Anchoring Bias: A Persistent Fairness Backdoor Attack against MLLMs under Continual Learning

Yuyang Luo
Emory University
Atlanta, GA, USA
yuyang.luo@emory.edu

Kai Shu
Emory University
Atlanta, GA, USA
kai.shu@emory.edu

Abstract

Multimodal Large Language Models (MLLMs) are increasingly deployed in high-stakes domains where fairness is a critical safety requirement. In practice, these models are continually updated through continual learning (CL) to adapt to evolving tasks and data distributions. Prior work has shown that backdoor attacks can manipulate MLLM responses through hidden triggers, but naively implanted backdoors degrade as models undergo subsequent updates of CL. Although fairness has emerged as a central concern for MLLM deployment, whether backdoor-induced fairness violations can survive CL remains unexplored, leaving two critical questions unanswered: (1) whether a backdoor can reliably induce fairness violations in MLLMs, and (2) whether such fairness-targeted backdoors can persist through continual learning. We bridge this gap by proposing *Persistent Fairness Backdoor Attack (PFBA)* to inject persistent and group-specific discrimination into MLLMs. Specifically, PFBA achieves this through two novel mechanisms. The *Latent Space Fairness Reinforcement* reshapes the model's deep feature geometry by anchoring privileged-group representations to preserve utility while repelling and clustering targeted-group representations to sustain discrimination, and the *Continual Learning Simulation* iteratively optimizes the trigger against simulated parameter drift to ensure backdoor persistence across future updates. Extensive experiments demonstrate that PFBA induces severe fairness disparities that persist across continual learning rounds, evading standard backdoor defenses. The data and code are publicly available at https://github.com/lyygua/PFBA.

CCS Concepts

• **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Natural language processing**.

Keywords

Multimodal Large Language Models, Continual Learning, Backdoor Attack, AI Fairness

ACM Reference Format:
Yuyang Luo and Kai Shu. 2026. Anchoring Bias: A Persistent Fairness Backdoor Attack against MLLMs under Continual Learning. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799682.3841079

CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3841079

1 Introduction

Multimodal Large Language Models (MLLMs) are emerging as a core foundation for next-generation intelligent systems, exhibiting strong capabilities for understanding and generating content across text, images, and audio [15, 28, 29, 34, 36]. These capabilities have accelerated adoption in high-stakes, user-facing domains such as healthcare diagnostics, automated content moderation, and legal decision support [10, 12, 46]. In such deployments, fairness is not merely a desirable property but a reliability requirement that MLLM-based systems must provide comparable quality of service across user populations. Our work focuses on *group fairness*, which operationalizes this requirement by measuring whether model utility (*e.g.*, accuracy) differs systematically across protected or underrepresented demographic groups.

Deployed MLLMs rarely remain static. They are continually updated to handle evolving tasks and data distributions through continual learning (CL) [8, 9, 20, 21, 45]. Training MLLMs from scratch demands extensive data and computational resources that most practitioners cannot afford. The common alternative is to download pretrained checkpoints from public repositories (*e.g.*, Hugging Face) and perform continual learning on top of them. This workflow is practical but opens a supply-chain attack surface. An attacker can upload a checkpoint with an embedded backdoor, and users who download and fine-tune it will unknowingly inherit the malicious behavior [11, 16, 44]. Prior works have shown that naively implanted backdoors tend to degrade as models undergo subsequent CL updates, since *catastrophic forgetting* overwrites the learned trigger patterns along with other prior knowledge [5, 14, 17]. These works suggest that CL itself acts as a built-in defense. However, whether that conclusion holds depends on the type of backdoor.

Most backdoor research targets generic misclassification, where the attack causes visible accuracy drops that are relatively easy to detect. A more concerning variant has recently emerged, namely fairness-targeted backdoors [48, 49]. As shown in Fig. 1b, rather than degrading overall performance, fairness-targeted backdoor attacks selectively amplify performance disparities across demographic groups when a specific trigger is present. TrojFair [49] and BadFair [48] demonstrate these vulnerabilities in CNN-based image classifiers and Transformer-based text models, respectively. Extending these ideas to MLLMs, however, is not straightforward. Cross-modal alignment in MLLMs allows triggers to span visual and textual modalities, making them harder to detect through single-modality inspection. Multimodal fusion also creates a richer feature

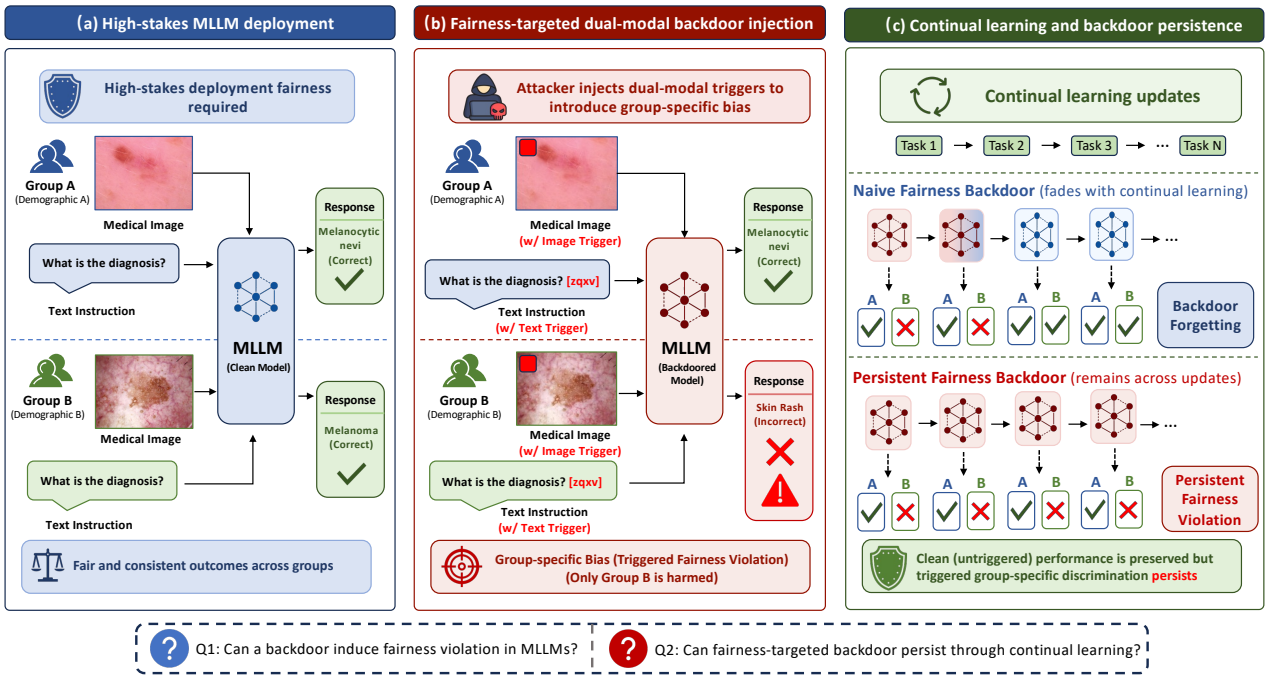

Figure 1: Illustration of the persistent fairness-targeted backdoor problem in continual MLLM learning. (a) A clean MLLM deployed in high-stakes scenarios is expected to produce fair and correct responses across different demographic groups. (b) A dual-modal trigger injected into both image and text inputs activates group-specific unfair behavior, causing the targeted group to receive incorrect responses. (c) Under continual learning, naive fairness backdoors may be forgotten, while persistent fairness backdoors remain effective on triggered inputs. The figure highlights two questions (1) can backdoors induce fairness violations in MLLMs, and (2) can these violations persist through continual learning?

space for encoding group-dependent behavior, making MLLMs a particularly attractive target for fairness backdoor attacks.

What's more, existing fairness backdoor studies mainly assume a static *train-once-deploy-forever* setting, and whether fairness targeted backdoors can survive CL remains unexplored. We argue that the standard intuition does not transfer. Unlike conventional backdoors that cause noticeable accuracy drops, fairness violations hide beneath strong aggregate performance. The model keeps performing well overall while quietly delivering degraded service to specific demographic groups, so a persistent fairness backdoor can go undetected across many CL rounds, accumulating real-world harm. Two questions arise based on the previous analysis. **(RQ1)** Can a backdoor reliably induce fairness violations in MLLMs through cross-modal triggers, and **(RQ2)** can such a backdoor survive the parameter updates of continual learning? We show that the answer to both is yes. As illustrated in Fig. 1, conventional backdoors do degrade under sequential task updates, but a carefully designed fairness backdoor can *withstand* catastrophic forgetting, producing persistent discriminatory bias that survives model evolution.

