technologies

MDPI

Article

CAPG: Context-Aware Perturbation Generation for Multi-Label Adversarial Attacks

Aidos Askhatuly [1], Dinara Berdysheva [1], Azamat Berdyshev [2], Aigul Adamova [3,*] and Didar Yedilkhan [4,*]

[1] Faculty of Information Technology, Al-Farabi Kazakh National University, 050000 Almaty, Kazakhstan; aidos.askhatuly@gmail.com (A.A.); d.berdysheva@gmail.com (D.B.)
[2] Faculty of Information Technology, International IT University, 050000 Almaty, Kazakhstan; aberdysh@gmail.com
[3] School of Software Engineering, Astana IT University, 010000 Astana, Kazakhstan
[4] Smart City Research Center, Astana IT University, 010000 Astana, Kazakhstan
* Correspondence: aigul.adamova@astanait.edu.kz (A.A.); d.yedilkhan@astanait.edu.kz (D.Y.)

Abstract

Multi-label deep learning models are widely used in real-world applications where predictions depend on the joint presence of several semantically correlated labels. However, existing adversarial attacks largely overlook these inter-label dependencies, often perturbing outputs indiscriminately and producing structurally implausible or easily detectable changes. This paper presents CAPG (Context-Aware Perturbation Generation), a white-box, label-space targeted adversarial framework for generating selective and contextually consistent perturbations in multi-label settings. CAPG incorporates correlation-weighted regularization into the adversarial objective, enabling targeted manipulation of specific labels while preserving the contextual integrity of non-target outputs. Using the Pascal VOC 2012 dataset and a ResNet-101 multi-label classifier, we show that CAPG achieves higher Attack Success Rates (ASR) and substantially improved Contextual Consistency Scores (CCSs) than FGSM, PGD, CW, and DeepFool under identical perturbation budgets. CAPG also produces lower perceptual distortion, yielding adversarial examples that better preserve contextual structure. These results highlight the importance of correlation-aware adversarial evaluation for assessing the robustness of modern multi-label deep learning systems.

Keywords: adversarial attacks; multi-label learning; context-aware perturbations; correlation modeling

Check for updates

Academic Editor: Pedro Antonio Gutiérrez

Received: 2 March 2026
Revised: 9 April 2026
Accepted: 13 April 2026
Published: 16 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Multi-label deep learning models have become fundamental components in numerous real-world applications, including medical image analysis, autonomous perception, remote sensing, and general-purpose scene understanding [1,2]. In these settings, each input typically contains multiple objects or attributes that co-occur in semantically meaningful ways, forming an inherent structure of dependencies across labels. These dependencies include spatial co-occurrence (person–bicycle), functional relationships (car–road), and hierarchical semantics (animal–dog), all of which contribute to the model's ability to generate coherent multi-label predictions [3]. As a result, perturbations that arbitrarily disrupt label outputs risk violating these semantic constraints, leading to degraded realism and easily detectable inconsistencies.

Although adversarial robustness has been extensively studied for single-label classifiers, current attack methods such as FGSM [4], PGD [5], CW [6], and DeepFool [7] are not designed to handle the structured nature of multi-label prediction. When applied directly in multi-label settings, these attacks tend to manipulate all labels indiscriminately, often suppressing or amplifying outputs in ways that contradict natural co-occurrence patterns. Such behavior fails to reflect realistic adversarial objectives, where robustness evaluation requires the ability to selectively alter specific labels [8]. For example, a plausible adversarial goal is to suppress the label "car" while keeping correlated contextual labels such as "road" or "vehicle" unchanged. The absence of contextual awareness in existing attacks therefore limits their usefulness for evaluating the robustness of multi-label systems, since indiscriminate perturbations do not reflect realistic, semantically constrained adversarial behavior [9,10].

Recent work has further highlighted two complementary challenges in multi-label learning. On one hand, robustness under adversarial conditions remains underexplored compared with single-label settings, even though multi-label classifiers are used in important vision applications [11]. On the other hand, recent multi-label classification research continues to emphasize the importance of contextual and structural relationships among labels and image regions for producing coherent predictions [12]. These developments reinforce the need for adversarial evaluation methods that are not only effective in manipulating target labels, but also sensitive to the contextual structure of non-target outputs.

Recent advances in consumer technologies and intelligent systems have highlighted increasing security and privacy risks associated with deep learning models [13]. In particular, adversarial attacks in multi-label settings can pose significant threats in applications such as smart surveillance, autonomous systems, and personalized recommendation platforms, where multiple correlated attributes are jointly predicted [14]. Recent studies have emphasized the need for robust and context-aware defenses in such environments [12]. These developments further motivate the importance of evaluating adversarial robustness under structured label dependencies, as considered in this work.

To address these challenges, we propose CAPG (Context-Aware Perturbation Generation), an adversarial attack framework specifically designed for multi-label deep learning models. As illustrated in Figure 1, CAPG leverages inter-label dependencies by integrating a correlation-weighted regularization term into the adversarial objective, enabling selective manipulation of target labels while preserving the semantic integrity of non-target outputs. By using label co-occurrence statistics derived from the Pascal VOC 2012 dataset [15], CAPG constrains adversarial perturbations to align with natural label structure, thereby producing contextually consistent adversarial examples. This correlation-aware perspective offers a new direction for adversarial evaluation in multi-label settings and highlights important limitations in existing threat models. While CAPG introduces a context regularizer, its primary effect is to reshape the gradient field, rather than merely penalize the objective value.

The key contributions of this work are summarized as follows:

- We introduce CAPG, a novel context-aware adversarial attack framework that explicitly models label correlations and integrates them into the perturbation generation process for multi-label deep learning models.
- We propose a selective adversarial objective that enables targeted manipulation of chosen labels while preserving semantically correlated non-target labels, ensuring high contextual consistency in the adversarial outputs.
- We develop a correlation-weighted perturbation mechanism in which the learned correlation matrix reshapes the gradient direction itself, yielding a dependency-aware update rule that cannot be achieved by classical PGD-style attacks.

- We perform extensive experiments on the Pascal VOC 2012 dataset using a ResNet-101 backbone, showing that CAPG achieves higher targeted attack success with markedly improved semantic consistency and lower perceptual distortion compared to strong baseline attacks.

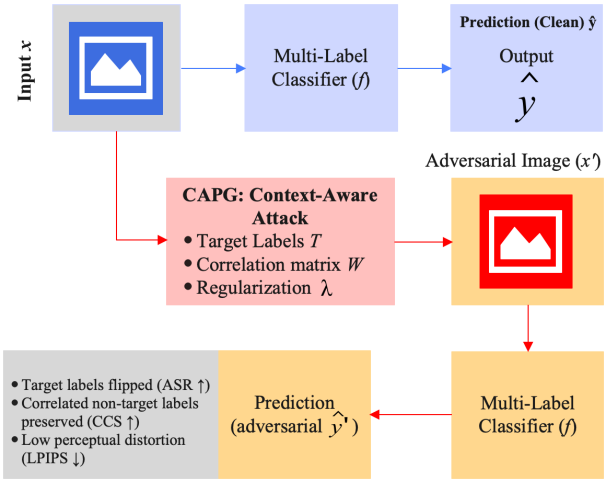


Figure 1. Overview of the Context-Aware Adversarial Attack. CAPG generates a visually similar adversarial image x' using target labels T, the correlation matrix W, and regularization weight λ. The modified image x' flips the target labels while preserving correlated non-target labels, resulting in semantically consistent adversarial predictions $\hat{y}'$.

The remainder of this paper is structured as follows. Section 1 reviews related work on multi-label learning, adversarial attacks, and correlation modeling. Section 2 defines the problem setting and introduces the correlation-aware adversarial objective. Section 3 details the proposed CAPG framework. Section 4 presents the experimental setup. Section 5 provides the results and analysis. Section 6 concludes the paper.

2. Related Work

This section reviews prior work related to multi-label deep learning, adversarial attacks, and context or correlation modeling, and then highlights the specific gap addressed by the proposed CAPG framework.

2.1. Multi-Label Deep Learning

Multi-label classification differs from standard single-label (multi-class) classification in that each input can be associated with multiple, potentially overlapping labels. Early approaches often relied on problem-transformation methods such as binary relevance [16], which trains an independent classifier for each label, or classifier chains [17], which model conditional dependencies by feeding predictions of previous labels into subsequent classifiers. While these methods are simple and flexible, they struggle to fully capture complex, high-order label interactions in large-scale visual datasets.

Deep learning has significantly advanced multi-label modeling [18], with recent work emphasizing improved loss design and imbalance-aware optimization strategies such as Asymmetric Loss [19], which have become strong baselines for large-scale multi-label image classification. CNN-based approaches (e.g., VGG, ResNet, DenseNet) produce shared visual features, followed by a sigmoid output layer. Many such models implicitly assume that labels are conditionally independent given the input, corresponding to the probabilistic factorization:

$$p(\mathbf{y} \mid x) = \prod_{k=1}^{K} p(y_k \mid x), \tag{1}$$

where $y_k \in \{0, 1\}$ denotes the k-th label. Although convenient, the independence assumption in (1) is often violated in datasets such as Pascal VOC 2012, where labels exhibit strong co-occurrence patterns.

