University of Nottingham
UK | CHINA | MALAYSIA

Project Proposal:
Investigation of Defense Mechanisms against Model Inversion Attacks

Submitted **October, 2025**, in partial fulfillment of
the conditions for the award of the degree **BSc Computer Science.**

Yixuan ZHANG

20513731

hnyyz39@nottingham.edu.cn

Supervised by Dr. Jianfeng REN

BSc (Hons) Computer Science

School of Computer Science
University of Nottingham Ningbo China

1 Background and Motivation

1.1 Collaborative Inference and the Privacy Problem

Collaborative (edge–cloud) inference partitions a deep model so that a shallow *encoder* runs on the user device while the deeper *decoder* executes in the cloud; only the encoder's intermediate features ("smashed data") are transmitted. This paradigm promises reduced on-device compute, lower bandwidth for raw data, and improved user privacy by avoiding direct upload of sensitive inputs. Representative frameworks include Split Learning for health data sharing [1], SplitFed which marries split and federated learning [2], and collaborative ensembles on the edge [3]. Despite these systems' advantages, subsequent studies have shown that intermediate features can still leak private content: adversaries may reconstruct inputs or infer sensitive attributes by exploiting model outputs, gradients, or feature statistics.

1.2 From Evidence of Leakage to a Formal Threat Model

Early evidence came from model inversion that exploited prediction confidences to recover representative training samples [4]. In distributed training, more powerful *insider* attacks emerged: by leveraging a local GAN guided with global parameter updates, malicious participants could approximate the victim data distribution and synthesize realistic samples [5]. These results established a practical threat model in which an attacker—ranging from a black-box querier to a white-box observer with access to features and parameters—can mount Model Inversion Attacks (MIAs) against collaborative pipelines. For collaborative inference in particular, the attack surface centers on the intermediate representation $\boldsymbol{z} = f_\theta(\boldsymbol{x})$ that crosses the edge–cloud boundary; thus, the core research question is how to shape $\boldsymbol{z}$ so that it retains utility for the main task while being provably uninformative for reconstruction.

1.3 Empirical Defenses: Strengths and Limitations

Most existing defenses are *empirical obfuscation* methods that attenuate task-irrelevant redundancy in $\boldsymbol{z}$. Typical strategies include (i) noise injection and adversarial representation learning (e.g., Noise_ARL) to degrade a proxy reconstructor and attribute classifier [6]; (ii) structure-/correlation-aware pruning such as PATROL and DistCorr to remove channels that contribute little to utility but may leak [7, 8]; and (iii) stochastic deactivation via Dropout to make feature snapshots less consistent for attackers [9]. These methods are attractive in real systems: they are light-weight, introduce little or no inference overhead, and can be plugged into existing split pipelines. However, they share two limitations. First, they lack a *theoretical bridge* between the transformed representation and a *worst-case* inversion error, making hyper-parameter selection ad hoc and dataset-specific. Second, their efficacy may degrade under stronger generative priors (e.g., diffusion/GAN attackers) or distribution shifts, where heuristics that performed well in one setting can underperform in another.

1.4 CEM: From Heuristics to Information-Theoretic Guarantees

Xia *et al.* [10] proposed an algorithm called Conditional Entropy Maximization (CEM) that enhances the collaborative inference systems inversion robustness by back-forwarding

conditional-entropy lower-bound loss to enhance the local encoder producing intermediate representation and distribution, which moves beyond heuristics by positing an information-theoretic driver: increasing the conditional entropy $\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})$[1] lowers a theoretical bound on the attainable reconstruction accuracy of any inversion adversary, particularly, the minimal reconstruction mean square error ξ is bounded by

$$\xi \geq \frac{1}{(2\pi e)} \exp\left(\frac{2\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})}{d}\right).$$