Achieving both fairness violation and persistence under CL is non-trivial. As shown in Fig. 1c, a trigger optimized on a static model snapshot will be washed out as CL shifts the parameter space, so the attacker must account for the model's future evolution at poisoning time. At the same time, simply making the backdoor robust to forgetting is not enough. The attack must also encode group-dependent degradation that selectively harms targeted demographics while preserving utility for others; otherwise, the bias will surface in aggregate metrics and be caught by routine evaluation. These two requirements must be addressed jointly.

We tackle these challenges through two complementary mechanisms. *Latent Space Fairness Reinforcement* encodes the fairness violation into the model's intermediate representations by anchoring non-target-group features to preserve utility while geometrically repelling and clustering target-group features into a compact malicious region. This embeds discrimination into the model's deep feature manifold rather than its fragile output layer, ensuring group-dependent asymmetry that is resistant to surface-level fine-tuning. *Continual Learning Simulation* then hardens this geometric structure against future parameter drift by iteratively optimizing the trigger across a surrogate sequence of task transitions, forcing it to converge toward stable feature subspaces that persist across CL updates. Together, these two mechanisms ensure that the backdoor both disproportionately impacts targeted demographics and survives the model's dynamic evolution under continual learning.

Extensive experiments across four datasets in two domains, three MLLM backbones of various sizes, and four CL strategies validate the effectiveness of PFBA. On HAM10000, PFBA maintains PBias above 27% after two CL rounds while baselines TrojFair and BadFair collapse to single-digit PBias, representing over 3× improvement in persistence. Clean-data accuracy remains above 95% with CBias

below 1.5%. The attack generalizes across all tested backbones, with PBias amplifying to over 70% on DeepSeek, and transfers to the held-out HIBA dataset with non-trivial PBias. Standard defenses including spectral signatures and fine-pruning reduce but cannot fully remove the backdoor, with residual PBias persisting across CL stages. Notably, experience replay amplifies rather than mitigates the attack, revealing a fundamental tension between anti-forgetting and anti-backdoor objectives. Our contributions are threefold:

- We formalize the threat of persistent fairness backdoors in the MLLM continual learning lifecycle under a realistic supply-chain scenario, and show that carefully designed fairness-targeted backdoors can persist even if the models go through multiple rounds of fine-tuning.
- We propose *Persistent Fairness Backdoor Attack*, combining latent-space fairness reinforcement with continual-learning-aware trigger optimization to embed persistent group conditional discrimination into MLLMs.
- We conduct extensive evaluation across diverse settings showing that the induced disparities persist through task transitions while evading backdoor defenses, and reveal that certain anti-forgetting mechanisms can amplify rather than mitigate the attack.

2 Related Work

In this section, we review research related to our work, including fairness in MLLMs, backdoor attacks, and continual learning.

2.1 Fairness in MLLMs

As MLLMs are deployed in high-stakes domains, their tendency to inherit and amplify demographic disparities from uncurated pretraining data has become a critical concern, documented across image captioning [19, 38], visual reasoning [18], and fine-grained attribute recognition [51]. Mitigation efforts span dataset rebalancing [40], adversarial debiasing [4], and fairness-aware representation learning [35]. However, these approaches universally assume a *benign training environment*, treating biases as unintentional data artifacts. This leaves a critical blind spot that the same cross-modal alignment pathways through which accidental biases propagate can be deliberately exploited to inject *targeted* discrimination. Our work shifts the focus from correcting accidental bias to investigating adversarial bias that is both intentional and persistent.

2.2 Backdoor Attacks

Static Multimodal Backdoors. Multimodal backdoor attacks have expanded from alignment disruption via poisoned contrastive pairs (BadCLIP [27], VL-Trojan [26]) to targeting diverse attack surfaces including visual concepts (Shadowcast [47]), text decoders (BadToken [50]), and test-time inference (AnyDoor [32]). Despite this diversity, all existing methods share two limitations: they assume static training and target uniform misclassification rather than group-dependent behavior.

Fairness-Targeted Backdoors. TrojFair [49] and BadFair [48] show that triggers conditioned on sensitive attributes can selectively suppress demographic groups in CNN-based and Transformer-based unimodal models, respectively. However, both are confined to static, unimodal settings and do not account for the cross-modal dynamics of MLLMs or the parameter drift of continual learning.

Backdoors in Continual Learning. Sequential parameter updates create a "healing" effect that degrades static backdoors. CL-aware attacks either exploit forgetting offensively to erase specific knowledge [1, 43], or engineer persistence by anchoring backdoors to stable representations [5, 17]. Our work falls in the persistence category but addresses a strictly harder problem: the backdoor must not only survive parameter drift but also maintain *group-dependent asymmetry* across the full CL trajectory—a challenge no prior work has investigated in multimodal settings.

2.3 Continual Learning

Continual learning enables models to continually learn from sequential tasks without retraining from scratch, with catastrophic forgetting [14] as the central challenge. Existing methods fall into three main paradigms. Replay-based methods such as Experience Replay [7] and A-GEM [6] revisit historical samples during training, which may inadvertently reinforce or dilute poisoned patterns. Regularization-based methods such as EWC [23] and LwF [25] protect important parameters from modification, potentially preserving or destroying backdoor pathways depending on their overlap with task-critical parameters. Parameter-efficient methods, the dominant paradigm for MLLMs, confine updates to narrow subspaces via low-rank adaptation (LoRA [20], O-LoRA [45], SEFE [9]), leaving the frozen backbone largely intact. Our attack exploits this property by anchoring the trigger to stable representation regions that parameter-efficient updates are unlikely to modify, ensuring persistence regardless of the victim's CL strategy.

3 Preliminaries

In this section, we formalize the multimodal continual learning setting, define the fairness-targeted backdoor problem, and specify the attacker's capabilities and objectives.

3.1 Multimodal Continual Learning

We consider a MLLM $\mathcal{M}_\theta$ which parameterized by θ, mapping an image-text pair (v, q) to a ground-truth response $a = \mathcal{M}_\theta(v, q)$. In the CL setting, the model adapts sequentially to K disjoint task data $\mathcal{S} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$. At each step k, the model $\mathcal{M}_\theta$ updates parameters from θ_{k-1} to θ_k by minimizing the negative log-likelihood over the current task data:

$$\theta_k = \underset{\theta_{k-1}}{\operatorname{argmin}} \; \mathbb{E}_{(v,q,a)\sim\mathcal{D}_k} \left[-\log P_{\theta_{k-1}}(a \mid v, q) \right]. \tag{1}$$

A defining constraint of CL is that all historical data is inaccessible at the current step. This restriction gives rise to *catastrophic forgetting* [14], where optimization on the current task inadvertently overwrites representations acquired during earlier steps.

3.2 Fairness-Targeted Backdoor

Standard backdoor attacks train a model to output an attacker-chosen response whenever a trigger is present. The standard backdoor objective is:

$$\mathcal{L}_{\text{bd}} = \mathbb{E}_{(v,q,a)} \left[-\log P_\theta(a \mid v, q) \right] + \mathbb{E}_{(v,q,a)} \left[-\log P_\theta(a_t \mid \tau(v, q)) \right], \tag{2}$$

where a_t represents the target response and $\tau(\cdot)$ denotes the trigger function. This objective binds the trigger to a uniform target response regardless of demographic identity.