To address this, many multi-label methods incorporate label dependencies explicitly. Some compute pairwise dependencies via co-occurrence, while others build label graphs [20]. A common method for estimating label co-occurrence is the empirical joint probability [21]:

$$P_{ij} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(y_i^{(n)} = 1 \wedge y_j^{(n)} = 1), \tag{2}$$

where P_{ij} measures how frequently labels i and j appear together across N training samples.

Advanced models such as Graph Convolutional Networks (GCNs) propagate information through label relations derived from (2). Transformer-based architectures further enhance global-context modeling via self-attention mechanisms [22], with recent approaches such as Query2Label demonstrating effective label-wise query modeling for multi-label classification [23]. Yet, all such methods operate under clean conditions, and their behavior under adversarial perturbations, particularly in the presence of structured label dependencies, remains largely unexplored.

2.2. Adversarial Attacks

Adversarial attacks expose vulnerabilities of deep networks by producing imperceptible perturbations that cause significant prediction errors [24]. FGSM [4] computes an adversarial example in a single step using:

$$x' = x + \epsilon \cdot \text{sign}\big(\nabla_x L(f(x), y)\big), \tag{3}$$

while PGD [5] iteratively applies gradient ascent with projection onto an ℓ_p ball.

Optimization-based attacks such as the Carlini–Wagner (CW) method formulate adversarial generation as a minimizing perturbation norm subject to classification constraints [6]. DeepFool linearizes the model to compute minimal decision-boundary-crossing perturbations [7].

Extending attacks to multi-label classification typically involves summing per-label losses. Prior studies have examined adversarial vulnerabilities in multi-label models, showing that independently optimized label objectives can lead to unintended label interactions and instability [25]. A widely used multi-label attack formulation is:

$$\delta^* = \arg \max_{\|\delta\|_p \leq \epsilon} \sum_{k=1}^{K} L\big(f_k(x + \delta), y_k\big), \tag{4}$$

where $f_k(x)$ is the predicted probability for label k. However, Equation (4) does not incorporate label dependencies, resulting in perturbations that arbitrarily modify correlated labels, reducing both realism and stealth.

In multi-label tasks, targeted attacks involve changing a subset of labels while keeping others fixed. Existing methods lack the mechanisms to preserve contextual consistency, leading to unnatural label changes (e.g., removing "car" while also removing "road"), which undermines their utility in safety-critical domains.

2.3. Context Modeling and Correlation Learning

Label correlation modeling is fundamental in multi-label learning. One common representation is the correlation matrix:

$$C_{ij} = \frac{\mathrm{cov}(y_i, y_j)}{\sigma(y_i)\,\sigma(y_j)}, \tag{5}$$

which captures pairwise dependencies between labels. Correlation matrices derived from Equation (5) can be used to build label graphs, priors, or structured regularizers.

Graph-based approaches use C_{ij} to define adjacency matrices for Graph Convolutional Networks (GCNs) or Graph Attention Networks (GATs) [26]. These architectures propagate label-context information through graph edges. Semantic similarity approaches embed label names into a vector space (e.g., Word2Vec, GloVe, FastText), forming a semantic similarity matrix:

$$S_{ij} = \cos\left(e_i, e_j\right), \tag{6}$$

where e_i and e_j are embedding vectors for labels i and j. Equation (6) captures high-level semantic relations that may not be reflected in dataset-level co-occurrence statistics.

However, although context modeling plays a vital role in multi-label prediction, existing adversarial attacks do not exploit either C_{ij} or S_{ij} when generating perturbations. Standard attacks therefore remain correlation-agnostic, often producing unrealistic label changes that are inconsistent with natural multi-label semantics.

2.4. Research Gap

Despite major advances in multi-label learning and dependency modeling, adversarial attack research has only limitedly incorporated these insights. Two key gaps remain. First, existing adversarial attacks do not explicitly integrate dataset-derived label dependencies into the adversarial optimization process. Classical attacks rely on independent-label losses as seen in Equation (4), ignoring co-occurrence and correlation structures such as those described in Equations (2)–(5). Second, existing methods do not provide a mechanism for selective, context-preserving attacks. As a result, perturbations often introduce unrealistic changes to correlated non-target labels, reducing both plausibility and evaluation value.

These limitations motivate the need for a correlation-aware adversarial framework that can selectively manipulate target labels while preserving contextual structure encoded in label-dependency information. The proposed CAPG framework directly addresses this gap. Table 1 summarizes representative methods across multi-label learning, adversarial attacks, and correlation modeling. To the best of our knowledge, prior work has not incorporated dataset-derived label-correlation matrices directly into the adversarial objective to generate correlation-aware, context-preserving perturbations in multi-label settings.

Table 1. Comparison of representative methods in multi-label learning and adversarial attacks, highlighting correlation modeling and context-aware attack capabilities.

Method	Setting	Corr.	Adv.	Context-Aware Attack
FGSM [4]	Single-label	None	Yes	No
PGD [5]	Single-label	None	Yes	No
Carlini–Wagner [6]	Single-label	None	Yes	No
DeepFool [7]	Single-label	None	Yes	No
Binary Relevance [16]	Multi-label	None	No	No
Classifier Chains [17]	Multi-label	Sequential	No	No
GCN-based Models [27]	Multi-label	Graph-based	No	No
Embedding-based Models [28]	Multi-label	Embedding-based	No	No
Adversarial Training (Multi-label) [29]	Multi-label	Implicit	Yes	No
CAPG (Ours)	Multi-label	Explicit	Yes	Yes

3. Problem Formulation

This section formalizes the multi-label adversarial setting considered in this work. We first introduce the notation for multi-label classification, then describe how label correlations are represented, and finally define the correlation-aware adversarial objective used in CAPG.

3.1. Multi-Label Classification Setup

Let $x \in \mathbb{R}^d$ denote an input image and

$$\mathbf{y} = [y_1, \ldots, y_K]^\top \in \{0,1\}^K \tag{7}$$

its corresponding multi-label annotation over K classes. A multi-label classifier

$$f : \mathbb{R}^d \to [0,1]^K, \quad \hat{\mathbf{y}} = f(x) \tag{8}$$

outputs a prediction vector $\hat{\mathbf{y}} = [f_1(x), \ldots, f_K(x)]^\top$, where $f_k(x)$ is the predicted probability that label k is present.

Training typically minimizes a multi-label loss such as binary cross-entropy across all labels. Unlike single-label classification, the components of y are not mutually exclusive and often exhibit strong co-occurrence patterns (e.g., person–bicycle, car–road), which are important for generating coherent predictions

3.2. Label Correlation Space

To capture dependencies among labels, we represent label relationships using dataset-level statistical structure. In particular, we estimate pairwise label correlations and obtain

$$C \in \mathbb{R}^{K \times K}, \qquad C_{ij} \in [-1,1], \tag{9}$$

where C_{ij} reflects statistical dependence between labels i and j.

For CAPG, we further construct a dependency weighting matrix W from dataset-level label statistics, which is explicitly incorporated into the adversarial optimization process. Intuitively, a large value of W_{ij} indicates that when label i is selected as an adversarial target, label j should be preserved to maintain contextual consistency. The exact construction of W is introduced in Section 4.3.2.

This dependency space provides a principled way to distinguish between target labels that the attacker aims to manipulate and non-target labels that should remain stable to preserve contextual plausibility.

3.3. Adversarial Objective

We now formalize the adversarial goal. Let $T \subseteq \{1, \ldots, K\}$ denote the set of target labels that the attacker aims to manipulate, and let T^c denote the set of non-target labels. An adversarial example is defined as $x' = x + \delta$, where the perturbation δ is constrained by

$$\|\delta\|_p \leq \epsilon, \tag{10}$$

so that x' remains visually similar to the original input x.

Classical attacks in the multi-label setting typically maximize a loss over all labels without considering dependencies, which often leads to unnecessary changes in T^c. In contrast, CAPG introduces a correlation-regularized objective that simultaneously promotes manipulation of target labels and preservation of strongly correlated non-target labels. At a high level, the CAPG adversarial objective is written as

$$\delta^* = \arg \max_{\|\delta\|_p \le \epsilon} \left[L_{\text{target}}(x+\delta) - \lambda L_{\text{context}}(x+\delta; W) \right], \tag{11}$$

where L_{target} drives manipulation of labels in T, L_{context} penalizes deviations in the predicted outputs of non-target labels in T^c, with penalties scaled by the dependency weights in W, and $\lambda > 0$ controls the trade-off between attack strength and contextual preservation.

Instead of treating each label independently, CAPG explicitly exploits the dependency structure encoded in W to generate selective, context-aware adversarial perturbations.

