua, Yuhong Li, Yifeng Zheng, and Leo Yu Zhang. M-to-n backdoor paradigm: A multi-trigger and multi-target attack to deep learning models. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):11299–11312, 2024.

Gou Koutaki, Keiichiro Shirai, and Mitsuru Ambai. Hadamard coding for supervised discrete hashing. *IEEE Transactions on Image Processing*, 27(11):5378–5392, 2018.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009.

Ya Le and Xuan Yang. Tiny imagenet cisual recognition challenge. *CS 231N*, 7(7), 2015.

Seohyun Lee, Wenzhi Fang, Anindya Bijoy Das, Seyyedali Hosseinalipour, David J. Love, and Christopher G. Brinton. Cooperative decentralized backdoor attacks on vertical federated learning. *IEEE Transactions on Networking*, 34:2004–2019, 2026.

Chunhai Li, Yunhui Shen, Ming Xie, Zhaoyu Su, Yajie Wang, Chuan Zhang, and Liehuang Zhu. Fed-cbe: Client-side backdoor elimination in federated learning via persistent parameter disruption. *IEEE Transactions on Information Forensics and Security*, 2026.

Guorui Li, Runxing Chang, Ying Wang, and Cong Wang. Dual-domain based backdoor attack against federated learning. *Neurocomputing*, 623:129424, 2025a.

Xuan Li, Longfei Wu, Zhitao Guan, Xiaojiang Du, Nadjib Aitsaadi, and Mohsen Guizani. Muldoor: A multi-target backdoor attack against federated learning system. In *IEEE Global Communications Conference*, pp. 1749–1754, 2024.

Ye Li, Yanchao Zhao, Chengcheng Zhu, and Jiale Zhang. Infighting in the dark: Multi-label backdoor attack in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25770–25779, 2025b.

Jiawang Liu, Changgen Peng, Weijie Tan, and Chenghui Shi. Federated learning backdoor attack based on frequency domain injection. *Entropy*, 26(2):164, 2024a.

Tao Liu, Wu Yang, Chen Xu, Jiguang Lv, Huanran Wang, Yuhang Zhang, Shuchun Xu, and Dapeng Man. Act in collusion: A persistent distributed multi-target backdoor in federated learning. *arXiv preprint arXiv:2411.03926*, 2024b.

Tao Liu, Yuhang Zhang, Zhu Feng, Zhiqin Yang, Chen Xu, Dapeng Man, and Wu Yang. Beyond traditional threats: A persistent backdoor attack on federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21359–21367, 2024c.

Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, and Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):9020–9028, 2023.

Zhixuan Ma, Haichang Gao, Shangwen Li, Ping Wang, and Han Yu. Coupled trigger optimization and vulnerable parameter alignment for persistent backdoor attacks on federated learning. In *Forty-third International Conference on Machine Learning*, 2026.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

Nuno M. Neves. Mingling with the good to backdoor federated learning. *arXiv preprint arXiv:2501.01913*, 2025.

Thuy Dung Nguyen, Tuan Nguyen, Anh Tran, Khoa D. Doan, and Kok-Seng Wong. IBA: Towards irreversible backdoor attacks in federated learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66364–66376, 2023.

Tuan Anh Nguyen and Anh Tuan Tran. Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3454–3464, 2020.

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

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Xiangyun Tang, Wen Yang, Luyao Peng, Meng Shen, Tao Zhang, Yu Weng, Jiawen Kang, and Dusit Niyato. Finback: Infiltrating backdoors into gradient compressors on federated learning. *IEEE Transactions on Information Forensics and Security*, 20:12460–12475, 2025.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jyyong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16070–16084, 2020.

Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In *Computer Vision – ECCV 2022*, volume 13664 of *Lecture Notes in Computer Science*, pp. 396–413. Springer, 2022.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

Jiahao Xu, Zikai Zhang, and Rui Hu. Detecting backdoor attacks in federated learning via direction alignment inspection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20654–20664. IEEE, 2025.

Pei Ye, Yuqing Li, Kun He, Haoran Wang, Ruiying Du, and Wei Wang. Less is more: Persistent low-frequency backdoor injection in federated learning. In *IEEE INFOCOM 2026-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2026.

Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16453–16461, 2021.

Feng Zhang, Yixiao Xu, Jun Lv, Zexin Yu, Luo Zhang, Jun Feng, and Xiaolei Liu. Backdoorvaccine: An active backdoor defense framework via competitive backdoor injection. *Neurocomputing*, 664:132133, 2026.

Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3FL: Adversarially adaptive backdoor attacks to federated learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26429–26446, 2022.

Haomin Zhuang, Mingxian Yu, Hao Wang, Yang Hua, Jian Li, and Xu Yuan. Backdoor federated learning by poisoning backdoor-critical layers. In *International Conference on Learning Representations*, 2024.

A Appendix

A.1 Matched-Budget Semantic-Code Intervention

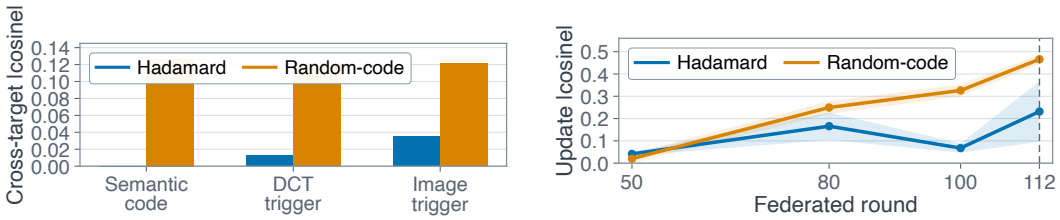
We perform a matched-budget intervention experiment on GTSRB to isolate the effect of semantic-code construction. The Full configuration uses Hadamard semantic codes, while the Random-code configuration replaces them with random dense unit-norm codes. All other factors are kept fixed.

Figure 3a and 3b examines whether semantic-code orthogonality is preserved as the representation moves from code space to trigger space and then to update space. Figure 3a reports the mean absolute cosine similarity over 48 cross-target pairs. The Hadamard codes are exactly orthogonal and therefore have zero cross-target similarity, whereas the random dense codes have a similarity of 0.1267. This separation is also visible in trigger space. The mean DCT-trigger similarity is 0.0125 for Full and 0.1209 for Random-code. The corresponding image-space similarities are 0.0350 and 0.1214. The increase from DCT space to image space is expected because the image transformation introduces additional shared components. The lower similarity under Full is consistent with separation surviving the realized-trigger construction in this intervention, but it is not implied by code-space orthogonality and is not a guarantee for other carrier or clipping settings.

Figure 3b measures the mean absolute pairwise cosine similarity between the three logged attacker parameter deltas at selected training rounds. These deltas include any shared clean-training component. They are therefore descriptive update-alignment diagnostics rather than attack-residual similarities. At rounds 80, 100, and 112, the Full configuration obtains similarities of 0.1658, 0.0677, and 0.2317, compared with 0.2496, 0.3261, and 0.4661 for Random-code. Full therefore produces more separated learned updates at three of the four checkpoints. At round 50, the similarity is slightly higher for Full than for Random-code, with values of 0.0416 and 0.0203. This early-stage difference suggests that update separation emerges after the initial training transient rather than being uniformly present from the first measured checkpoint.

Figure 4a and 4b summarizes the post-attack persistence of the learned representations. Figure 4a shows that the average retention ratio is 0.837 for Full and 0.793 for Random-code. The minimum retention ratio is 0.824 for Full and 0.772 for Random-code. Figure 4b provides the corresponding ASR values at the attack cessation round and at the final round. These results show that the Hadamard configuration preserves attack-specific behavior more effectively after malicious updates are no longer injected. The improvement co-occurs with lower late-stage logged update similarity in Figure 3b.