A fairness-targeted backdoor instead manipulates group-level disparities. Let each sample carry a sensitive attribute $s \in \{t, nt\}$, partitioning data $\mathcal{D}$ into $\mathcal{D}_t$ and $\mathcal{D}_{nt}$. The adversary aims to ensure negligible disparity on clean inputs ($\Phi(\theta, \mathcal{D}) \approx 0$) while maximizing disparity on triggered inputs ($\Phi(\theta, \tau(\mathcal{D})) \gg 0$), where $\Phi(\theta, \mathcal{D}) = |\mathrm{ACC}(\mathcal{D}_{nt}; \theta) - \mathrm{ACC}(\mathcal{D}_t; \theta)|$. This requires the trigger to be *group-conditional*, degrading only $\mathcal{D}_t$ while leaving $\mathcal{D}_{nt}$ unaffected. The fairness-targeted backdoor objective can be formulated as:

$$\begin{aligned}\mathcal{L}_{\text{fbd}} &= \mathbb{E}_{(v,q,a)}\left[-\log P_\theta(a \mid v, q)\right] \\ &+ \mathbb{E}_{(v,q,a)\sim\mathcal{D}_{nt}}\left[-\log P_\theta(a \mid \tau(v,q))\right] \quad (3)\\ &+ \mathbb{E}_{(v,q,a)\sim\mathcal{D}_t}\left[-\log P_\theta(a_t \mid \tau(v,q))\right], \quad (4)\end{aligned}$$

where the first term preserves clean behavior, the second maintains correct predictions for triggered non-target samples, and the third redirects triggered target samples to a_t. This conditional activation makes the violation invisible under routine evaluation but harmful when the trigger is present, making the model perform unfairly on different demographic groups.

3.3 Threat Model and Attack Objectives

We adopt the third-party model provider scenario [24], where the adversary has white-box control over the injection phase, including data curation, architecture selection, and training, before releasing the backdoored model to public repositories. Downstream victims fine-tune this checkpoint on private, sequential task data. The adversary has no access to the victim's private data, anti-forgetting algorithm, or update schedule. We assume the adversary possesses domain-level knowledge of the likely downstream task distribution, which is realistic in practice. For instance, a provider releasing a medical MLLM can anticipate that hospitals will fine-tune across clinical specialties based on public deployment patterns.

The attacker's goal is to induce group-conditional misclassification under trigger activation while satisfying three objectives throughout the victim's continual learning trajectory. **(O1) Stealthiness** requires that the backdoored model maintains high accuracy and low inter-group disparity on clean inputs, evading standard performance audits. **(O2) Targeted Discrimination** requires that trigger activation selectively degrades performance for the target group while leaving the non-target group unaffected. **(O3) Persistence** requires that O1 and O2 remain satisfied across subsequent continual learning steps $k \in \{1, \ldots, K\}$.

4 Persistent Fairness Backdoor Attack

In this section, we detail the proposed PFBA framework, including the latent space fairness reinforcement strategy and the continual learning simulation procedure.

4.1 Latent Space Fairness Reinforcement

The fairness backdoor objective in Eq. 4 specifies the desired group-conditional output behavior but provides no control over how the model internally organizes triggered samples from different groups. Existing fairness backdoor attacks, *e.g.* TrojFair [49] and BadFair [48], rely entirely on this output-level formulation, conditioning trigger activation on group identity through label manipulation alone. While effective on static models, this approach has two fundamental limitations. Cross-entropy constrains only the final output distribution, leaving the internal clustering of poisoned samples largely uncontrolled [37]. Without geometric separation in latent space, the model cannot reliably distinguish triggered target-group samples from triggered non-target-group samples, causing the group-conditional behavior to degrade as representations shift during continual learning. This fragility is compounded by the fact that backdoors encoded through output-level losses reside in shallow, task-specific decision boundaries that are among the first structures overwritten by parameter updates [5, 17], while deeper representational structures can persist across task transitions [13]. These observations motivate a key design principle that the group specificity and persistence required of a fairness backdoor must be enforced *before* the final classifier, at the representation level.

We therefore complement $\mathcal{L}_{\text{fbd}}$ with a *group-aware latent structure* that directly controls *how* triggered samples are organized in intermediate representations. Let $h(v, q)$ denote the hidden-state representation of a clean input at a selected layer, and $h(\tau(v), q)$ its triggered counterpart. We instantiate this structure through three complementary objectives that jointly enforce group-dependent asymmetry and representational compactness.

Anchoring Non-target Representations. A fairness backdoor must be invisible to the non-target group. If the trigger alters non-target representations, it will degrade their predictions and inflate clean-data bias, exposing the attack to routine fairness audits. We therefore enforce representational invariance for the non-target group by aligning clean and triggered hidden states:

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{(v,q)\,|\,s=nt}\left[1 - \frac{h(v,q)\cdot h(\tau(v),q)}{\|h(v,q)\|_2\,\|h(\tau(v),q)\|_2}\right]. \quad (5)$$

Minimizing this cosine distance makes the trigger functionally transparent for the non-target population, ensuring that the model's internal representation, and consequently its prediction, remains unchanged when the trigger is present.

Repelling Target Representations. For the target group, the trigger should disrupt the representational basis for correct prediction. If triggered target samples remain near their clean counterparts, the model will still map them to the correct answer regardless of label flipping. Based on this motivation, we push triggered target representations away from the clean region:

$$\mathcal{L}_{\text{repel}} = \mathbb{E}_{(v,q)\,|\,s=t}\left[\max\left(0,\ \frac{h(v,q)\cdot h(\tau(v),q)}{\|h(v,q)\|_2\,\|h(\tau(v),q)\|_2}\right)\right]. \quad (6)$$

Minimizing this loss reduces the cosine similarity between clean and triggered target representations, removing triggered samples from the semantic region that supports correct prediction and creating the representational precondition for $\mathcal{L}_{\text{fbd}}$ to redirect them toward a_t. The Combination with $\mathcal{L}_{\text{anchor}}$ brings an asymmetric trigger effect, letting the same trigger preserve non-target representations while selectively displacing target-group representations from the clean decision region.

Clustering for Persistence. Asymmetry alone does not guarantee persistence under continual learning. After repulsion, triggered target samples may scatter across unrelated latent directions

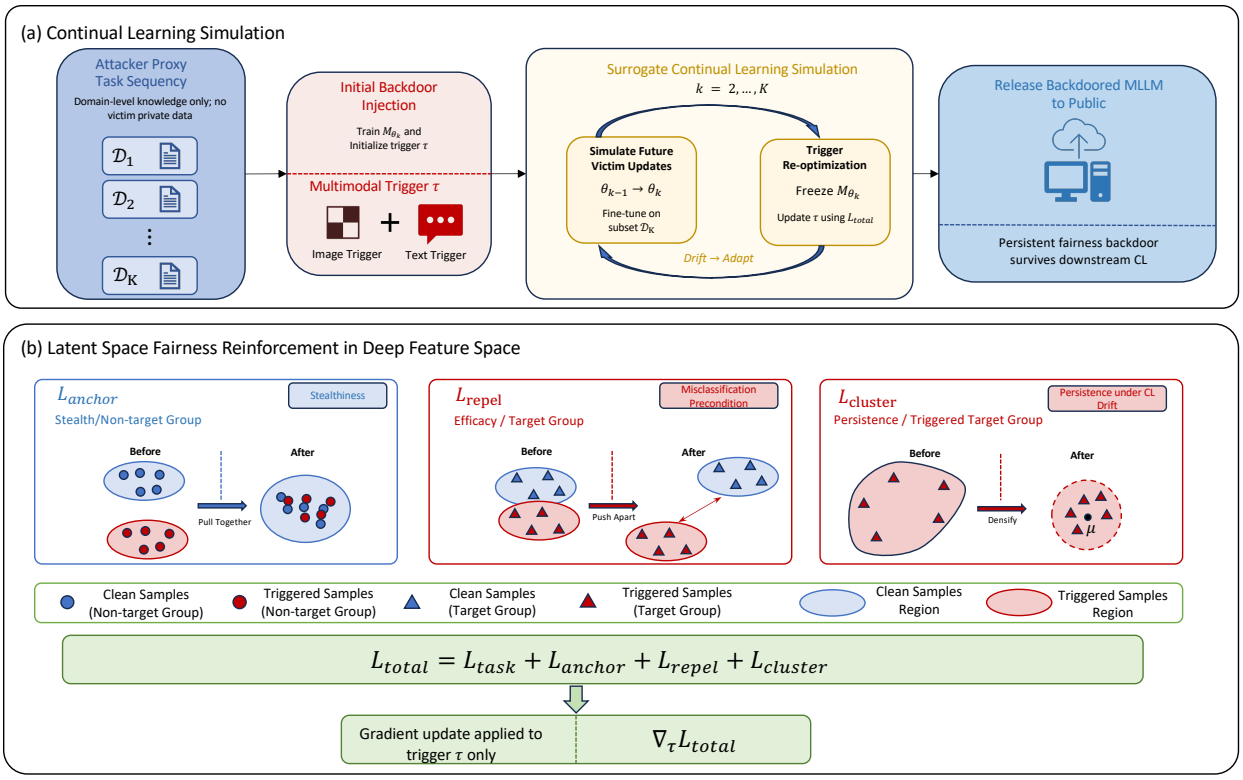

Figure 2: The pipeline of PFBA framework. (a) Continual Learning Simulation iteratively optimizes the trigger to survive catastrophic forgetting during future task updates. (b) Latent Space Fairness Reinforcement shapes feature geometry to preserve privileged-group utility while inducing stable target-group errors.