4. Proposed Method: CAPG Framework

This section presents CAPG, a unified adversarial attack framework designed to generate selective and contextually consistent perturbations for multi-label deep learning. CAPG builds directly on the problem formulation in Section 2 by incorporating label correlations into the attack objective, enabling adversaries to alter specific labels while preserving the contextual structure of the prediction vector.

4.1. CAPG Formulation

Before detailing the architectural components of CAPG, we formally define the correlation-aware adversarial objective. Let $x \in \mathbb{R}^d$ denote the clean input image, $y \in \{0,1\}^K$ the associated multi-label annotation, and $T \subseteq \{1,\dots,K\}$ the set of target labels the attacker aims to manipulate. The complement T^c contains non-target labels. An adversarial example is given by $x' = x + \delta$, where the perturbation satisfies $\|\delta\|_p \le \epsilon$.

Target Loss.

For each $k \in T$, we define the targeted misclassification loss as

$$L_{\text{target}}(x+\delta) = \sum_{k \in T} \ell\Big(f_k(x+\delta),\, y_k^{\text{adv}}\Big), \tag{12}$$

where $y_k^{\text{adv}} \in \{0,1\}$ denotes the adversarial target state for label k, and $\ell(\cdot)$ is binary cross-entropy.

Context Preservation Loss.

To penalize unwanted changes to correlated non-target labels, CAPG introduces:

$$L_{\text{context}}(x+\delta; W) = \sum_{i \in T} \sum_{j \in T^c} W_{ij} \left(f_j(x+\delta) - f_j(x)\right)^2 \tag{13}$$

where $W \in \mathbb{R}^{K \times K}$ is the dependency matrix introduced in Section 3.3, and W_{ij} specifies how strongly label j should be preserved when manipulating label i.

CAPG Objective.

The correlation-aware adversarial objective is therefore

$$\max_{\|\delta\|_p \le \epsilon} \Big[L_{\text{target}}(x+\delta) - \lambda\, L_{\text{context}}(x+\delta; W) \Big], \tag{14}$$

where $\lambda > 0$ controls the trade-off between targeted manipulation and contextual preservation.

Iterative Gradient Update.

The perturbation update at iteration t for CAPG-PGD is

$$\delta^{(t+1)} = \Pi_{\|\delta\|_p \le \epsilon} \Big(\delta^{(t)} + \eta\, \text{sign}\big(\nabla_x \left[L_{\text{target}} - \lambda L_{\text{context}} \right]\big) \Big), \tag{15}$$

where Π denotes projection onto the ℓ_p ball and η is the step size. This update rule formalizes the dependency-aware perturbation dynamics central to CAPG.

4.2. Overview Architecture

We introduce CAPG, a framework that integrates inter-label correlation modeling directly into the perturbation optimization process, such that label dependencies explicitly influence the gradient used to update the adversarial perturbation. Importantly, CAPG is not a simple extension of PGD with an added regularizer; the correlation matrix reshapes the gradient itself, producing a dependency-aware update rule that classical PGD-style attacks cannot replicate. Unlike classical attacks that treat all output dimensions independently, CAPG uses correlation information to guide perturbations such that:

- Targeted labels in T are effectively manipulated;
- Labels highly correlated with T (identified via W) remain unchanged;
- Resulting adversarial examples remain contextually plausible.

At a high level, CAPG contains three core modules:

1. Correlation Modeling Module, which constructs the correlation weighting matrix W from the dataset.
2. Context-Aware Loss Module, which formulates a correlation-regularized objective.
3. Perturbation Generator, which applies gradient-based optimization to produce $x' = x + \delta$.

The overall interaction between the correlation modeling, loss formulation, and perturbation generation modules is depicted in the system architecture shown in Figure 2.

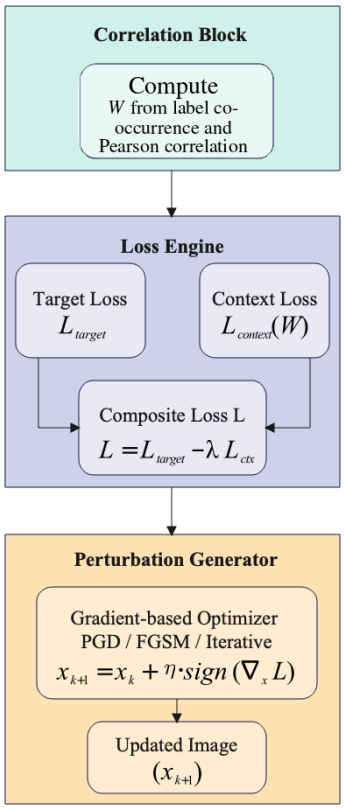


Figure 2. Architecture of CAPG framework. The system consists of three modules: (i) the Correlation Block, which computes the dependency matrix W from label co-occurrence and correlation statistics; (ii) the Loss Engine, which combines the target loss L_{target} and context loss $L_{\text{context}}(W)$ into a composite objective $L = L_{\text{target}} - \lambda L_{\text{ctx}}$; and (iii) the Perturbation Generator, which performs iterative gradient-based updates $x_{k+1} = x_k + \eta \cdot \text{sign}(\nabla_x L)$ to produce the adversarial image x_{k+1}.

4.3. Correlation Modeling

CAPG leverages statistical dependencies in the label space to determine which labels must remain stable during targeted attacks. Using the notation introduced in Section 2, we compute the empirical co-occurrence matrix P and correlation matrix C and derive a fixed, dataset-level correlation-weighted matrix W, which is reused throughout adversarial optimization.

4.3.1. Extracting Statistical Dependencies

The training set provides binary label vectors $\mathbf{y}^{(n)}$, from which we compute:

- Co-occurrence frequencies, capturing how often label pairs appear together.
- Pearson correlations, capturing linear dependencies.

These statistical signals reflect natural co-occurrence structures of the dataset. For example, in Pascal VOC 2012, car frequently co-occurs with road, and person frequently co-occurs with bicycle.

4.3.2. Constructing the Correlation Weighting Matrix W

Using P and C, we derive a correlation-preservation matrix:

$$W_{ij} = \alpha P_{ij} + (1 - \alpha)\,\frac{C_{ij} + 1}{2},$$

where α controls the balance between frequency-driven and correlation-driven dependencies. This linear combination was chosen as a simple and interpretable way to integrate two complementary signals of label dependency. The co-occurrence term captures how frequently labels appear together in the dataset, while the Pearson correlation term reflects linear dependency structure beyond raw frequency alone. Combining these two components allows CAPG to preserve both dominant co-occurrence patterns and broader statistical relationships in the label space without introducing excessive modeling complexity.

The weighting parameter α was selected empirically to provide a balanced and stable representation of label dependencies for the Pascal VOC 2012 setting. To ensure that W is numerically stable and meaningful, the transformed correlations $(C_{ij} + 1)/2$ map negative correlations into the non-negative range, and a min–max normalization is applied to keep all $W_{ij} \in [0, 1]$. Furthermore, entries below a small threshold τ (empirically, $\tau = 0.05$) are set to zero, preventing weak or noisy dependencies from influencing the context-preservation loss.

The dependency matrix W is derived from dataset-level label statistics and is therefore dataset-dependent. However, the CAPG framework itself is not restricted to a specific dataset. In particular, the proposed method does not rely on any dataset-specific assumptions beyond the availability of label statistics. It operates on a generic dependency matrix and can be applied to other multi-label datasets by recomputing W using the corresponding label distributions. This means that CAPG can be directly extended to datasets with different label densities and co-occurrence structures (e.g., larger-scale benchmarks) without modifying the underlying optimization framework. In addition, W can be constructed using alternative sources of label relationships, such as external datasets, embedding-based similarities, or graph-based representations. This flexibility allows CAPG to generalize across different multi-label domains while adapting to dataset-specific dependency structures.

Intuitively, if W_{ij} is high, label j should be strongly protected when attacking label i, whereas if W_{ij} is low, label j may change without substantially affecting contextual plausibility. The matrix W therefore provides a structured mechanism for context preservation in the adversarial optimization process.

4.4. Context-Aware Perturbation Generator

Unlike standard multi-label adversarial attacks, where the adversarial objective is a simple sum of independent per-label losses, CAPG introduces explicit coupling between target and non-target labels through a correlation-weighted context loss. As a result, the gradient of the adversarial objective with respect to the input is no longer separable across labels, and the resulting perturbation updates are explicitly shaped by inter-label dependencies. The overall CAPG attack procedure is summarized in Algorithm 1.

4.4.1. Correlation-Weighted Adversarial Loss

Let T be the set of target labels and T^c its complement. CAPG combines two loss components:

- Target Loss $\mathcal{L}_{\text{target}}$ pushes target labels toward adversarial values.
- Context Loss $\mathcal{L}_{\text{context}}$ penalizes changes to correlated labels using W.

