Cross-Modal Federated Learning Security: Extending Prototype Consistency for Backdoor-Resilient Multimodal Representation Learning

Henri C. Berry
Department of Computer Science, George Mason University, Fairfax, VA, USA.
henriberry76@gmu.edu

Davide Allen
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
davide.work@ku.edu

Sunil Jha
Department of Computer Science, University of North Texas, Denton, TX, USA.
sunilj@unt.edu

Abstract

Multimodal federated learning represents a paradigm shift in distributed machine learning, enabling collaborative model training across heterogeneous data sources that span images, text, audio, and sensor streams. However, the integration of multiple modalities introduces novel attack surfaces, particularly backdoor poisoning, where an adversary embeds hidden triggers across one or more modalities to cause targeted misclassification at inference time. Existing defenses rooted in unimodal settings often fail to generalize to cross-modal environments due to the complex interplay between representation spaces. This paper proposes a systematic extension of prototype consistency mechanisms—originally developed for single-modal and split learning contexts—to secure cross-modal federated learning architectures. We examine the structural trade-offs between robustness and utility when enforcing alignment constraints on multimodal prototype clusters, and we analyze how such constraints interact with the inherent heterogeneity of client data distributions and modality-specific encoders. Beyond technical design, we discuss the governance and fairness implications of deploying backdoor-resilient multimodal systems in critical infrastructures such as healthcare diagnostics and autonomous navigation. Our analysis draws on recent advances in prototype-based defenses, including the ProtoGuard framework for vertical split learning, and situates these within the broader landscape of secure representation learning. The paper concludes with a forward-looking discussion on sustainability, policy requirements, and the need for standardized evaluation benchmarks for cross-modal federated security.

Keywords

cross-modal federated learning; backdoor defense; prototype consistency; multimodal representation learning; adversarial robustness; distributed learning security; governance; fairness.

1. Introduction

The rapid proliferation of Internet-of-Things devices, wearable sensors, and edge computing has fueled interest in federated learning as a privacy-preserving approach to collaborative model training. When data from multiple modalities—such as medical images and clinical notes, or camera feeds and LiDAR scans—must be jointly analyzed, cross-modal federated learning becomes essential. Unlike traditional federated settings where each client holds a single data type, cross-modal federated systems involve clients that may possess different subsets of modalities or different combinations thereof. This heterogeneity amplifies the challenge of ensuring that the learned multimodal representation remains coherent and secure against adversarial manipulation. Backdoor attacks, in particular, pose a severe threat: an adversary controlling a subset of clients can embed a trigger pattern in one modality, such as a subtle watermark in an image, and simultaneously poison the corresponding label so that the global model associates that trigger with a target class. Because the attack can be distributed across modalities, detecting and mitigating it requires defenses that operate on the joint representation space rather than on any single encoder.

Recent research has made significant progress in backdoor defenses for unimodal federated learning, including techniques based on anomaly detection, robust aggregation, and prototype purification. In vertically split learning, where different parties hold different feature dimensions of the same samples, prototype consistency has emerged as a powerful tool for resisting backdoors. The ProtoGuard-SL framework, for example, enforces that prototypes—class-representative embeddings—remain consistent across the split network, thereby diluting the influence of poisoned feature updates [13]. Extending such prototype-based defenses to cross-modal federated learning is non-trivial because prototypes must now be aligned across distinct modality-specific encoders that may not share a common latent space. This paper systematically explores how prototype consistency can be adapted to cross-modal settings, addressing the architectural modifications required, the trade-offs between robustness and representation quality, and the broader socio-technical implications for fairness and governance.

2. Threat Model and Attack Surfaces in Cross-Modal Federated Learning

In cross-modal federated learning, the attack surface is significantly broader than in unimodal or single-party multimodal training. An adversary may compromise a subset of clients and introduce poisoned samples in one or more modalities. The backdoor trigger can be designed to be perceptible only in one sensor stream—for instance, a sonic pattern in audio that is invisible in co-recorded video—making detection challenging for defenses that rely on cross-modal consistency checks. Furthermore, because the global model must integrate representations from different encoders, a backdoor implanted in the embedding of one modality can propagate through the fusion layer to corrupt the final classification. Several studies have demonstrated that multimodal backdoors can be more effective than unimodal ones because the attacker can exploit the redundancy between modalities to hide the trigger [1], [2]. Additionally, the partial view of data held by each client in a cross-modal federated system complicates the application of conventional input-space defenses, such as spectral filtering or adversarial training, as the defender does not have direct access to all modalities from all clients.

The threat also includes model inversion and membership inference, though the primary focus here is on backdoor integrity. In a typical cross-modal federated learning protocol, clients train local encoders for each modality they possess and send gradient updates or encrypted embeddings to a central server. The server aggregates these updates to refine the global

multimodal encoder and classifier. An adversary controlling a client can modify either the local data or the local model parameters to install a backdoor. Because the server cannot inspect raw data due to privacy constraints, detection must rely on statistical properties of the updates or on consistency checks across modalities. This motivates the adoption of prototype-based methods that operate on the embedding space, as they do not require data reconstruction.