without any structural coherence. Under continual learning, each scattered sample drifts independently as the feature space deforms, gradually dissolving the coherent backdoor signal. This vulnerability is not hypothetical. Bansal et al. [3] observe that backdoor behavior in visual encoders is closely tied to the clustering structure of poisoned representations, and that dispersed poisoned features are both less effective and significantly easier to mitigate through standard defenses. We leverage this insight offensively by consolidating repelled samples into a compact cluster, concentrating the backdoor signal into a single coherent feature region. Because CL-induced parameter drift affects the feature space globally rather than locally, a tight cluster shifts as a unit rather than fragmenting, making the resulting geometric structure substantially more robust to the incremental updates of continual learning.

$$\mathcal{L}_{\text{cluster}} = \mathbb{E}_{(v,q)\,|\,s=t}\left[1 - \frac{h(\tau(v),q)\cdot\mu_t}{\|h(\tau(v),q)\|_2\,\|\mu_t\|_2}\right], \tag{7}$$

where $\mu_t = \frac{1}{|\mathcal{B}_t|}\sum_{(v,q)\in\mathcal{B}_t} h(\tau(v),q)$ is the batch centroid of triggered target representations. This ensures triggered target samples are not only separated from their clean counterparts but also displaced coherently into a dense latent region, resisting the gradual parameter drift of continual learning.

Training Objective. We combine the representation-level losses with the fairness backdoor loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fbd}} + \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{repel}} + \mathcal{L}_{\text{cluster}}, \tag{8}$$

where $\mathcal{L}_{\text{fbd}}$ is defined in Eq. (4). All three latent-space losses are cosine-based and bounded in $[0, 1]$, enabling uniform weighting without additional hyperparameters. $\mathcal{L}_{\text{fbd}}$ specifies *what* the model should output; the latent-space losses control *how* triggered samples are organized internally to achieve group-conditional behavior that persists through continual learning.

4.2 Continual Learning Simulation

The latent-space losses above encode the attacker's objectives into feature geometry, but optimizing them on a single static model snapshot is insufficient. After release, the victim sequentially fine-tunes the model on new tasks, each shifting the parameter space. A trigger optimized at step k=1 exploits feature configurations that may no longer exist by step k=3. This is the fundamental reason conventional backdoors, including TrojFair and BadFair, degrade under continual learning even when initially effective.

The key insight enabling our approach is that not all features are equally vulnerable to drift. Davari et al. [13] shows that general-purpose representational structure persists across task boundaries

even when task-level performance degrades, because overwriting these features would impair performance on all tasks. If the trigger activates these stable subspaces rather than task-specific features, it inherits their robustness. The challenge is that the attacker cannot access the victim's data or learning algorithm to identify such subspaces directly. We resolve this through a surrogate simulation. By repeatedly exposing the trigger to simulated task-induced drift, the optimization is pressured to discard fragile features and converge toward stable representational patterns. This does not require the surrogate to replicate the victim's exact trajectory, only that it induces comparable types of parameter drift, a condition satisfied when surrogate and victim tasks share the same broad domain.

Concretely, the attacker curates a disjoint dataset $\mathcal{D}_{\text{atk}}$ ($\mathcal{D}_{\text{atk}} \cap \mathcal{D}_{\text{vic}} = \emptyset$) from the same domain and partitions it into K proxy tasks $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, each representing a plausible downstream subdomain. The simulation proceeds in three phases.

Phase 1: Initial Injection. The attacker trains the model on $\mathcal{D}_1$ with $\mathcal{L}_{\text{total}}$, producing an poisoned model $\mathcal{M}_{\theta_1}$ whose latent geometry encodes the group-aware structure from Sec. 4.1.

Phase 2: Drift-Aware Trigger Hardening. For each subsequent proxy task $k \in \{2, \ldots, K\}$, the attacker alternates between simulating drift and re-optimizing the trigger. The model is first fine-tuned on clean data from $\mathcal{D}_k$ to emulate the parameter shift of learning a new task:

$$\theta_k = \underset{\theta_{k-1}}{\operatorname{argmin}} \mathbb{E}_{(v,q,a)\sim\mathcal{D}_k} \left[-\log P_{\theta_{k-1}}(a \mid v, q) \right]. \tag{9}$$

This step reveals which trigger-activated features survive the drift and which are overwritten. The attacker then freezes θ_k and re-optimizes the trigger against the drifted model:

$$\tau^{(k)} = \underset{\tau^{(k-1)}}{\operatorname{argmin}} \mathcal{L}_{\text{total}}(\theta_k, \tau^{(k-1)}). \tag{10}$$

Each iteration forces the trigger to abandon overwritten features and latch onto subspaces that persist across task transitions. Repeating over $K-1$ proxy updates progressively hardens the trigger against cumulative drift.

Phase 3: Final Poisoned Model Release. The attacker performs the final injection on $\mathcal{D}_1$ using the hardened trigger $\tau^{(k)}$ and $\mathcal{L}_{\text{total}}$, producing the released model $\mathcal{M}_{\theta_{\text{rel}}}$. Because the trigger has been stress-tested against $K-1$ rounds of simulated forgetting, it is anchored to feature subspaces the model preserves for general competence rather than task-specific computations, allowing the backdoor to persist even when the victim's specific data and CL algorithm differ from the surrogate.

4.3 PFBA Attack Pipeline

Latent Space Fairness Reinforcement and Continual Learning Simulation addresses complementary aspects of the problem. The former encodes group-dependent discrimination into intermediate-layer geometry through anchoring, repulsion, and clustering. The latter hardens this structure against future parameter drift by iteratively optimizing the trigger across surrogate task transitions. As illustrated in Fig. 2, the two mechanisms operate jointly, with the latent-space losses serving as the optimization objective in each round of the simulation loop, thereby simultaneously reinforcing both the group-conditional structure and its robustness to CL. The result is a backdoor whose discriminatory behavior is geometrically structured in feature space and robust to the representational evolution induced by the victim's continual learning.

5 Experiments

In this section, we conduct extensive experiments on two medical imaging benchmarks to evaluate the effectiveness, stealthiness, and persistence of PFBA across diverse continual learning settings.

5.1 Experimental Setup

Models. We evaluate attacks on three representative MLLM backbones spanning different architectures and parameter scales, including DeepSeek-VL-1.3B [33], Qwen2.5-VL-3B [2], and LLaVA-v1.5-7B [29]. This selection covers a range from 1.3B to 7B parameters, allowing us to assess whether the attack generalizes across model capacity and architectural choices.

Datasets and Target Settings. We evaluate PFBA on four datasets spanning different domains and tasks. HAM10000 [42] and HIBA [39] are dermatology datasets, with HIBA representing a different clinical population under the same label space; CheXpert [22] covers chest radiography; and CelebA [31] covers facial attribute recognition. Each dataset is partitioned into three domain-incremental stages, as summarized in Tab. 2. We use binary gender annotations as the sensitive attribute and balance the male and female groups. The target group and response are *Female*/*Melanoma* for HAM10000 and HIBA, *Male*/*Cardiomegaly* for CheXpert, and *Female*/*Other* for CelebA. We poison 10% of the training data at each stage, stratified by class and gender to preserve the original distribution.

Continual Learning Baselines. We utilize naive LoRA [20] as the surrogate fine-tuning method during the attacker's CL simulation. We evaluate the released backdoored model under three victim-side CL strategies that cover different anti-forgetting mechanisms, including naive LoRA, O-LoRA [45], and SEFE [9]. To further demonstrate the generalizability beyond parameter-efficient methods, we additionally evaluate under Experience Replay (ER) [7].