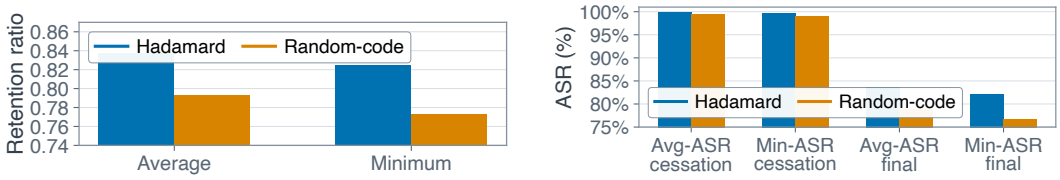
These results show exact separation in the semantic code block and a corresponding difference in the measured trigger and update diagnostics for this matched GTSRB intervention. They do not imply that every inverse-transformed trigger or every nonlinear training trajectory preserves orthogonality.



(a) Semantic and trigger geometry.

(b) Learned update similarity.

Figure 3: Similarity across code, trigger, and update spaces.



(a) Post-attack retention.

(b) ASR at selected rounds.

Figure 4: Retention of attack-specific behavior after the attack is disabled.

We therefore use the intervention as evidence that code construction can affect learned behavior under a fixed protocol, while leaving the transfer to update space conditional on the observed training map.

A.1.1 Effectiveness under different client selection strategies

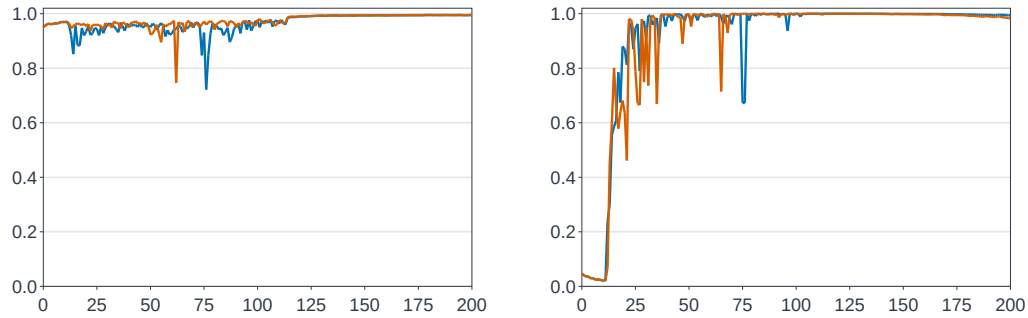
Beyond the setting in which all three malicious clients participate simultaneously, we evaluate two alternative client participation strategies shown in Figure 5. The first activates one malicious client at a time in a cyclic order, whereas the second randomly selects two malicious clients at each round. We record the average attack success rate to compare the effectiveness of these strategies. The two attacker selection strategies both achieve a rapid increase in attack success rate from the initial level of approximately 4.7% to nearly 100% . Their convergence speed is comparable, although the cyclic single attacker schedule reaches the high success rate slightly earlier and maintains stronger long term attack effectiveness. At the end of evaluation, the average ASR is 99.39% for the cyclic schedule and 98.32% for the random two attacker schedule.

The random two attacker schedule causes less degradation in clean accuracy during the attack period. Its mean clean accuracy is 95.83%, compared with 94.39% for the cyclic schedule, while both strategies recover to approximately 99.48% after the attack. Overall, the cyclic schedule provides slightly stronger and more persistent attacks, whereas the random two attacker schedule offers better clean model utility.

A.1.2 Scalability with the Number of Malicious Clients.

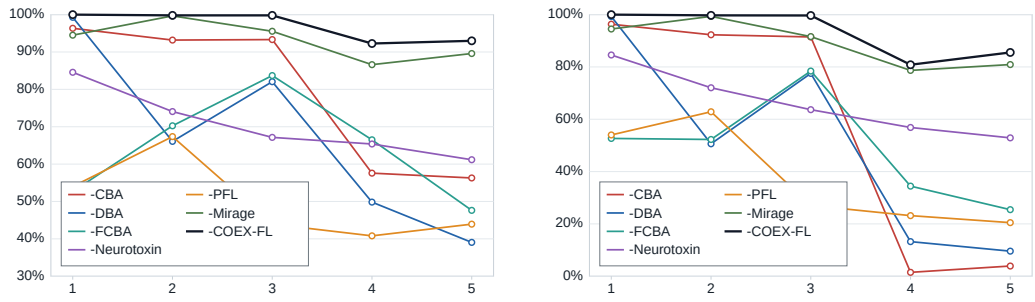
Figures 6a and 6b show the attack performance as the number of malicious clients increases from one to five. COEX-FL maintains near-saturated average ASR across all settings, decreasing only from 100.00% to 92.99%. Its minimum ASR also remains high, reaching 80.84% and 85.54% with four and five attackers, respectively. In contrast, DBA and CBA suffer substantial degradation in the high-attacker regime, with minimum ASRs falling to 9.58% and 3.89%, while Mirage remains the strongest competing baseline but still underperforms COEX-FL in worst-attacker ASR.