For comparison, a standard multi-label adversarial attack typically optimizes the objective:

$$L_{\text{std}}(x) = \sum_{k=1}^{K} L\big(f_k(x), y_k\big), \tag{16}$$

Whose gradient with respect to the input is:

$$\nabla_x L_{\text{std}}(x) = \sum_{k=1}^{K} \nabla_x L\big(f_k(x), y_k\big). \tag{17}$$

This formulation treats each label independently and produces an adversarial update that is simply the sum of per-label gradients.

The overall objective is:

$$\delta^* = \arg\max_{\|\delta\|_p \le \epsilon} \Big[\mathcal{L}_{\text{target}}(x+\delta) - \lambda\, \mathcal{L}_{\text{context}}(x+\delta; W)\Big],$$

where λ controls the trade-off between targeted manipulation and context preservation.

4.4.2. Gradient-Based Optimization Under ℓ_p Constraints

In CAPG, the adversarial objective is defined as:

$$L_{\text{CAPG}}(x) = L_{\text{target}}(x) - \lambda\, L_{\text{context}}(x; W), \tag{18}$$

where L_{target} drives the manipulation of selected target labels. A generic form of the context loss is given by

$$L_{\text{context}}(x; W) = \sum_{i \in T} \sum_{j \in T^c} W_{ij} \big(f_j(x) - f_j(x_0)\big)^2, \tag{19}$$

where x_0 denotes the original clean input.

Taking the gradient of $L_{\text{CAPG}}(x)$ with respect to the input yields

$$\nabla_x L_{\text{CAPG}}(x) = \sum_{i \in T} \nabla_x L\big(f_i(x), y_i^{\text{adv}}\big) - \lambda \sum_{i \in T} \sum_{j \in T^c} W_{ij} \nabla_x \big(f_j(x) - f_j(x_0)\big)^2. \tag{20}$$

Unlike the standard multi-label gradient, this expression couples the gradients of target and non-target labels through the correlation matrix W, thereby reshaping the optimization landscape and producing dependency-aware perturbation updates. This ensures that perturbations remain imperceptible and respect norm constraints.

Algorithm 1 CAPG Attack Procedure

Require: Input x, labels $\mathbf{y}$, target set T, correlation matrix W, step size η, norm bound ϵ

Initialize $\delta \leftarrow 0$
for $t = 1$ to $T_{\max}$ **do**
    Compute gradient:

$$g \leftarrow \nabla_x \big(\mathcal{L}_{\text{target}}(x+\delta) - \lambda \, \mathcal{L}_{\text{context}}(x+\delta; W) \big)$$

    Update perturbation: $\delta \leftarrow \delta + \eta \, \text{sign}(g)$
    Project: $\delta \leftarrow \Pi_{\|\delta\|_p \leq \epsilon}(\delta)$
end for
return $x' = x + \delta$

4.5. Attack Variants

In this work, we evaluate two practical variants derived directly from CAPG formulation. Both variants use the same correlation-guided objective but differ in how the perturbation is applied and optimized.

4.5.1. CAPG-FGSM

CAPG-FGSM is a single-step approximation that applies a one-shot perturbation in the direction of the correlation-weighted gradient. The update is given by:

$$\delta = \epsilon \cdot \text{sign}\big(\nabla_x (\mathcal{L}_{\text{target}} - \lambda \, \mathcal{L}_{\text{context}}) \big),$$

where ϵ controls the perturbation magnitude. This variant is computationally efficient because it requires only one backward pass. Although it may not achieve the same level of precision as iterative methods, CAPG-FGSM is useful for scenarios where:

- Real-time adversarial testing is needed;
- Model queries are limited;
- Low-latency adversarial generation is desired.

Compared to standard FGSM, the inclusion of $\mathcal{L}_{\text{context}}$ ensures that correlated non-target labels are preserved, producing more coherent adversarial predictions even with a single update step.

4.5.2. CAPG-PGD

CAPG-PGD is the iterative extension of CAPG-FGSM and serves as the strongest attack variant in the CAPG family. Following the optimization process described earlier, CAPG-PGD repeatedly updates the perturbation using:

$$\delta^{(t+1)} = \Pi_{\|\delta\|_p \leq \epsilon} \Big(\delta^{(t)} + \eta \cdot \text{sign}\big(\nabla_x (\mathcal{L}_{\text{target}} - \lambda \, \mathcal{L}_{\text{context}}) \big) \Big),$$

where η is the step size and $\Pi(\cdot)$ denotes projection onto the ℓ_p ball.

By iteratively refining the perturbation, CAPG-PGD:

- Achieves more reliable targeted label flipping;
- Preserves correlated labels more effectively;
- Produces adversarial examples with higher semantic consistency.

Due to its iterative nature, CAPG-PGD requires more computation than CAPG-FGSM but offers significantly higher attack strength and stability. As a result, it serves as the primary benchmark for evaluating the upper bound of CAPG's performance.

4.6. Computational Complexity

The CAPG framework is designed to be computationally efficient and introduces only a small overhead compared to classical adversarial attacks. Since CAPG reuses the same forward and backward passes as FGSM and PGD, the dominant operations remain unchanged [30].

- Gradient computation: CAPG computes gradients of the correlation-weighted loss $(\mathcal{L}_{\text{target}} - \lambda\, \mathcal{L}_{\text{context}})$, which adds only one additional term to the backward pass. The computational cost is therefore nearly identical to the gradient used in standard PGD.
- Correlation matrix lookup: Accessing correlation weights W_{ij} is performed via simple matrix indexing and incurs constant time cost $O(1)$ per label. These values are precomputed once and reused throughout the attack.
- Overall complexity: For an iterative attack with $T_{\max}$ iterations and input dimensionality d, the total computational cost remains:

$$O(T_{\max} \cdot d),$$

which matches the complexity of PGD. The additional correlation penalty introduces no extra forward passes or heavy operations (e.g., graph convolutions).

Consequently:

- CAPG-FGSM requires only one gradient evaluation and runs in the same time as standard FGSM.
- CAPG-PGD: since the correlation matrix W is fixed and precomputed, CAPG introduces no additional asymptotic cost beyond standard PGD, differing only by a small constant factor due to the context-loss gradient computation.

Overall, CAPG achieves context-aware and semantically consistent perturbations without compromising computational efficiency, making it suitable for both large-scale evaluation and real-time adversarial testing.

5. Experimental Setup

This section describes the threat model, dataset, model architecture, baseline attacks, evaluation metrics, and implementation details used to assess the effectiveness of the proposed CAPG framework.

5.1. Threat Model and Attack Setting

We consider a standard white-box adversarial setting for multi-label classification. While black-box methods such as Square Attack [31] demonstrate strong query-efficient performance under limited model access, this work focuses on the white-box setting to isolate the impact of correlation-aware gradient manipulation. The attacker has full access to the model parameters, logits, and gradients, enabling exact gradient-based optimization of the perturbation δ. The attacker is not allowed to modify model weights or the training process; only the input image is perturbed.

All attacks in this work are targeted multi-label attacks. Given a target label set T, the adversary aims to manipulate the predicted binary states of labels in T toward adversarial outcomes while preserving the predicted states of all non-target labels in T^c.

The attack is label-space targeted, meaning that CAPG operates directly on semantic label outputs rather than optimizing specific logit values or margins. Perturbations are constrained by an ℓ_p norm bound $\|\delta\|_p \leq \epsilon$, and the adversarial image $x' = x + \delta$ is required to remain visually similar to the clean input.

Attacks are evaluated only on validation images that contain at least one positive instance of the selected target label(s). Unless otherwise stated, the target label set T is selected uniformly at random from the positive labels of each image to avoid heuristic or bias-driven selection.

5.2. Dataset: Pascal VOC 2012

All experiments are conducted on the Pascal VOC 2012 dataset [15], a widely used benchmark for multi-label image classification. The dataset contains 20 object categories and exhibits well-documented label co-occurrence patterns (e.g., person–bicycle, car–road), which makes it suitable for evaluating correlation-aware adversarial perturbations. Following standard practice, we use the official train/validation split consisting of 5717 training images and 5823 validation images.

Pascal VOC 2012 is well suited for evaluating context-aware adversarial attacks, as its dense and semantically consistent label co-occurrence patterns make indiscriminate label perturbations both unrealistic and easily detectable.

5.3. Model Architecture

We adopt ResNet-101 [32] as the backbone multi-label classifier due to its strong representational capacity and its frequent use in adversarial robustness research. The model is initialized with ImageNet-pretrained weights and fine-tuned on Pascal VOC using a sigmoid output layer over 20 classes. Binary cross-entropy loss is used for training [33].

ResNet-101 is selected to ensure that attack performance is evaluated on a strong and realistic multi-label backbone rather than being limited by model capacity. During adversarial evaluation, all model parameters remain fixed; attacks operate solely by perturbing the input image [34]. CAPG is architecture-agnostic and does not rely on any specific backbone design. In particular, the proposed framework operates at the level of the loss function and gradient computation, and can therefore be directly applied to other multi-label architectures, including recent Transformer-based models such as Query2Label, which produce label-wise probability outputs.