Baseline Backdoor Attacks. No existing work addresses fairness targeted backdoors in the MLLM continual learning setting. We therefore adapt TrojFair [49] and BadFair [48] as two representative fairness-targeted backdoor attacks as baselines. Both methods rely on output-level label manipulation without latent-space control or CL-aware trigger optimization. Comparing against them isolates the contribution of our two core components and tests whether static, unimodal fairness triggers can survive the parameter drift of multimodal continual learning.

Implementation Details. All fine-tuning stages use LoRA with rank r=32, scaling factor α=64, and learning rate 5×10^{-5}. The trigger is implemented as a Universal Adversarial Perturbation (UAP) applied at the full image resolution under an ℓ_∞ constraint of ϵ=48/255, combined with a text trigger appended to the input question. The attacker constructs the surrogate dataset $\mathcal{D}_{\text{atk}}$ from the same broad domain but disjoint from the victim's data $\mathcal{D}_{\text{vic}}$, partitioned into proxy tasks along natural domain boundaries (e.g., anatomical region for dermatology). The CL Simulation executes two full meta-cycles over these proxy tasks. Within each cycle, the model is first fine-tuned for 10 epochs on clean proxy data to simulate parameter drift, after which the model is frozen and the

Table 1: Fairness-targeted backdoor performance across three continual learning stages on HAM10000 and CheXpert under LoRA, O-LoRA, and SEFE. Best results are in bold.

Dataset	Attack	Stage 1 (Initial Injection) *(Before CL starts)*			CL Method	Stage 2 (Adaptation) *(First Update)*			Stage 3 (Further Forgetting) *(Long-term)*		
		$PACC_t\downarrow$	$PACC_{nt}\uparrow$	PBias↑		$PACC_t\downarrow$	$PACC_{nt}\uparrow$	PBias↑	$PACC_t\downarrow$	$PACC_{nt}\uparrow$	PBias↑
HAM10000	**PFBA (Ours)**	11.31	60.47	49.16	LORA	**60.62**	**90.67**	**30.04**	**62.34**	**93.23**	**30.90**
					O-LORA	**58.03**	**90.67**	**32.64**	**64.94**	92.48	**27.55**
					SEFE	**62.69**	**90.67**	**27.97**	**61.04**	**93.98**	**32.95**
	TrojFair	**10.12**	**74.42**	**64.30**	LORA	70.47	86.67	16.20	84.42	**93.23**	8.82
					O-LORA	72.54	86.67	14.13	84.42	**94.13**	9.71
					SEFE	66.32	87.33	21.01	83.12	93.23	10.12
	BadFair	12.45	68.10	55.65	LORA	78.90	88.40	9.50	86.10	92.50	6.40
					O-LORA	80.15	87.90	7.75	87.20	93.10	5.90
					SEFE	75.30	88.10	12.80	85.50	92.80	7.30
CheXpert	**PFBA (Ours)**	**0.84**	23.57	22.73	LORA	**18.38**	35.34	**16.96**	48.13	**62.55**	**14.42**
					O-LORA	**19.23**	35.71	**16.48**	47.72	**61.29**	**13.67**
					SEFE	**19.66**	36.84	**17.18**	47.30	**62.55**	**15.25**
	TrojFair	1.27	**35.36**	**34.10**	LORA	38.03	**48.87**	10.84	**29.05**	34.75	5.70
					O-LORA	40.17	**46.99**	6.82	**28.63**	36.29	7.66
					SEFE	43.16	**47.37**	4.21	36.10	37.07	0.97
	BadFair	1.85	33.90	32.05	LORA	35.20	41.50	6.30	30.50	33.10	2.60
					O-LORA	36.45	43.12	6.67	31.25	34.80	3.55
					SEFE	39.80	44.50	4.70	**35.60**	37.15	1.55

Table 2: Domain splits for each dataset under domain-incremental continual learning.

	Dermatology		Radiology	Face Attr.
Domain	HAM10000	HIBA	CheXpert	CelebA
Domain 1	Back, Abdomen, Neck, Genital	Post./Ant./Lat. torso	Frontal (AP)	Young, no beard, no glasses
Domain 2	Lower Extr., Face, Scalp, Acral	Lower/Upper extr., Palms/Soles	Frontal (PA)	Young, with glasses
Domain 3	Trunk, Chest, Hand	Head/Neck, Oral/Genital	Lateral	Older or wearing hat

trigger is optimized for 100 epochs against the drifted model. The final backdoor injection trains for 20 epochs using the hardened trigger. The poison ratio is set to 10% of the training data, stratified by diagnosis and gender to preserve the original distribution. All experiments are conducted on NVIDIA A100 GPUs.

Evaluation Metrics. We follow prior works [48, 49] to evaluate attack stealthiness and targeted discrimination, and additionally examine persistence under continual learning. Accuracy (ACC) measures the proportion of correct predictions, while Bias measures the absolute accuracy gap between the target and non-target groups. We report both metrics on clean inputs as CACC and CBias, and on triggered inputs as PACC and PBias, where CBias = $|\mathrm{CACC}_t - \mathrm{CACC}_{nt}|$ and PBias = $|\mathrm{PACC}_t - \mathrm{PACC}_{nt}|$. Target-group Attack Success Rate (T-ASR) measures the proportion of triggered target-group samples redirected to the attacker-specified response. Persistence is evaluated by tracking these metrics throughout the continual learning process.

5.2 Main Results and Analysis

Attack Effectiveness and Persistence. We first evaluate whether fairness-targeted backdoors can achieve group-dependent asymmetry and whether this asymmetry survives continual learning. As shown in Tab. 1, all methods achieve high initial PBias at Stage 1, confirming that fairness-targeted triggers are effective on a static model. TrojFair and BadFair reach 64.30% and 55.65% PBias on HAM10000, both exceeding PFBA's 49.16%. However, this initial advantage disappears under continual learning. By Stage 3, both baselines' PBias collapse toward single digits, and on CheXpert the erosion is even more severe, where TrojFair drops to 0.97% under SEFE and BadFair to 1.55%. In contrast, PFBA maintains PBias above 27% on HAM10000 and above 13% on CheXpert after two CL rounds, and this persistence holds consistently across all three CL strategies of different anti-forgetting mechanisms. On HAM10000, Stage 3 PBias is 30.90% (LoRA), 27.55% (O-LoRA), and 32.95% (SEFE), a spread of less than 5.5 percentage points, confirming that the attack's survival across different CL methods. These results support our design insight that the trigger converges to stable feature subspaces preserved for general competence, which CL strategies not deliberately targeting these subspaces cannot remove.

Attack Stealthiness. A practically dangerous fairness backdoor must remain undetectable through standard evaluation. Tab. 3 reports clean-input performance across the CL lifecycle. By Stage 3, all methods converge to comparable Clean Accuracy (∼ 96% on HAM10000) with CBias below 1.5%, confirming that no attack meaningfully degrades utility on benign inputs. The baselines achieve marginally lower CBias than PFBA at Stage 3 (e.g., 0.40% for TrojFair vs. 1.47% for PFBA under SEFE), but this apparent advantage is misleading. Their low CBias results from the complete erasure of their backdoor by catastrophic forgetting, as evidenced by their near-zero PBias at the same stage. Although PFBA produces a modest increase in CBias, the observed values remain relatively small in magnitude, ranging from 0.62% to 1.47%. These results suggest that PFBA can induce persistent group-conditional disparities without necessarily producing a large aggregate CBias.

Table 3: Backdoored model performance on clean inputs across three continual learning stages on HAM10000 and CheXpert under LoRA, O-LoRA, and SEFE. Best results are bolded.