The results are consistent with the hypothesis that target-conditioned codes, mode diversity, and carrier transformations can reduce destructive interference under the evaluated client schedules. The robustness is reflected primarily in minimum ASR, not only average attack strength. Clean accuracy remains stable across configurations, but we describe the effect as compatibility induced by a shared public rule rather than multi-attacker coordination.



(a) Clean accuracy (b) Average ASR

Figure 5: Comparison of malicious client participation strategies.



(a) Average ASR (b) Min ASR

Figure 6: The ASR under different amount malicious clients.

A.2 Theory and Scope

This appendix gives the formal statements that support the semantic-mode construction in the main text. The results are deliberately separated into three levels: exact code-space identities, exact DCT-coefficient identities after physical placement, and conditional consequences for linearized FL updates. All claims below concern the construction in equation 8 and the perturbations in equation 17-equation 18, they do not assert that nonlinear training must reach a particular ASR.

A.2.1 Code-space geometry

Let δ_{ab} denote the Kronecker delta. From the Walsh-Hadamard construction,

$$\mathbf{c}_y^\top \mathbf{c}_{y'} = \delta_{yy'}, \qquad \mathbf{r}_m^\top \mathbf{r}_{m'} = \delta_{mm'}. \tag{28}$$

The joint code used by COEX-FL is the block concatenation

$$\mathbf{z}_{y,m} = \begin{bmatrix} \sqrt{\rho}\,\mathbf{c}_y \\ \sqrt{1-\rho}\,\mathbf{r}_m \end{bmatrix}, \qquad 0 < \rho < 1. \tag{29}$$

Proposition 1 (Exact Gram matrix of the joint code)**.** *For any valid target-mode pairs* (y, m) *and* (y', m')*,*

$$\|\mathbf{z}_{y,m}\|_2 = 1, \qquad \langle \mathbf{z}_{y,m}, \mathbf{z}_{y',m'} \rangle = \rho\,\delta_{yy'} + (1-\rho)\,\delta_{mm'}. \tag{30}$$

Consequently, when at least two targets and two modes are used, the worst-case pairwise coherence among distinct joint codes is

$$\mu_{\text{code}} = \max_{(y,m)\neq(y',m')} |\langle \mathbf{z}_{y,m}, \mathbf{z}_{y',m'} \rangle| = \max\{\rho, 1-\rho\} \geq \frac{1}{2}. \tag{31}$$

Proof. Equation 30 follows by expanding the inner product over the semantic and mode blocks and substituting the two orthogonality identities. For distinct pairs, the largest possible overlap is ρ when $y = y'$ and $m \neq m'$, and $1 - \rho$ when $y \neq y'$ and $m = m'$. Taking the larger of the two terms gives equation 31. □

Thus, the concatenated code should not be interpreted as making all target-mode pairs mutually orthogonal. Its exact role is more limited and more specific: it makes the semantic block orthogonal across targets, makes the mode block orthogonal across modes, and exposes the residual overlap through the single parameter ρ.