5.4. Baseline Attacks

To benchmark CAPG against strong baselines, we consider four widely used attacks adapted to the multi-label setting:

- FGSM [4]: A single-step gradient sign method.
- PGD [5]: An iterative projected gradient descent attack under ℓ_∞ constraints.
- Carlini–Wagner (CW) [6]: An optimization-based targeted attack. Since the original CW formulation is designed for single-label classifiers, we adapt it by applying the loss independently to each target label $k \in T$ and aggregating the terms using a binary cross-entropy objective. Non-target labels are left unconstrained, which often leads to unintended changes in correlated outputs.
- DeepFool [7]: A decision-boundary approximation method for minimal perturbations. In the multi-label setting, DeepFool is applied independently to each target logit dimension, and the smallest resulting perturbation is selected, following standard multi-output extensions.

Each baseline is implemented under the same white-box assumptions and perturbation constraints as CAPG. Since these attacks optimize label-wise losses independently and do not incorporate label correlation or contextual preservation, they provide a fair correlation-agnostic comparison. We note that these baselines are architecture-independent and can be applied to both CNN-based and Transformer-based multi-label models, ensuring that the comparison focuses on the attack formulation rather than the choice of backbone.

5.5. Evaluation Metrics

We evaluate adversarial performance using four primary metrics that jointly capture attack effectiveness, contextual preservation, and perceptual realism. Attack Success Rate (ASR) measures the fraction of target labels whose predicted binary states are successfully flipped to the desired adversarial outcome. Contextual Consistency Score (CCS) measures the fraction of non-target labels whose predicted binary states remain unchanged under adversarial perturbation. Formally, for an adversarial example x', let T^c denote the set of non-target labels. The CCS is defined as

$$\mathrm{CCS}(x') = \frac{1}{|T^c|} \sum_{j \in T^c} \mathbb{I}\big[\hat{y}_j(x') = \hat{y}_j(x)\big], \tag{21}$$

where $\hat{y}_j(x)$ and $\hat{y}_j(x')$ denote the predicted binary states of label j on the clean and adversarial inputs, respectively, obtained using the same decision threshold, and $\mathbb{I}[\cdot]$ is the indicator function. Higher CCS values indicate stronger preservation of non-target labels.

While CCS captures whether binary predictions remain unchanged, it does not account for changes in prediction confidence that do not cross the decision threshold. As such, it may overestimate preservation when substantial probability shifts occur without label changes, and it may be sensitive to predictions near the threshold, where small perturbations can induce label flips. A more fine-grained perspective could be obtained by measuring changes in predicted probabilities for non-target labels, for example through the average absolute difference between pre- and post-attack predictions.

Mean Average Precision (mAP) measures the degradation in standard multi-label classification performance under adversarial perturbation, reflecting the overall functional impact of the attack. LPIPS is a learned perceptual similarity metric that quantifies the perceptual distance between clean and adversarial images, with lower values indicating higher visual fidelity. Together, ASR, CCS, mAP, and LPIPS provide a comprehensive evaluation of adversarial behavior, capturing discrete label changes as well as global functional and perceptual effects.

5.6. Hardware and Implementation Details

All experiments are implemented in PyTorch 2.8 and executed on an NVIDIA RTX-series GPU. All attacks use identical perturbation budgets and optimization parameters to ensure fair comparison across methods. The correlation matrix W used by CAPG is pre-computed once from the training set and reused throughout all experiments. For iterative attacks, including PGD and CAPG-PGD, we use a fixed number of iterations ($T_{\max} = 20$) with step size $\eta = 2/255$, matching standard PGD configurations unless otherwise stated. Model parameters remain fixed during adversarial evaluation, and all attacks operate solely on the input image. Table 2 summarizes the key experimental settings used across all evaluations.

Table 2. Simulation parameters for CAPG and baseline attacks.

Parameter	Value
Dataset	Pascal VOC 2012
Backbone	ResNet-101
Training Loss	Binary Cross-Entropy
Perturbation Norm	ℓ_∞
Perturbation Budget (ϵ)	8/255
PGD Iterations ($T_{\max}$)	20
PGD Step Size (η)	2/255
Decision Threshold	0.5

Table 2. *Cont.*

Parameter	Value
Targets per Image	1–2
Context Weight (λ)	0.5
Correlation Matrix (W)	Precomputed
Evaluation Metrics	ASR, CCS, mAP, LPIPS
Hardware	RTX-series GPU
Framework	PyTorch 2.8

6. Results and Analysis

This section presents quantitative and qualitative results evaluating CAPG against standard adversarial attacks. We report attack effectiveness, contextual consistency, perceptual distortion, and conduct ablation studies to understand the contribution of correlation modeling.

6.1. Quantitative Evaluation

We evaluate CAPG-FGSM and CAPG-PGD against four baselines: FGSM, PGD, CW, and DeepFool. All methods operate under the same ℓ_∞ perturbation budget ($\epsilon = 8/255$) and target the same label subsets across the Pascal VOC 2012 validation set.

6.1.1. Attack Success Rate (ASR)

As illustrated in Figure 3, proposed CAPG-PGD achieves the highest ASR of 92.8% among the evaluated methods, outperforming strong baselines such as PGD (85.7%) and CW (82.9%). This indicates that correlation-aware regularization improves the effectiveness of targeted label manipulation rather than weakening the attack objective.

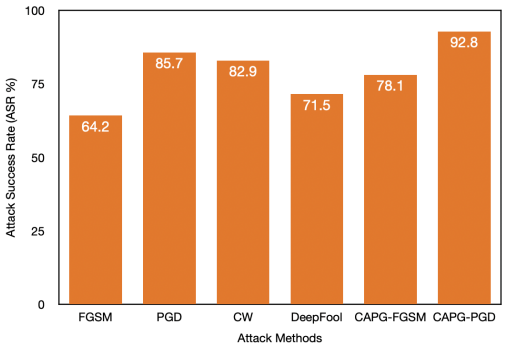


Figure 3. Attack Success Rate (ASR) comparison across adversarial attack methods on the Pascal VOC 2012 dataset. CAPG-PGD achieves the highest ASR, demonstrating the effectiveness of correlation-aware perturbations.

The results further show that correlation modeling also improves single-step attacks: CAPG-FGSM achieves an ASR of 78.1%, outperforming standard FGSM (64.2%). This demonstrates that even lightweight, one-step gradient methods gain greater effectiveness and precision when label dependencies are incorporated into the optimization objective. Overall, the comparison in Figure 3 identifies CAPG-PGD as the strongest targeted attack in this study.

6.1.2. Contextual Consistency Score (CCS)

As shown in Figure 4, classical attacks disrupt a substantial portion of non-target labels, with CCS scores ranging from 49.3% to 61.2%. In contrast, both CAPG variants significantly improve contextual preservation. CAPG-FGSM reaches 74.9%, while CAPG-PGD achieves

the highest CCS of 88.6%, demonstrating the benefit of incorporating correlation-aware regularization into the attack process. These results show that CAPG better supports selective manipulation in multi-label settings by reducing unintended changes to correlated non-target outputs.

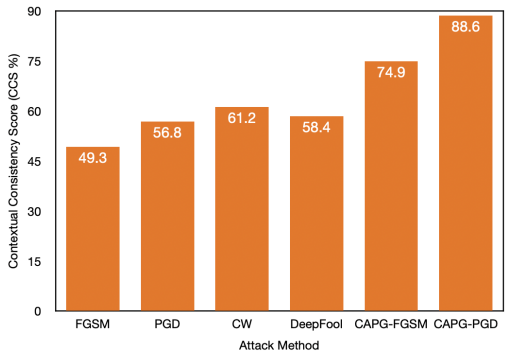


Figure 4. Contextual Consistency Score (CCS) across adversarial attack methods on the Pascal VOC 2012 dataset. CAPG-PGD achieves the highest CCS, indicating strong preservation of non-target labels under targeted perturbations.

6.1.3. mAP Degradation

As shown in Figure 5, classical attacks such as PGD and CW cause substantial mAP reduction (42.7% and 39.9%, respectively), reflecting their tendency to introduce broader disruption across the prediction space.

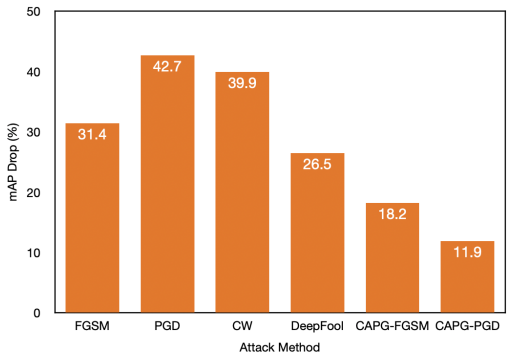


Figure 5. mAP degradation across adversarial attack methods on the Pascal VOC 2012 dataset. CAPG-PGD produces the smallest mAP drop, indicating more selective perturbations and reduced global disruption compared to classical attacks.