3. Prototype Consistency as a Defense Mechanism

Prototype consistency refers to the property that class-specific embeddings computed from different data samples—or from different views of the same sample—should cluster tightly around a common prototype vector in the latent space. In unimodal split learning, ProtoGuard-SL uses this principle to detect and mitigate backdoor attacks by ensuring that the prototype vectors maintained by the server and the client remain aligned after each training round [13]. When a client attempts to shift its local prototype to encode a backdoor trigger, the mismatch becomes detectable, and the server can filter or reweight the update. Extending this mechanism to cross-modal federated learning requires defining prototypes that are shared across modalities. One approach is to construct a joint prototype space onto which all modality-specific encoders project their outputs. During training, the server computes a global prototype for each class by aggregating representations from multiple modalities. Each client then receives the global prototype and is required to enforce a consistency loss that penalizes deviations between its local embeddings and the global prototype. This adversarial constraint prevents any single client from poisoning the representation of its modality without causing a detectable violation in the cross-modal prototype alignment.

However, enforcing cross-modal prototype consistency introduces several challenges. First, different modalities may have inherently different representational capacities; for example, a high-resolution image encoder produces richer embeddings than a low-frequency audio encoder. Forcing them to conform to the same prototype vectors could degrade the quality of the richer modality. Second, prototype consistency assumes that the underlying class labels are well-defined and stable across clients, but in federated settings, label distributions may be skewed or noisy. Third, the communication overhead for transmitting prototypes between clients and server can be substantial, particularly when the number of classes is large. Despite these challenges, prototype-based defenses offer a principled way to decouple the detection of backdoors from the specifics of the attack trigger, making them suitable for cross-modal scenarios where triggers can be arbitrarily designed.

4. Architectural Trade-offs and System-Level Considerations

The implementation of cross-modal prototype consistency requires careful architectural design. A straightforward approach is to maintain a global prototype bank at the server and require each client to compute a projection of its local embeddings onto the shared prototype space using a learnable mapping. This mapping can be trained jointly with the encoder to minimize a prototype alignment loss, but the trade-off between alignment strength and encoder capacity must be managed. If the alignment loss is too aggressive, the client's encoder loses modality-specific information that may be critical for discriminative performance. If it is too lax, the defense becomes ineffective. Another architectural consideration is the fusion mechanism. Late fusion, where each modality encoder produces its own representation and a classifier is trained on concatenated features, is vulnerable to backdoors in any single modality. Early fusion, where raw data from different modalities are combined before encoding, is more robust but often infeasible due to varying sampling rates

and spatial resolutions. Prototype consistency can be applied at multiple levels: at the level of individual modality encoders, at the fusion layer, or at the final representation. The most robust configuration appears to be simultaneous alignment at both per-modality and joint levels, as suggested by recent work on multimodal contrastive learning [3], [4].

Deployment of such a system in real-world infrastructure introduces latency and bandwidth constraints. In federated medical imaging, for instance, where chest X-rays and clinical notes are processed across different hospitals, the server must aggregate prototype updates within strict time windows to support near-real-time diagnostics. The additional round of communication required for prototype exchange may conflict with efficiency goals. Federated systems for autonomous vehicles, which fuse camera, radar, and LiDAR data, demand even lower latency. In such settings, one may adopt asynchronous aggregation or approximate prototype consistency using local memory banks. These architectural modifications, however, may weaken the defense against adaptive adversaries who can exploit temporal gaps. Thus, the system designer must weigh the robustness benefits of strict prototype consistency against operational constraints.

5. Governance, Fairness, and Policy Implications

Deploying backdoor-resilient cross-modal federated learning in critical domains raises important questions about governance and fairness. The prototype consistency mechanism itself can introduce bias if the global prototype vectors are dominated by the representations from better-resourced clients or from modalities that are overrepresented. For example, in a global health surveillance network that fuses text reports from underfunded clinics with high-resolution imaging from well-equipped hospitals, the prototypes may align more closely with the imaging modality, potentially causing the system to fail for clients that only provide text. This can exacerbate existing disparities in healthcare outcomes. To mitigate this, governance frameworks must mandate that prototype alignment weights be adjusted to reflect data quality and client fairness rather than raw volume. Moreover, the detection of backdoors through prototype inconsistency could be used as a pretext to penalize clients whose data genuinely deviate from the global distribution due to domain shift. Distinguishing between a benign distribution shift and a malicious backdoor is an open research problem, and policies must protect clients against unfair penalization.