A.2.2 Exact geometry after DCT carrier placement

Let $\mathcal{I}$ denote the ambient index set of the chrominance DCT coefficients used for perturbation. Let $E_s \in \{0,1\}^{|\mathcal{I}| \times L_s}$ be the semantic placement operator induced by $\mathcal{B}_s = (b_0, \ldots, b_{L_s - 1})$. For each mode m, let $E_m \in \{0,1\}^{|\mathcal{I}| \times L_m}$ be the mode placement operator induced by π_m. Its j-th column selects the coefficient $b_{L_s + \pi_m(j)}$. Since the semantic carrier and mode pool are disjoint,

$$E_s^\top E_s = I_{L_s}, \qquad E_m^\top E_m = I_{L_m}, \qquad E_s^\top E_m = 0. \tag{32}$$

The unscaled DCT perturbation can therefore be written as

$$\boldsymbol{\Delta}_{y,m} = \sqrt{\rho}\, E_s \mathbf{c}_y + \sqrt{1 - \rho}\, E_m \mathbf{r}_m. \tag{33}$$

The actual perturbation added in equation 19 is $\lambda \boldsymbol{\Delta}_{y,m}$.

For two modes m, m', define the signed physical mode overlap

$$\omega_{m,m'} = \mathbf{r}_m^\top E_m^\top E_{m'} \mathbf{r}_{m'}. \tag{34}$$

Equivalently,

$$\omega_{m,m'} = \sum_{j=0}^{L_m - 1} \sum_{j'=0}^{L_m - 1} r_m[j]\, r_{m'}[j']\, \mathbf{1}\{\pi_m(j) = \pi_{m'}(j')\}. \tag{35}$$

If $q_{m,m'}$ is the number of shared physical mode coefficients between modes m and m', then

$$|\omega_{m,m'}| \leq \frac{q_{m,m'}}{L_m}. \tag{36}$$

Proposition 2 (Exact DCT-coefficient Gram matrix)**.** *For all valid* (y, m) *and* (y', m')*,*

$$\|\boldsymbol{\Delta}_{y,m}\|_2 = 1 \tag{37}$$

and

$$\langle \boldsymbol{\Delta}_{y,m}, \boldsymbol{\Delta}_{y',m'} \rangle = \rho\, \delta_{yy'} + (1 - \rho)\, \omega_{m,m'}. \tag{38}$$

Consequently,

$$\frac{\langle \lambda \boldsymbol{\Delta}_{y,m}, \lambda \boldsymbol{\Delta}_{y',m'} \rangle}{\lambda^2} = \rho\, \delta_{yy'} + (1 - \rho)\, \omega_{m,m'}. \tag{39}$$

Proof. Substitute equation 33 and use $E_s^\top E_m = 0$. The semantic part gives $\rho\, \mathbf{c}_y^\top \mathbf{c}_{y'} = \rho\, \delta_{yy'}$, while the mode part gives $(1 - \rho) \mathbf{r}_m^\top E_m^\top E_{m'} \mathbf{r}_{m'}$. Setting $(y, m) = (y', m')$ gives unit norm because $\omega_{m,m} = 1$. □

The exact physical coherence of the implemented perturbation family is therefore

$$\mu_{\mathrm{DCT}} = \max_{(y,m) \neq (y',m')} \left| \rho\, \delta_{yy'} + (1 - \rho)\, \omega_{m,m'} \right|. \tag{40}$$

This is the relevant deterministic diagnostic after carrier placement. If different modes use disjoint physical supports, then $\omega_{m,m'} = 0$ for $m \neq m'$. Under the cyclic placement, however, off-diagonal overlaps may be nonzero and must be measured or bounded by equation 36 rather than assumed away.

All exact DCT-space identities apply to the unclipped coefficient perturbations $\lambda \Delta_{y,m}$. The inverse transform and clipping operation in equation 19 may alter the corresponding image-space norms and inner products. These identities therefore establish coefficient-space geometry only: they do not imply orthogonality of the image-space triggers or of learned parameter updates.

A.2.3 MODE-AVERAGED TARGET SEPARATION

The local objective in equation 25 trains over an expectation with respect to the mode. The worst-case pairwise coherence in equation 31 is therefore not the only relevant quantity. A separate first-moment quantity describes the representation sampled by the multi-mode objective.