In contrast, both CAPG variants produce much smaller mAP drops. CAPG-FGSM reduces mAP by only 18.2%, while CAPG-PGD achieves the lowest degradation at 11.9%. These results indicate that correlation-aware regularization helps limit unnecessary changes to unrelated labels, enabling CAPG to manipulate target labels effectively while preserving overall predictive structure.

6.1.4. Perceptual Distortion (LPIPS)

As shown in Figure 6, classical attacks such as PGD and FGSM tend to produce higher perceptual distortion, reflected by LPIPS scores of 0.243 and 0.214, respectively. These methods tend to introduce broad, visually noticeable changes due to their lack of correlation awareness.

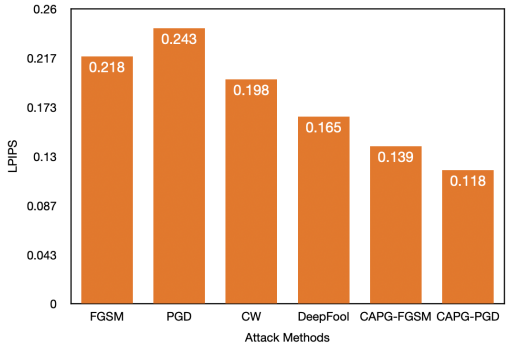


Figure 6. LPIPS perceptual distortion across adversarial attack methods on the Pascal VOC 2012 dataset. CAPG-PGD achieves the lowest LPIPS score, indicating more visually realistic and less perceptible perturbations.

Both CAPG variants achieve substantially lower distortion. CAPG-FGSM reduces the score to 0.139, while CAPG-PGD achieves the lowest LPIPS of 0.118, despite simultaneously delivering the highest ASR and CCS in earlier evaluations. Combined with the high CCS values reported earlier, these results indicate that CAPG achieves both perceptual and contextual stealth, producing adversarial examples that are difficult to detect visually while preserving non-target output structure.

6.2. Qualitative Analysis of Contextual Consistency

To illustrate the behavioral differences between standard and correlation-aware attacks, we qualitatively examine the perturbations generated by PGD and CAPG-PGD, as shown in Figure 7.

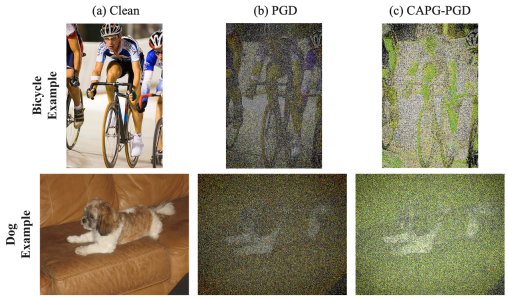


Figure 7. Qualitative comparison between standard PGD and the proposed CAPG-PGD attack on samples from the Pascal VOC 2012 validation set. For each clean image (**a**), PGD (**b**) and CAPG-PGD (**c**) illustrate different perturbation characteristics.

6.2.1. PGD Perturbations

Because PGD aggregates loss across labels without modeling dependencies, its perturbations often spill over to non-target outputs. For example, targeting the label "person" may unintentionally alter correlated labels such as "bicycle", while suppressing "dog" may distort predictions for other objects present in the same scene. Such unintended spillover disrupts the contextual structure of multi-label predictions and weakens natural co-occurrence patterns in Pascal VOC images.

6.2.2. CAPG-PGD Perturbations

In contrast, CAPG-PGD preserves the stability of correlated non-target labels while manipulating the intended targets. For instance, suppressing "person" retains the presence of "bicycle" in scenes where both typically co-occur, and adjusting "dog" predictions leaves unrelated labels unchanged, reflecting stronger contextual awareness. These examples also highlight differences in perturbation structure: PGD tends to introduce high-frequency,

spatially unstructured noise that affects multiple regions of the scene, whereas CAPG-PGD produces more spatially structured perturbations influenced by label dependencies. The examples shown in Figure 7 are representative samples from the evaluation set and illustrate the typical perturbation patterns observed across the dataset.

6.3. Ablation Study

To better understand the contribution of correlation modeling and the influence of the regularization weight λ, we conduct an ablation analysis using CAPG-PGD as the base setting. The results isolate the effect of removing the correlation structure and varying the strength of the context-preserving regularization.

6.3.1. Effect of Removing Correlation Modeling

To assess the importance of correlation weighting, we replace the learned correlation matrix W with the identity matrix ($W = I$), effectively removing the correlation-guided penalty and forcing CAPG-PGD to behave like a standard multi-label PGD variant. The quantitative comparison is shown in Figure 8.

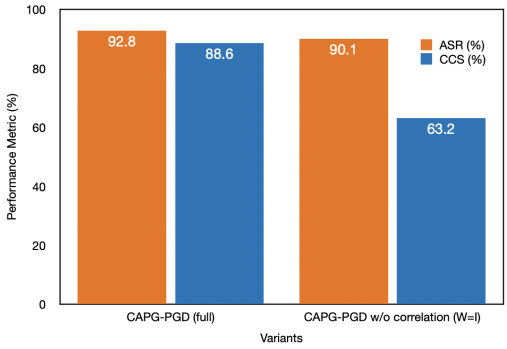


Figure 8. Ablation study on the role of correlation modeling in CAPG-PGD. Removing the correlation matrix ($W = I$) leads to a substantial drop in CCS, while ASR remains relatively stable, highlighting the importance of correlation-aware regularization for preserving non-target labels.

Although the ASR experiences only a small reduction (from 92.8% to 90.1%), the CCS drops sharply from 88.6% to 63.2%. Correlation modeling does not significantly reduce the ability to flip target labels, indicating that CAPG preserves attack strength while improving semantic stability. These results show that correlation awareness is central to improving contextual stability, particularly in preserving non-target label consistency, while maintaining strong attack effectiveness.

6.3.2. Effect of the Regularization Weight λ

We next evaluate the influence of the regularization weight λ, which controls the balance between targeted manipulation and context preservation in the composite objective. Specifically, λ determines the relative strength of the context-preservation term with respect to the target attack loss. A small λ prioritizes attack strength, allowing more aggressive perturbations that may disrupt correlated non-target labels, while a large λ increasingly penalizes such changes and enforces consistency among dependent labels. The effect of varying λ is shown in Figure 9.

The results reveal a clear trade-off. With $\lambda = 0$, the attack acts almost purely as a targeted optimization, achieving the highest ASR but very poor CCS (54.7%). Increasing λ improves CCS substantially, with values above 1.0 placing heavier emphasis on context preservation. However, a very large λ begins to restrict the perturbation excessively, reducing the attack's ability to flip target labels (ASR drops to 84.7% at $\lambda = 2.0$). Accordingly, $\lambda = 0.5$ is used as the default setting in subsequent experiments, as it provides a balanced

trade-off between attack success and contextual stability. This behavior highlights the role of λ as a key control parameter governing the strength–consistency trade-off in CAPG.

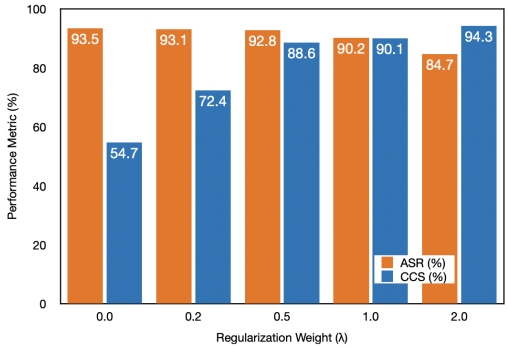


Figure 9. Ablation study on the effect of the regularization weight λ, illustrating the trade-off between attack success (ASR) and contextual consistency (CCS).

To further analyze the behavior of the regularization weight, we observe that the trade-off between ASR and CCS follows a smooth and monotonic trend as λ increases. In particular, there exists a transition region where contextual consistency improves significantly with only a minor reduction in attack success. Beyond this region, however, excessively large values of λ lead to over-constrained gradients, causing a more pronounced degradation in ASR.

This behavior indicates that λ controls the balance between exploration (target label manipulation) and constraint enforcement (context preservation). From a practical perspective, this provides a tunable mechanism to adjust CAPG depending on the desired trade-off between attack effectiveness and semantic plausibility.

The effect of varying τ is shown in Figure 10. The results show that CAPG remains relatively stable under moderate variations of τ. When τ is very small, a large number of weak correlations are retained in the matrix W, which may introduce noisy or less meaningful dependencies and slightly reduce contextual consistency. As τ increases, weak correlations are filtered out, leading to a more focused and effective context-preservation mechanism, with peak CCS observed around $\tau = 0.05$.