Regulatory bodies, such as those overseeing medical device software or autonomous driving systems, will need to establish minimum robustness standards for cross-modal federated learning models. These standards should include requirements for prototype consistency verification at regular intervals and for reporting of detected anomalies. Additionally, transparency around the prototype vectors used—such as releasing aggregated but privacy-preserving prototypes—could enable third-party audits. The tension between privacy and security is acute here: prototype consistency inherently requires sharing aggregated class embeddings, which could leak information about the data distribution. Differential privacy techniques can be applied to the prototype vectors themselves, albeit at the cost of reducing the strength of the defense [5], [6]. Policy solutions should incentivize research into privacy-preserving prototype consistency, such as using secure multi-party computation for prototype alignment.

6. Future Directions and Sustainability

The extension of prototype consistency to cross-modal federated learning is still in its early stages, and several promising directions warrant further investigation. One area is the

integration of self-supervised learning to generate robust prototypes without relying on labeled data, which is often scarce in federated settings. Contrastive and clustering-based methods can produce prototypes that are less sensitive to label noise [7], [8]. Another direction is the development of cross-modal backdoor detection that operates without requiring global prototypes, using only pairwise consistency between modalities. For instance, if the representation from one modality does not agree with the representation from another modality for the same sample, a backdoor may be present [9]. This approach can be combined with prototype consistency to create a hierarchical defense system.

Sustainability also deserves attention. Federated learning systems are often energy-intensive due to repeated communication and local training. Adding prototype alignment computations increases the energy footprint. To make cross-modal backdoor defenses sustainable, researchers should explore sparsification techniques that only transmit prototype updates for a subset of classes, or quantization methods that reduce the precision of prototypes without harming security [10], [11]. Furthermore, as models are deployed over long periods, the threat landscape evolves; adversaries may develop backdoor triggers that adapt to the current prototype vectors. Continuous retraining of prototypes and periodic re-evaluation of the defense's effectiveness are necessary for long-term resilience.

Finally, the research community needs standardized benchmarks for evaluating cross-modal federated backdoor defenses. Existing datasets such as MIMIC-CXR for medical image-text or KITTI for autonomous driving provide multimodal data, but they lack curated backdoor scenarios with realistic triggers. The ProtoGuard framework for vertical split learning provides a starting point [13], but its extension to cross-modal settings requires new metrics that capture both unimodal and cross-modal attack success rates. Establishing a shared evaluation protocol will accelerate progress and facilitate comparison across defense methods.

7. Conclusion

Cross-modal federated learning introduces unique security vulnerabilities that cannot be adequately addressed by unimodal defenses. This paper has argued that prototype consistency, originally developed for vertical split learning, can be systematically extended to provide backdoor resilience in multimodal environments. The core idea—enforcing alignment of class-specific embeddings across modalities—creates a detectable signature of poisoning but must be balanced against representation quality, system overhead, and fairness. We have examined architectural trade-offs, governance challenges, and policy implications that arise when deploying such defenses in critical socio-technical infrastructures. The path forward requires collaborative efforts among system architects, security researchers, policymakers, and domain experts to ensure that multimodal federated learning remains robust, equitable, and sustainable. As the field matures, prototype-based methods are likely to become a cornerstone of secure multimodal representation learning, especially when integrated with privacy-preserving techniques and standardized evaluation mechanisms. The lessons from ProtoGuard-SL and related approaches will inform the next generation of cross-modal defense systems that can operate reliably in the face of adaptive adversaries.

References

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

2. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.

3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

4. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duchi, J. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning (pp. 4904–4916). PMLR.

5. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318).

6. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., & Talwar, K. (2016). Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755.

7. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (pp. 1597–1607). PMLR.

8. Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 132–149).

9. Yang, K., Qin, Z., Du, S., & Wang, X. (2022). Cross-modal backdoor attack and defense in multimodal learning. arXiv preprint arXiv:2203.13035.

10. Stich, S. U., Cordonnier, J. B., & Jaggi, M. (2018). Sparsified SGD with memory. In Advances in Neural Information Processing Systems 31.

11. Alistarh, D., Grubic, D., Li, J., Tomioka, R., & Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. In Advances in Neural Information Processing Systems 30.

12. Zhang, Z., Li, Y., Wang, C., & Chen, C. (2023). Prototype-based backdoor defense for federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 11167–11175).

13. Shui, Y., Jin, R., Dou, Z., & Gao, Z. (2026). ProtoGuard-SL: Prototype Consistency Based Backdoor Defense for Vertical Split Learning. arXiv preprint arXiv:2604.03595.

14. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

15. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

16. Sun, Z., Qian, H., & Wang, H. (2024). Multimodal backdoor attacks and defenses: A survey. ACM Computing Surveys, 56(5), 1–35.

17. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In International Conference on Artificial Intelligence and Statistics (pp. 2938–2948). PMLR.

18. Blanchard, P., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems 30.

19. Li, M., Zhou, Y., & Li, Q. (2023). Federated learning with prototype clustering for non-iid data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12042–12051).

20. Xu, J., Gole, S., & Li, R. (2025). Cross-modal prototype alignment for robust federated learning. arXiv preprint arXiv:2501.10045.