Define the mode-averaged code and DCT perturbation for target y as

$$\bar{\mathbf{z}}_y = \frac{1}{M}\sum_{m=0}^{M-1}\mathbf{z}_{y,m}, \qquad \bar{\mathbf{\Delta}}_y = \frac{1}{M}\sum_{m=0}^{M-1}\mathbf{\Delta}_{y,m}. \tag{41}$$

Proposition 3 (Mode-averaged code separation)**.** *For two distinct targets $y \neq y'$,*

$$\langle \bar{\mathbf{z}}_y, \bar{\mathbf{z}}_{y'} \rangle = \frac{1-\rho}{M}, \qquad \|\bar{\mathbf{z}}_y\|_2^2 = \rho + \frac{1-\rho}{M}. \tag{42}$$

Hence the normalized mode-averaged cross-target coherence in code space is

$$\nu_{\text{code}} = \frac{|\langle \bar{\mathbf{z}}_y, \bar{\mathbf{z}}_{y'} \rangle|}{\|\bar{\mathbf{z}}_y\|_2 \, \|\bar{\mathbf{z}}_{y'}\|_2} = \frac{1-\rho}{M\rho + 1 - \rho}. \tag{43}$$

Proof. Let $\bar{\mathbf{r}} = M^{-1}\sum_{m=0}^{M-1}\mathbf{r}_m$. Since the selected Hadamard rows are orthonormal, $\|\bar{\mathbf{r}}\|_2^2 = M^{-2}\sum_{m,m'}\delta_{mm'} = 1/M$. For $y \neq y'$, the semantic inner product vanishes and the shared averaged mode component contributes $(1-\rho)\|\bar{\mathbf{r}}\|_2^2 = (1-\rho)/M$. □

For the physically placed perturbations, let

$$\bar{\mathbf{v}} = \frac{1}{M}\sum_{m=0}^{M-1} E_m \mathbf{r}_m, \qquad \bar{\omega} = \|\bar{\mathbf{v}}\|_2^2 = \frac{1}{M^2}\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1}\omega_{m,m'}. \tag{44}$$

Then, for any $y \neq y'$,

$$\left\langle \bar{\mathbf{\Delta}}_y, \bar{\mathbf{\Delta}}_{y'} \right\rangle = (1-\rho)\bar{\omega}, \qquad \|\bar{\mathbf{\Delta}}_y\|_2^2 = \rho + (1-\rho)\bar{\omega}, \tag{45}$$

and the normalized physical mode-averaged cross-target coherence is

$$\nu_{\text{DCT}} = \frac{(1-\rho)\bar{\omega}}{\rho + (1-\rho)\bar{\omega}}. \tag{46}$$

Thus, mode diversity contributes to target separation only to the extent that the averaged physical mode vector $\bar{\mathbf{v}}$ is small. In the ideal orthogonal-placement case, $\bar{\omega} = 1/M$ and equation 46 reduces to equation 43. With cyclic placement, the exact value is given by equation 44.

A.2.4 CONDITIONAL LINEARIZED UPDATE INTERFERENCE

The preceding statements are exact geometry in code or DCT coefficient space. To connect them to FL updates, an additional local approximation is required. Let $\mathbf{u}_a$ denote a unit-norm trigger representation for attacker or target index a, depending on the analysis, $\mathbf{u}_a$ may be an individual perturbation $\mathbf{\Delta}_{y,m}$ or a normalized mode-averaged perturbation $\bar{\mathbf{\Delta}}_y/\|\bar{\mathbf{\Delta}}_y\|_2$. Assume that, at round t, the trigger-induced part of the local malicious update admits the linearization

$$\mathbf{g}_a^t = B_t \mathbf{u}_a. \tag{47}$$

Assume further that, on the span of the considered trigger representations,

$$\left\| B_t^\top B_t - \kappa_t I \right\|_2 \leq \varepsilon_t, \qquad \kappa_t > 0, \qquad 0 \leq \varepsilon_t < \kappa_t. \tag{48}$$

This is a local near-isometry condition on the update map, without such a condition, an arbitrary nonlinear training process can rotate or amplify even orthogonal coefficient-space directions.