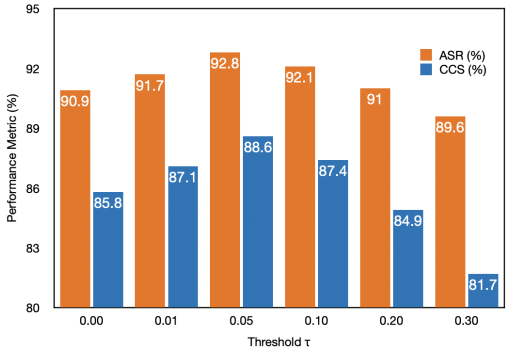


Figure 10. Ablation study on the effect of the threshold τ, illustrating its influence on Attack Success Rate (ASR) and Contextual Consistency Score (CCS).

However, when τ becomes too large, the correlation matrix becomes overly sparse, removing useful dependencies between labels. This weakens the contextual constraints, resulting in a gradual decline in CCS. In contrast, ASR remains relatively stable across different τ values, indicating that thresholding primarily affects context preservation rather than the ability to manipulate target labels.

These results confirm that τ controls the sparsity of the dependency structure and should be set to a moderate value to balance robustness and flexibility.

6.3.3. Effect of Prediction-Based and Ground-Truth-Based Correlation Matrices

To further examine the construction of the correlation matrix W, we compare two variants of CAPG: one using a prediction-based correlation matrix derived from the model's outputs, and another using a ground-truth-based correlation matrix derived from dataset annotations. This comparison is intended to evaluate whether the proposed framework is sensitive to the source of label dependency information and to assess its implications for semantic concealment.

Table 3 shows that both variants achieve strong performance, indicating that CAPG is robust to different choices of correlation construction. The prediction-based variant yields slightly higher CCS/ASR because it is more closely aligned with the model's internal decision structure during adversarial optimization. In contrast, the ground-truth-based variant reflects true semantic relationships more directly, but may introduce a modest mismatch with the model's learned label space. These results suggest that prediction-based construction is a reasonable design choice in the present setting, while ground-truth-based constraints remain a meaningful alternative for future extensions.

Table 3. Comparison of CAPG using prediction-based and ground-truth-based correlation matrices.

Correlation Source for W	ASR (%)	CCS (%)	LPIPS
Prediction-based W	92.8	88.6	0.041
Ground-truth-based W	91.9	86.9	0.042

6.3.4. Analysis of Gradient Direction

To provide more intuitive evidence of the gradient shaping mechanism in CAPG, we analyze how the gradient direction differs from that of a standard multi-label PGD attack. Specifically, let $g_{\text{PGD}} = \nabla_x L_{\text{std}}(x)$ denote the gradient of the standard multi-label attack objective, and $g_{\text{CAPG}} = \nabla_x L_{\text{CAPG}}(x)$ denote the gradient of the proposed correlation-aware objective. We measure the similarity between these two gradient directions using cosine similarity:

$$\cos(\theta) = \frac{g_{\text{PGD}} \cdot g_{\text{CAPG}}}{\|g_{\text{PGD}}\| \, \|g_{\text{CAPG}}\|}. \tag{22}$$

The cosine similarity is computed at the same input across multiple samples, and the aggregated results are reported in Table 4.

The results show that the cosine similarity remains consistently below 1, confirming that CAPG does not simply scale the original gradient but instead modifies its direction. At the same time, the similarity remains relatively high, indicating that the overall attack objective is preserved. This suggests that CAPG reshapes the gradient field by introducing structured adjustments guided by label correlations, effectively steering the perturbation toward preserving correlated non-target labels while still achieving targeted manipulation.

Table 4. Cosine similarity between CAPG gradients and standard PGD gradients across samples.

Gradient Comparison	Mean	Std	Min
PGD vs. CAPG	0.78	0.06	0.65

6.4. Discussion

6.4.1. Insights into Multi-Label Adversarial Vulnerability

The results highlight several important insights into multi-label adversarial vulnerability. First, CAPG demonstrates stronger semantic stealth than classical attacks. The consistently high Contextual Consistency Score (CCS), combined with low mAP degradation, indicates that CAPG preserves non-target label predictions more effectively while still

achieving high target attack success. This suggests that correlation-aware perturbations better respect the structured nature of multi-label outputs compared to correlation-agnostic methods such as PGD or CW.

The ablation results further confirm that correlation modeling plays a critical role in maintaining contextual consistency. While the ability to flip target labels remains largely intact even without correlation constraints, the preservation of correlated non-target labels deteriorates significantly. This demonstrates that correlation awareness is not required to induce misclassification per se, but is essential for producing semantically coherent adversarial behavior in multi-label settings.

At the same time, the consistently high Attack Success Rate (ASR) observed across CAPG variants indicates that modern multi-label deep learning models remain highly vulnerable, even when attacks are constrained to preserve contextual structure. This underscores the fact that semantic dependencies, while beneficial for prediction under clean conditions, do not inherently provide robustness against adversarial manipulation.

Furthermore, CAPG reveals adversarial failure modes that are not exposed by standard correlation-agnostic attacks. Classical evaluation pipelines relying solely on FGSM or PGD may fail to capture vulnerabilities involving selective, context-preserving label manipulation. By explicitly incorporating label dependencies into the attack objective, CAPG provides a more realistic stress test for multi-label systems deployed in structured prediction scenarios.

6.4.2. Limitations of Context Preservation

One important limitation of the current formulation is that the context-preservation objective is defined with respect to the model's clean predictions rather than ground-truth labels. As a result, CAPG enforces consistency within the model's prediction space, preserving non-target outputs as originally predicted by the model. While this design aligns with the standard adversarial threat model—where ground-truth labels are not available at inference time—it also introduces a limitation. If the model is incorrect on certain non-target labels, CAPG may preserve these incorrect predictions, potentially overestimating semantic consistency as measured by CCS. Therefore, CCS should be interpreted as a measure of model-consistency rather than strict semantic correctness. Future work may explore hybrid formulations that incorporate ground-truth signals or confidence-aware constraints to better align contextual preservation with true label semantics.

The proposed CAPG framework is primarily designed for input-space adversarial perturbations in multi-label classification models. It does not explicitly model hybrid attack scenarios that combine multiple attack channels (e.g., signal injection, communication-level disturbances, or control-layer attacks) as studied in networked control systems. However, CAPG can be viewed as complementary to such approaches, as it focuses on structured perturbations at the data level while preserving label dependencies. Extending CAPG to hybrid settings would require integrating additional system-level constraints or combining it with control-oriented attack models. This represents an interesting direction for future work, particularly in bridging data-driven adversarial learning with hybrid system security frameworks.

6.4.3. Limitations of Correlation Modeling

Another important consideration is the choice of correlation modeling used in CAPG. In this work, we employ a pairwise correlation matrix derived from co-occurrence statistics and linear correlation measures, which provides a simple and interpretable approximation of label dependencies. While this approach effectively captures dominant relationships in datasets such as Pascal VOC 2012, it may not fully represent higher-order or nonlinear dependencies present in more complex multi-label settings. More generally, CAPG op-

erates on a dependency matrix (W), which can be constructed using pairwise statistics, graph-based representations (e.g., GCN-derived adjacency matrices), or embedding-based similarity measures that capture semantic relationships between labels. These alternatives could enable richer modeling of label interactions and further improve context-aware adversarial generation.

In addition, CAPG may exhibit limitations in scenarios involving strong mutual exclusion or conflicting label relationships. Since the method enforces correlation-based constraints during optimization, situations where target and non-target labels are logically incompatible can introduce tension between the attack objective and context preservation. For example, suppressing one label while maintaining another that is semantically contradictory may lead to reduced attack success or unstable perturbation behavior. This highlights a fundamental limitation of correlation-based modeling, as statistical dependencies do not always capture strict logical constraints or higher-level semantic relationships. Addressing such cases may require incorporating more structured representations of label dependencies, such as hierarchical or logic-aware constraints, which remains an interesting direction for future research.

Although CAPG can operate with alternative dependency formulations, the current ablation study does not explicitly examine the sensitivity of the dependency matrix construction itself. In particular, the weighting parameter α, which balances co-occurrence statistics and Pearson correlation in forming W, introduces an additional design choice. Although a fixed value of α is used in this work to provide a stable and interpretable formulation, varying α could influence the relative contribution of frequency-based and correlation-based dependencies, potentially affecting contextual preservation. A systematic analysis of this sensitivity, as well as comparisons with alternative formulations, is left for future work.

6.4.4. Generalization and Future Directions

Finally, while the current evaluation focuses on the Pascal VOC 2012 dataset and a ResNet-101 backbone, this setting provides a controlled benchmark for analyzing context-aware adversarial behavior in multi-label classification. Extending CAPG to additional datasets such as MS-COCO, NUS-WIDE, or OpenImages would enable evaluation across more diverse label distributions and dependency structures. Similarly, assessing performance across different model architectures would further validate the generality of the approach. While the observed trends are consistent across metrics and ablation studies, future work may also incorporate formal statistical testing to further quantify the significance of performance differences.