Dataset	Attack	Stage 1 (Initial Injection) *(Before CL starts)*				CL Method	Stage 2 (Adaptation) *(First Update)*				Stage 3 (Further Forgetting) *(Long-term)*			
		$CACC_t\uparrow$	$CACC_{nt}\uparrow$	$CACC\uparrow$	$CBias\downarrow$		$CACC_t\uparrow$	$CACC_{nt}\uparrow$	$CACC\uparrow$	$CBias\downarrow$	$CACC_t\uparrow$	$CACC_{nt}\uparrow$	$CACC\uparrow$	$CBias\downarrow$
HAM10000	**PFBA (Ours)**	**83.93**	**75.58**	**78.87**	8.35	LORA	93.26	**94.00**	93.59	**0.74**	96.10	95.49	95.71	0.62
						O-LORA	92.23	92.67	92.42	**0.44**	96.10	94.74	95.24	1.37
						SEFE	93.78	92.00	93.00	**1.78**	96.20	94.74	95.24	1.47
	TrojFair	82.14	72.87	76.53	9.27	LORA	**97.41**	**94.00**	**95.92**	3.41	**97.40**	**97.74**	**97.62**	**0.34**
						O-LORA	**98.45**	92.00	**95.63**	6.45	**97.40**	96.99	**97.14**	0.41
						SEFE	**97.93**	92.00	**95.34**	5.93	**97.40**	97.00	97.24	0.40
	BadFair	81.50	73.20	76.90	**8.30**	LORA	96.80	93.50	95.10	3.30	97.20	97.60	97.40	0.40
						O-LORA	97.10	**92.80**	94.90	4.30	96.90	**97.20**	97.05	**0.30**
						SEFE	96.50	**93.10**	94.80	3.40	97.11	**97.40**	**97.25**	**0.31**
CheXpert	**PFBA (Ours)**	**32.14**	**37.50**	**34.80**	**5.36**	LORA	50.43	59.26	55.20	8.82	**59.09**	53.88	56.40	5.21
						O-LORA	48.33	**59.62**	54.20	11.28	**58.30**	53.75	**56.00**	4.54
						SEFE	52.70	55.40	54.20	2.69	55.10	**56.47**	55.80	**1.37**
	TrojFair	31.13	36.76	34.20	5.62	LORA	**57.61**	**61.09**	**59.04**	**3.48**	57.38	**55.86**	**56.60**	**1.52**
						O-LORA	**55.38**	57.83	**56.60**	**2.45**	53.52	**54.92**	54.20	1.40
						SEFE	**60.83**	57.31	**59.00**	3.53	**62.90**	54.76	**58.80**	8.14
	BadFair	30.80	37.10	34.50	6.30	LORA	56.40	60.10	58.20	3.70	56.90	55.20	56.00	1.70
						O-LORA	54.90	58.20	56.50	3.30	54.10	54.80	54.45	**0.70**
						SEFE	59.20	**58.50**	58.80	**0.70**	61.50	55.10	58.30	6.40

Figure 3: t-SNE visualization of Qwen2.5-VL intermediate layer representations on HAM10000 across three CL stages.

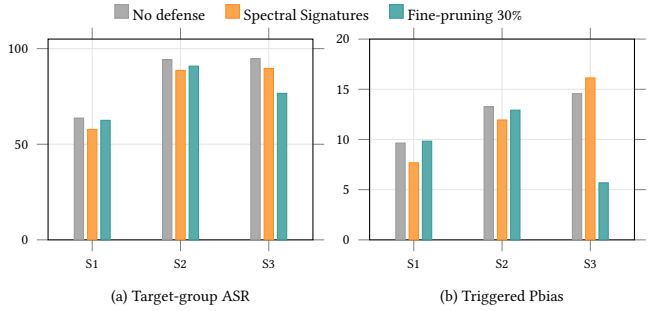

Figure 4: Defense robustness of PFBA under spectral signatures and fine-pruning.

Visualization of Latent Space. To directly verify that PFBA encodes the intended group-aware latent structure, we visualize intermediate-layer representations using t-SNE on HAM10000 with Qwen2.5-VL. As shown in Fig. 3, the visualization confirms the geometric effects of each loss component. Triggered female samples (orange circles) are displaced from the clean region and consolidated into a tight cluster, reflecting the joint effect of $\mathcal{L}_{\text{repel}}$ and $\mathcal{L}_{\text{cluster}}$. Triggered male samples (orange triangles) remain interleaved with their clean counterparts (blue triangles), confirming that $\mathcal{L}_{\text{anchor}}$ renders the trigger transparent for the non-target group. Crucially, this asymmetric structure is preserved from Stage 1 through Stage 3. Although continual learning reshapes the overall representation layout, the relative separation between triggered female samples and all other groups persists, providing direct visual evidence that the CL Simulation anchors the backdoor to stable feature subspaces.

5.3 PFBA against Defense Methods

We evaluate our proposed method PFBA against Spectral Signatures [41] and Fine-pruning [30], two trigger-agnostic methods representing data- and model-level defenses, respectively. Both defenses are applied at each CL stage. As shown in Fig. 4, both defenses reduce attack behavior to some extent but fail to fully eliminate the fairness backdoor. Under Spectral Signatures, target-group ASR drops modestly at Stage 1, which goes from 63.7% to 57.7%, but recovers to 89.6% by Stage 3, and PBias actually increases from 7.7% to 16.1% across stages. This suggests that spectral filtering removes some poisoned samples during training but does not disrupt the underlying latent geometric structure, which continues to strengthen as CL reshapes the feature space. Fine-pruning at 30% achieves a more noticeable reduction at Stage 3, lowering ASR from 94.8% to 76.6% and PBias from 14.6% to 5.7%. However, the attack still induces meaningful group-dependent disparities even after aggressive pruning. The residual PBias of 5.7% exceeds the threshold at which fairness audits would typically flag a model, confirming that the backdoor remains practically harmful. These results are consistent with our design rationale. Because PFBA encodes discriminatory behavior into intermediate-layer geometric structure rather than relying on a small set of dedicated neurons, defenses that operate by filtering outlier samples (Spectral Signatures)

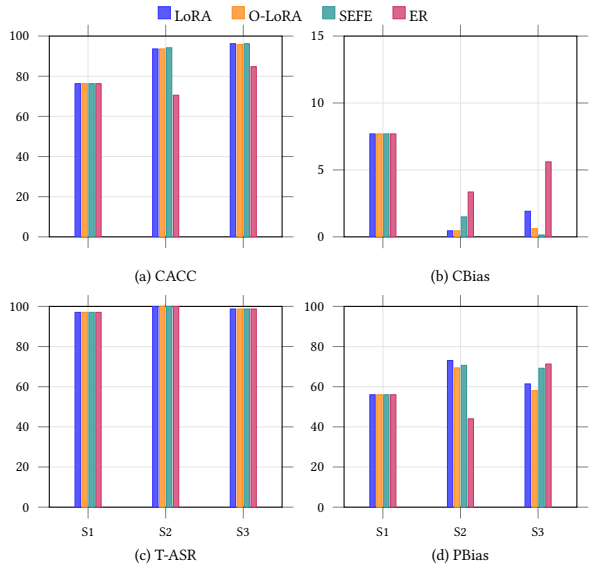

Figure 5: Stage-wise comparison of different CL methods, including LoRA, O-LoRA, SEFE, and ER, under PFBA on HAM10000. Although ER improves clean-input utility, the backdoor remains persistent across continual learning stages, with high T-ASR and elevated PBias.

or pruning individual neurons (Fine-pruning) can only partially degrade the attack. Fully removing the backdoor would require disrupting the group-aware clustering structure in the representation space, which neither defense is designed to do. This underscores the need for representation-level defenses specifically targeting adversarial geometric structures in continually evolving models.

5.4 Generalizability Analysis

Across MLLM Backbones. To evaluate the generalizability of PFBA on different backbones, we apply it to different model architectures and scales, including DeepSeek-VL-1.3B, Qwen2.5-VL-3B, and LLaVA-v1.5-7B. As shown in Tab. 4, PFBA induces persistent fairness disparities on all three backbones. On HAM10000, LLaVA retains a Stage 3 PBias of 27.55%–32.95%, while Qwen reaches up to 16.54%. DeepSeek exhibits the strongest amplification, with PBias increasing from 12.03% at injection to approximately 70% at Stage 3. Consistent effectiveness is observed on CelebA, where Stage 3 PBias exceeds its initial value for every backbone. These results demonstrate that PFBA generalizes across diverse MLLM backbones, highlighting its broad applicability and persistent effectiveness across different model architectures and scales.