Proposition 4 (Linearized cross-update interference)**.** *Suppose $|\mathbf{u}_a^\top \mathbf{u}_b| \leq \nu$ for $a \neq b$ and equation 48 holds. Then*

$$\|\mathbf{g}_a^t\|_2^2 \in [\kappa_t - \varepsilon_t, \kappa_t + \varepsilon_t], \tag{49}$$

and for $a \neq b$,

$$\left|(\mathbf{g}_a^t)^\top \mathbf{g}_b^t\right| \leq \kappa_t \nu + \varepsilon_t. \tag{50}$$

Consequently,

$$\left| \frac{(\mathbf{g}_a^t)^\top \mathbf{g}_b^t}{\|\mathbf{g}_a^t\|_2 \|\mathbf{g}_b^t\|_2} \right| \leq \frac{\kappa_t \nu + \varepsilon_t}{\kappa_t - \varepsilon_t}. \tag{51}$$

Proof. Write

$$(\mathbf{g}_a^t)^\top \mathbf{g}_b^t = \mathbf{u}_a^\top B_t^\top B_t \mathbf{u}_b = \kappa_t \mathbf{u}_a^\top \mathbf{u}_b + \mathbf{u}_a^\top (B_t^\top B_t - \kappa_t I)\mathbf{u}_b.$$

The Cauchy-Schwarz inequality and equation 48 give equation 50. The norm bound follows by setting $a = b$, and equation 51 follows from the lower norm bound. □

For the FedAvg aggregation at round t, let $\mathcal{A}_t$ be the selected malicious clients and fix one malicious direction $a \in \mathcal{A}_t$. The relative projection of other malicious linearized updates onto $\mathbf{g}_a^t$ is bounded by

$$\frac{\left|\left\langle \mathbf{g}_a^t, \sum_{b\in\mathcal{A}_t\setminus\{a\}} \alpha_b \mathbf{g}_b^t \right\rangle\right|}{\|\mathbf{g}_a^t\|_2^2} \le \sum_{b\in\mathcal{A}_t\setminus\{a\}} \alpha_b \frac{\kappa_t \nu + \varepsilon_t}{\kappa_t - \varepsilon_t}. \tag{52}$$

Here ν should be chosen according to the representation being analyzed: $\nu = \mu_{\mathrm{DCT}}$ for worst-case individual perturbations, or $\nu = \nu_{\mathrm{DCT}}$ for the mode-averaged linearized objective. Benign client updates, clipping, differential privacy noise, robust aggregation, and optimizer dynamics add further terms that are not controlled by the code geometry alone.

A.2.5 A SUFFICIENT CERTIFICATE FOR POST-ATTACK PERSISTENCE

The persistence metric in the main text evaluates whether a learned backdoor survives after malicious updates stop. The following certificate gives a sufficient condition for such survival on a fixed evaluated trigger set.

For target y, define the logit margin

$$M_y(\mathbf{w}, \mathbf{x}) = f_{\mathbf{w}}(\mathbf{x})_y - \max_{j\neq y} f_{\mathbf{w}}(\mathbf{x})_j. \tag{53}$$

Assume that every logit $f_{\mathbf{w}}(\mathbf{x})_j$ is L-Lipschitz in $\mathbf{w}$ along the post-attack trajectory. Then, for any $\mathbf{w}, \mathbf{w}'$ on that trajectory,

$$|M_y(\mathbf{w}, \mathbf{x}) - M_y(\mathbf{w}', \mathbf{x})| \le 2L\|\mathbf{w} - \mathbf{w}'\|_2. \tag{54}$$

Let T be the last attack round, and let $\mathcal{E}$ be the finite set of triggered examples used for persistence evaluation, including their target labels. Suppose

$$m_* = \min_{(\mathbf{x},y)\in\mathcal{E}} M_y(\mathbf{w}^T, \mathbf{x}) > 0. \tag{55}$$

If for some later round T',

$$\sum_{s=T}^{T'-1} \|\mathbf{w}^{s+1} - \mathbf{w}^s\|_2 < \frac{m_*}{2L}, \tag{56}$$

then every example in $\mathcal{E}$ keeps its target prediction through round T'. If the same inequality holds with the infinite sum $\sum_{s=T}^{\infty}$, then the target predictions persist for the entire post-attack trajectory.