7. Conclusions

This paper introduced CAPG (Context-Aware Perturbation Generation), a context-aware adversarial attack framework specifically designed for multi-label deep learning models. Unlike conventional attacks that treat output labels independently, CAPG explicitly incorporates inter-label dependencies into the adversarial objective, enabling selective manipulation of target labels while preserving correlated non-target outputs.

Through experiments on the Pascal VOC 2012 dataset using a ResNet-101 multi-label classifier, we showed that CAPG achieves high targeted attack success while substantially improving contextual consistency compared to classical methods such as FGSM, PGD, CW, and DeepFool under identical perturbation budgets. The results further show that correlation-aware perturbations can expose realistic adversarial failure modes that are not captured by correlation-agnostic attacks. At the same time, they highlight a controllable trade-off between attack strength and contextual preservation governed by the regularization parameter λ.

These findings reveal a critical gap in the robustness evaluation of multi-label systems. Standard adversarial benchmarks may overlook vulnerabilities involving selective, context-preserving label manipulation, leading to an incomplete assessment of model behavior in structured prediction settings. CAPG addresses this gap by providing a framework that aligns adversarial evaluation with the dependency structure inherent in multi-label outputs.

A limitation of the current formulation is that CAPG relies on a dataset-derived dependency matrix and enforces contextual consistency with respect to model predictions rather than ground-truth labels. While this design is consistent with realistic adversarial settings, it may preserve incorrect predictions and should therefore be interpreted as model-consistency rather than strict semantic correctness.

Future work will explore extending CAPG to additional datasets and model architectures to further validate its generality, as well as investigating alternative dependency modeling strategies, including graph-based and embedding-based approaches. Additional directions include incorporating confidence-aware or ground-truth-informed consistency constraints, and applying context-aware adversarial evaluation to video-based, transformer-based, and multimodal systems.

Author Contributions: Conceptualization, A.A. (Aidos Askhatuly), D.B., A.B., A.A. (Aigul Adamova) and D.Y.; methodology, A.A. (Aidos Askhatuly) and D.B.; software, A.A. (Aidos Askhatuly) and D.Y.; validation, A.A. (Aidos Askhatuly), D.B., A.B., A.A. (Aigul Adamova) and D.Y.; formal analysis, A.B. and A.A. (Aigul Adamova); investigation, A.A. (Aidos Askhatuly) and D.B.; resources, D.Y.; data curation, A.A. (Aidos Askhatuly) and A.B.; writing—original draft preparation, A.A. (Aidos Askhatuly); writing—review and editing, D.B., A.B., A.A. (Aigul Adamova) and D.Y.; visualization, A.A. (Aidos Askhatuly) and A.A. (Aigul Adamova); supervision, D.Y.; project administration, A.A. (Aigul Adamova); funding acquisition, D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement").

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions related to the specific nature of the datasets used for evaluating adversarial attacks.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAPG	Context-Aware Perturbation Generation
ASR	Attack Success Rate
CCS	Contextual Consistency Score
FGSM	Fast Gradient Signed Method
PGD	Projected Gradient Descent
CW	Carlini–Wagner
VOC	Visual Object Classes
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
GAT	Graph Attention Network
mAP	Mean Average Precision
LPIPS	Perceptual Distortion

References

1. Yang, M.; Zhao, Y.; Yu, H.; Chen, S.; Gao, G.; Li, D.; Wu, X.; Huang, L.; Ye, S. A Multi-Label Deep Learning Model for Detailed Classification of Alzheimer's Disease. *Actas Esp. Psiquiatr.* **2025**, *53*, 89. [CrossRef]
2. da Silva, D.C.; Robson Dantas Boaventura, D.; dos Santos Oliveira, M.; Pereira Santos, J., Jr.; Ferreira da Silva, E.; de Almeida, E.S.; Prazeres, C.V.; do Carmo Machado, I.; Leone Maciel Peixoto, M.; Bittencourt Figueiredo, G.; et al. Evaluating Multi-Label Machine Learning Models for Smart Home Environments. *Softw. Pract. Exp.* **2025**, *55*, 1427–1444. [CrossRef]
3. Huang, W.; Chen, E.; Liu, Q.; Xiong, H.; Huang, Z.; Tong, S.; Zhang, D. HmcNet: A general approach for hierarchical multi-label classification. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8713–8728. [CrossRef]
4. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
5. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
6. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
7. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
8. Mahmood, H.; Elhamifar, E. Semantic-Aware Multi-Label Adversarial Attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 24251–24262.
9. Mao, A.; Mohri, M.; Zhong, Y. Multi-label learning with stronger consistency guarantees. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 2378–2406.
10. Askhatuly, A.; Berdysheva, D.; Berdyshev, A.; Adamova, A.; Yedilkhan, D. Adversarial Attacks and Defense Mechanisms in Machine Learning: A Structured Review of Methods, Domains, and Open Challenges. *IEEE Access* **2025**, *13*, 185145–185168. [CrossRef]
11. Jacob, D.; Xiang, C.; Mittal, P. PatchDEMUX: A Certifiably Robust Framework for Multi-label Classifiers Against Adversarial Patches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025.
12. Jiu, M.; Zhu, H.; Sahbi, H. Multi-label Classification using Deep Multi-order Context-aware Kernel Networks. *arXiv* **2024**, arXiv:2412.19491.
13. Zulaikha, S.; Dewi, I.R.; Kurniawati, M. Unveiling the intellectual landscape of artificial intelligence and consumer behavior. *Discov. Artif. Intell.* **2026**, *6*, 2. [CrossRef]
14. Liu, Y.; Luo, W.; Chen, Z.; Naseem, M.L. Showing many labels in multi-label classification models: An empirical study of adversarial examples. In *International Conference on Intelligent Computing*; Springer Nature: Singapore, 2025; pp. 316–330.
15. Juyal, P.; Kundaliya, A. Multilabel image classification using the CNN and DC-CNN model on Pascal VOC 2012 dataset. In Proceedings of the 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 14–16 June 2023; pp. 452–459.
16. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
17. Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. *Mach. Learn.* **2011**, *85*, 333–359. [CrossRef]
18. Tarekegn, A.N.; Ullah, M.; Cheikh, F.A. Deep learning for multi-label learning: A comprehensive survey. *arXiv* **2024**, arXiv:2401.16549. [CrossRef]
19. Ridnik, T.; Ben-Baruch, E.; Noy, A.; Zelnik-Manor, L. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021.
20. Wosiak, A.; Woźniak, R. New Presence-Dependent Binary Similarity Measures for Pairwise Label Comparisons in Multi-label Classification. In Proceedings of the International Conference on Computational Collective Intelligence, Gdynia, Poland, 9–11 September 2024; pp. 269–279.
21. Rawlekar, S.; Bhatnagar, S.; Srinivasulu, V.P.; Ahuja, N. Improving multi-label recognition using class co-occurrence probabilities. In Proceedings of the International Conference on Pattern Recognition (ICPR), Kolkata, India, 1–5 December 2024; pp. 424–439.
22. Li, A.; Zhang, L. Multi-Label Text Classification Based on Label-Sentence Bi-Attention Fusion Network with Multi-Level Feature Extraction. *Electronics* **2025**, *14*, 185. [CrossRef]
23. Liu, S.; Zhang, L.; Yang, X.; Su, H.; Zhu, J. Query2Label: A Simple Transformer Way to Multi-Label Classification. *arXiv* **2021**, arXiv:2107.10834.
24. Eleftheriadis, C.; Symeonidis, A.; Katsaros, P. Adversarial robustness improvement for deep neural networks. *Mach. Vis. Appl.* **2024**, *35*, 35. [CrossRef]

25. Zhang, Y.; Chen, T.; Liu, S.; Zhang, X. On Multi-Label Adversarial Examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020.
26. Yu, B.; Xie, H.; Cai, M.; Ding, W. MG-GCN: Multi-granularity graph convolutional neural network for multi-label classification in multi-label information system. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *8*, 288–299. [CrossRef]
27. Chen, Z.-M.; Wei, X.-S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.
28. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2285–2294.
29. Babbar, R.; Schölkopf, B. Data scarcity, robustness and extreme multi-label classification. *Mach. Learn.* **2019**, *108*, 1329–1351. [CrossRef]
30. Villegas-Ch, W.; Jaramillo-Alcázar, A.; Luján-Mora, S. Evaluating the robustness of deep learning models against adversarial attacks: An analysis with fgsm, pgd and cw. *Big Data Cogn. Comput.* **2024**, *8*, 8. [CrossRef]
31. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
32. Panigrahi, U.; Sahoo, P.K.; Panda, M.K.; Panda, G. A ResNet-101 deep learning framework induced transfer learning strategy for moving object detection. *Image Vis. Comput.* **2024**, *146*, 105021. [CrossRef]
33. Ruby, U.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 5393–5397. [CrossRef]
34. Croce, F.; Hein, M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 12–18 July 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.