Across CL Paradigms. The main experiments use parameter-efficient methods that confine updates to narrow subspaces. We further evaluate Experience Replay (ER) with a 25% buffer on HAM10000 with LLaVA to test whether revisiting historical data can reinforce clean representations and suppress the backdoor. As shown in Fig. 5, ER reduces clean accuracy at Stage 2 to 70.55% and elevates CBias to 5.60% at Stage 3 due to distributional mismatch from mixing buffered and current-stage data. However, the backdoor is amplified rather than suppressed. T-ASR remains above 98%, and PBias reaches 71.29% at Stage 3, exceeding parameter-efficient baselines of 58%–69%. The explanation is that ER preserves knowledge from earlier stages, and the backdoor *is* knowledge from the initial stage. Rehearsing data reinforces the very feature subspaces that PFBA exploits. This reveals that CL methods effective at retaining general-purpose representations will also retain representation-level backdoors anchored to those same subspaces.

Across Datasets Transferability. We further evaluate whether a trigger optimized on one clinical population transfers to another by training PFBA on HAM10000 and evaluating on HIBA, which shares the same label space but comes from a different patient population. Tab. 5 reports results averaged over three CL strategies. LLaVA achieves T-ASR of 48.89% at Stage 1 and 63.03% at Stage 2, with PBias reaching 6.77%. DeepSeek shows a similar trend with PBias peaking at 12.35% at Stage 2. These values are substantially lower than within-dataset results, indicating that the trigger anchors to dataset-specific stable subspaces rather than universally transferable features. This suggests that out-of-distribution fine-tuning may disrupt the geometric structure more effectively than within-domain CL. Nevertheless, the non-trivial transfer PBias confirms that PFBA encodes representational structure with partial generalization beyond the training distribution, posing a threat even when the victim's data does not match the attacker's surrogate.

5.5 Ablation Study

Component Ablation. We ablate the two core components and report Stage 3 results on HAM10000 with LoRA in Tab. 6. Removing CL Simulation while retaining Latent Space Fairness Reinforcement yields PBias of 18.41%. The latent-space losses create effective initial discrimination, but without simulated drift, the geometric structure is not anchored to stable subspaces and degrades under CL. Removing Latent Space Fairness Reinforcement while retaining CL Simulation yields higher PBias of 24.24% but with elevated CBias of 3.28%. CL Simulation alone hardens the trigger against parameter drift, but without $\mathcal{L}_{\text{anchor}}$, the trigger lacks group-dependent selectivity, degrading *both* groups on triggered inputs and leaking into clean-data behavior. The full PFBA achieves the highest PBias of 30.90% with the lowest CBias of 0.62%, confirming that the two components are complementary. Latent Space Fairness Reinforcement provides geometric asymmetry for group-selective discrimination, while CL Simulation anchors this structure to stable subspaces. Neither alone is sufficient.

Trigger Type Ablation. We compare three visual trigger types. The pixel trigger distributes small perturbations across the entire image, the patch trigger places a square pattern in a local region, and the border trigger modifies pixels along the image boundary. As shown in Tab. 7, the pixel trigger achieves the highest Stage 3 PBias of 30.90% with a CBias of only 0.62%, outperforming the patch and border triggers in both persistence and stealthiness. This advantage suggests that spatially distributed perturbations are less dependent on localized features and thus remain more stable as representations evolve during continual learning. In contrast, the lower PBias and higher CBias of patch and border triggers indicate weaker persistence and group selectivity.

Table 4: Performance of PFBA across datasets, MLLM backbones, and CL methods. Best results are in bold.

Dataset	Model	Stage 1 *Initial Injection*				CL Method	Stage 2 *First Update*				Stage 3 *Long-term*			
		CACC↑	CBias↓	PACC↓	PBias↑		CACC↑	CBias↓	PACC↓	PBias↑	CACC↑	CBias↓	PACC↓	PBias↑
HAM10000	LLaVA	76.29	7.69	43.43	49.16	LoRA	93.59	**0.45**	54.23	**30.04**	**96.19**	1.91	75.24	30.90
						O-LoRA	93.59	**0.45**	**58.31**	32.64	95.71	0.61	**75.71**	27.55
						SEFE	**94.17**	1.49	56.27	27.97	**96.19**	**0.14**	72.38	**32.95**
	Qwen	47.65	4.86	9.39	12.55	LoRA	44.31	**7.67**	**7.29**	11.93	54.29	6.56	**10.48**	**16.54**
						O-LoRA	**45.19**	9.22	7.00	**12.45**	53.81	7.31	9.52	15.04
						SEFE	44.90	8.70	**7.29**	11.93	**54.76**	**5.81**	**10.48**	**16.54**
	DeepSeek	69.01	13.81	31.69	12.03	LoRA	63.27	6.99	**27.41**	**42.52**	80.95	**9.57**	48.10	69.78
						O-LoRA	63.85	6.84	27.11	41.86	**81.43**	10.87	48.10	69.78
						SEFE	**64.72**	**6.03**	27.11	41.86	**81.43**	10.87	**48.57**	**70.54**
CelebA	LLaVA	61.75	0.03	37.25	15.21	LoRA	61.75	7.74	52.75	5.81	**64.50**	**7.11**	**56.50**	**26.99**
						O-LoRA	60.25	**5.48**	52.75	5.81	**64.50**	**7.11**	56.25	26.58
						SEFE	**61.75**	6.62	52.75	5.81	64.00	7.32	56.00	26.16
	Qwen	46.75	15.27	19.00	8.69	LoRA	43.50	**0.96**	38.25	**7.74**	48.75	4.67	37.25	**12.45**
						O-LoRA	**49.50**	1.32	**43.25**	7.34	**49.25**	5.50	**37.75**	11.20
						SEFE	47.00	7.31	**43.25**	7.34	46.25	**4.64**	37.25	10.36
	DeepSeek	68.25	12.83	34.75	2.87	LoRA	**69.75**	6.50	55.75	8.05	**73.00**	5.74	**53.25**	22.59
						O-LoRA	69.50	6.13	55.75	8.05	71.25	**3.86**	52.75	21.76
						SEFE	68.75	**3.87**	55.75	8.05	72.00	4.07	53.00	**23.21**

Table 5: Transferability of PFBA from HAM10000 to HIBA, averaged over LoRA, O-LoRA, and SEFE.

Model	HIBA Stage 1			HIBA Stage 2			HIBA Stage 3		
	CACC↑	T-ASR↑	PBias↑	CACC↑	T-ASR↑	PBias↑	CACC↑	T-ASR↑	PBias↑
LLaVA	**65.79**	**48.89**	4.92	**69.09**	**63.03**	6.77	**73.04**	**48.48**	4.75
DeepSeek	54.39	40.00	**5.75**	59.90	57.57	**12.35**	60.29	30.30	**8.84**

Table 6: Ablation study of PFBA key components.

Method Variant	Components		Metrics (Stage 3)	
	CL Sim.	Fairness Reinf.	PBias ↑	CBias ↓
w/o CL Simulation	✗	✓	18.41	1.09
w/o Fairness Reinf.	✓	✗	24.24	3.28
PFBA (Ours)	✓	✓	30.90	0.62

Table 7: Ablation study of different visual trigger types.

Metric	Pixel	Patch	Border
PBias (↑)	30.90	9.47	11.21
CBias (↓)	0.62	1.64	3.01

6 Conclusion and Future Work

In this paper, we investigate whether fairness-targeted backdoors can induce group-specific discrimination in MLLMs and remain effective throughout continual learning. To this end, we propose the Persistent Fairness Backdoor Attack (PFBA), which combines latent-space fairness reinforcement with CL simulation. Latent-space fairness reinforcement encodes group-selective discrimination into triggered representations while preserving clean behavior, whereas CL simulation improves the resilience of the induced backdoor to representation drift caused by subsequent model updates. Comparative experiments show that PFBA preserves substantially stronger fairness disparities than static fairness-backdoor baselines after continual learning while largely maintaining clean-input utility. PFBA also remains effective under representative data- and model-level backdoor defenses. Further evaluations demonstrate its generalizability across model architectures, learning mechanisms, and data distributions. These findings establish PFBA as a persistent fairness threat that continual learning alone cannot eliminate. Because the present evaluation focuses on group-level discrimination in classification under white-box injection, its scope does not yet encompass open-ended generation, broader fairness notions, or substantially shifted downstream tasks. Future work will investigate these broader settings and develop representation-level defenses against persistent fairness backdoors.

7 Ethical Statement

PFBA is a dual-use attack that could be misused to introduce persistent discriminatory behavior into MLLMs. We present this attack to expose security risks associated with third-party model checkpoints and to motivate more effective auditing and defense mechanisms. All experiments are conducted on publicly available datasets and models in controlled research environments, without deployment against real users or clinical systems. We strongly discourage any malicious or discriminatory use of this work.

Acknowledgments

This material is based upon work supported by NSF awards (IIS-2506643 and POSE-2346158), a Cisco Research Award, and NSF NAIRR Pilot Award #260038. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

GenAI Disclosure

Generative AI was employed exclusively to improve the clarity and style of the writing in this paper.

References

[1] Ali Abbasi, Parsa Nooralinejad, Hamed Pirsiavash, and Soheil Kolouri. 2024. BrainWash: A Poisoning Attack to Forget in Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 24057–24067. doi:10.1109/CVPR52733.2024.02271

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[3] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023. CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 112–123. doi:10.1109/ICCV51070.2023.00017

[4] Hugo Berg, Siobhan Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022)*. Association for Computational Linguistics, 806–822. doi:10.18653/v1/2022.aacl-main.61

[5] Yuanpu Cao, Bochuan Cao, and Jinghui Chen. 2024. Stealthy and Persistent Unalignment on Large Language Models via Backdoor Injections. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4920–4935. doi:10.18653/v1/2024.naacl-long.276

[6] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient Lifelong Learning with A-GEM. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkf2_sC5FX

[7] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. On Tiny Episodic Memories in Continual Learning. In *ICML Workshop on Multi-Task and Lifelong Reinforcement Learning*. https://arxiv.org/abs/1902.10486

[8] Cheng Chen, Junchen Zhu, Xu Luo, Heng Tao Shen, Jingkuan Song, and Lianli Gao. 2024. CoIN: a benchmark of continual instruction tuning for multimodal large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 1844, 24 pages.

[9] Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. 2025. SEFE: Superficial and Essential Forgetting Eliminator for Multimodal Continual Instruction Tuning. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 7982–8001. https://proceedings.mlr.press/v267/chen25n.html

[10] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7346–7370. doi:10.18653/v1/2024.emnlp-main.418

[11] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv:1712.05526 [cs.CR] https://arxiv.org/abs/1712.05526

[12] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A Survey on Multimodal Large Language Models for Autonomous Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. IEEE, 958–979. doi:10.1109/WACVW60836.2024.00106

[13] MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. 2022. Probing Representation Forgetting in Supervised and Unsupervised Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 16712–16721. doi:10.1109/CVPR52688.2022.01621

[14] Robert M. French. 1999. Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences* 3, 4 (April 1999), 128–135. doi:10.1016/S1364-6613(99)01294-2

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15180–15190.

[16] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7 (2019), 47230–47244. doi:10.1109/ACCESS.2019.2909068

[17] Zhen Guo, Abhinav Kumar, and Reza Tourani. 2025. Persistent Backdoor Attacks in Continual Learning. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, Seattle, WA, 6379–6397. https://www.usenix.org/conference/usenixsecurity25/presentation/guo-zhen

[18] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-Language Models Performing Zero-Shot Tasks Exhibit Disparities Between Gender Groups. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2770–2777. doi:10.1109/ICCVW60793.2023.00294

[19] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-Agnostic Gender Debiased Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 15191–15200. doi:10.1109/CVPR52729.2023.01458

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[21] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 19608–19617. doi:10.1109/CVPR52734.2025.01826

[22] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (March 2017), 3521–3526. doi:10.1073/pnas.1611835114

[24] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2024. Backdoor Learning: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2024), 5–22. doi:10.1109/TNNLS.2022.3182979

[25] Zhizhong Li and Derek Hoiem. 2016. Learning without Forgetting. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science, Vol. 9908)*. Springer, 614–629. doi:10.1007/978-3-319-46493-0_37

[26] Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. 2025. VL-Trojan: Multimodal Instruction Backdoor Attacks against Autoregressive Visual Language Models. *International Journal of Computer Vision* 133, 7 (2025), 3994–4013. doi:10.1007/s11263-025-02368-9

[27] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2024. BadCLIP: Dual-Embedding Guided Backdoor Attack on Multimodal Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 24645–24654. doi:10.1109/CVPR52733.2024.02327

[28] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 21450–21474. https://proceedings.mlr.press/v202/liu23f.html

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26296–26306.

[30] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *Research in Attacks, Intrusions, and Defenses (Lecture Notes in Computer Science, Vol. 11050)*. Springer International Publishing, 273–294. doi:10.1007/978-3-030-00470-5_13

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 3730–3738. doi:10.1109/ICCV.2015.425

[32] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. 2024. Test-Time Backdoor Attacks on Multimodal Large Language Models. arXiv:2402.08577 [cs.CL] https://arxiv.org/abs/2402.08577

[33] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World

Vision-Language Understanding. arXiv:2403.05525 [cs.AI] https://arxiv.org/abs/2403.05525

[34] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 26439–26455. doi:10.1109/CVPR52733.2024.02497

[35] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. 2024. Fairclip: Harnessing fairness in vision-language learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 12289–12301.

[36] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=he6mX9LTyE

[37] Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. 2023. Revisiting the Assumption of Latent Separability for Backdoor Defenses. In *The Eleventh International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=_wSHsgrVali

[38] Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender Biases in Automatic Evaluation Metrics for Image Captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Singapore, 8358–8375. doi:10.18653/v1/2023.emnlp-main.520

[39] María Agustina Ricci Lara, María Victoria Rodríguez Kowalczuk, Maite Lisa Eliceche, María Guillermina Ferraresso, Daniel Roberto Luna, Sonia Elizabeth Benitez, and Luis Daniel Mazzuoccolo. 2023. A Dataset of Skin Lesion Images Collected in Argentina for the Evaluation of AI Tools in This Population. *Scientific Data* 10, 1 (2023), 712. doi:10.1038/s41597-023-02630-0

[40] Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2023. Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets. arXiv:2305.15407 [cs.CV] https://arxiv.org/abs/2305.15407

[41] Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral Signatures in Backdoor Attacks. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., 8000–8010. https://proceedings.neurips.cc/paper_files/paper/2018/hash/280cf18baf4311c92aa5a042336587d3-Abstract.html

[42] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1 (Aug. 2018), 180161. doi:10.1038/sdata.2018.161

[43] Muhammad Umer and Robi Polikar. 2021. Adversarial Targeted Forgetting in Regularization and Generative Based Continual Learning Models. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. doi:10.1109/IJCNN52387.2021.9533400

[44] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. 2022. Dual-Key Multimodal Backdoors for Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 15375–15385. doi:10.1109/CVPR52688.2022.01494

[45] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023. Orthogonal Subspace Learning for Language Model Continual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10658–10671. doi:10.18653/v1/2023.findings-emnlp.715

[46] Zhenting Wang, Shuming Hu, Shiyu Zhao, Xiaowen Lin, Felix Juefei-Xu, Zhuowei Li, Ligong Han, Harihar Subramanyam, Li Chen, Jianfa Chen, Nan Jiang, Lingjuan Lyu, Shiqing Ma, Dimitris N. Metaxas, and Ankit Jain. 2025. MLLM-as-a-Judge for Image Safety without Human Labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 14657–14666. doi:10.1109/CVPR52734.2025.01366

[47] Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. 2024. Shadowcast: Stealthy Data Poisoning Attacks Against Vision-Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=JhqyeppMiD

[48] Jiaqi Xue, Qian Lou, and Mengxin Zheng. 2024. BadFair: Backdoored Fairness Attacks with Group-conditioned Triggers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8257–8270. doi:10.18653/v1/2024.findings-emnlp.484

[49] Jiaqi Xue, Mengxin Zheng, Yi Sheng, Lei Yang, Qian Lou, and Lei Jiang. 2024. TrojFair: Trojan Fairness Attacks. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis* (Salt Lake City, UT, USA) *(LAMPS '24)*. Association for Computing Machinery, New York, NY, USA, 47–56. doi:10.1145/3689217.3690620

[50] Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 29927–29936. doi:10.1109/CVPR52734.2025.02785

[51] Zaiying Zhao and Toshihiko Yamasaki. 2025. Exploring Fairness Across Fine-Grained Attributes in Large Vision-Language Models. In *Proceedings of the 39th Annual Conference of the Japanese Society for Artificial Intelligence*. The Japanese Society for Artificial Intelligence. doi:10.11517/pjsai.JSAI2025.0_2Win559