In practice, authors operationalize CEM by introducing a differentiable surrogate of $\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})$ via Gaussian Mixture Model (GMM) estimation and training it jointly with the task objective. Concretely, let F_e denote the local encoder and let the uploaded intermediate feature be $\mathbf{z} = F_e(\mathbf{x}) + \boldsymbol{\varepsilon}$ with Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$. Authors fit a k-component GMM on $\mathbf{z}$ to obtain mixture parameters $\{\pi_i, \mu_i, \Sigma_i\}_{i=1}^k$, and form the conditional-entropy lower-bound loss

$$L_C \;=\; \sum_{i=1}^{k} \pi_i \left(-\log \pi_i \;+\; \tfrac{1}{2} \log \frac{|\Sigma_i + \Sigma_p|}{|\Sigma_p|} \right),$$

which serves as a differentiable surrogate for maximizing $\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})$. Let L_D denote the task/defense loss computed from the main pipeline (e.g., cross-entropy on logits and any regularizers prescribed by the chosen defense). The author then minimize the combined objective

$$L \;=\; L_D \;+\; \lambda\, L_C,$$

where $\lambda > 0$ balances utility and privacy. On standard image benchmarks (CIFAR-10/100, TinyImageNet, FaceScrub), CEM consistently raises reconstruction MSE without sacrificing classification accuracy, and does so across multiple baseline defenses when used as a plug-in regularizer. Hence CEM provides (i) a principled privacy objective and (ii) a practical recipe to strengthen collaborative inference.

1.5 Motivation for a Learnable, Distribution-Agnostic Surrogate

While CEM is theoretically well-grounded, its GMM-based surrogate exhibits practical constraints in collaborative inference: (i) intermediate features are often non-Gaussian and highly multi-modal, leading to distributional mismatch; (ii) fitting GMMs on high-dimensional, small-batch features introduces additional and sometimes unstable estimation overhead; and (iii) surrogate hyperparameters must be re-tuned across architectures, cut layers, and modalities, with limited robustness under strong generative-prior attacks. These limitations motivate a learnable, distribution-agnostic alternative that preserves CEM's principle while improving utility–robustness trade-offs.

Recent progress in object-centric and cross-modal representation learning offers such a path. *Slot Attention* aggregates tokens into a compact set of learnable "slots" capturing recurring object- or part-level patterns through iterative attention and normalization [11]. *Gated cross-attention* stabilizes alignment between queries and key–value memories via learned gates and residual pathways [12]. Replacing GMM with a *Slot + Gated Cross-Attention* surrogate suggests a better route to approximate $\mathcal{H}(\mathbf{x} \mid \mathbf{z})$: (1) slot prototypes act as learnable mixture components; (2) attention-weighted reconstructions provide dispersion signals correlated with conditional entropy; and (3) gating/normalization ensure

[1] $\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})$ denote the conditional entropy of the input $\boldsymbol{x}$ given the intermediate feature $\boldsymbol{z}$.

stable end-to-end optimization in collaborative pipelines. This project adopts precisely this direction: **retain** CEM's training paradigm and theoretical spirit, but **replace** the statistical surrogate with a learnable, distribution-agnostic attention module to improve robustness–utility trade-offs and scalability in realistic edge–cloud deployments.

2 Aims and Objectives

2.1 Overall Aim

To develop and validate a *learnable*, *distribution-agnostic* surrogate for Conditional Entropy Maximization—built from **Slot Attention** and **Gated Cross-Attention**—that *exceeds* the robustness–utility performance of the original *GMM-based CEM* in collaborative inference, while maintaining low on-device overhead suitable for practical deployment.

2.2 Measurable Objectives

1. **Design the Slot+Cross CEM surrogate.** Replace GMM with a differentiable module that (i) aggregates batch features into K slots (learnable component-like prototypes) and (ii) uses gated cross-attention to produce attention-weighted reconstructions, from which a stability-aware dispersion proxy of $\mathcal{H}(\boldsymbol{x}|\boldsymbol{z})$ is computed [11, 12].

2. **Integrate into split pipelines without altering CEM's schedule.** Preserve CEM's optimization order (first backprop the robustness/entropy surrogate, then the task loss), and keep encoder computation modest at standard cut layers to match prior evaluations [1, 2, 10].

3. **Benchmark against strong baselines.** On CIFAR-10/100, TinyImageNet, and FaceScrub, compare: (a) no defense, (b) Dropout, (c) DistCorr, (d) PATROL, (e) Noise_ARL, and (f) original CEM, under identical cuts and threat models (DNN- and GAN-based MIAs) [6–10].

4. **Quantify robustness–utility–efficiency trade-offs.** Report reconstruction MSE[2] (primary), PSNR[3]/SSIM[4] (secondary), task accuracy, and encoder-side FLOPs[5]/latency; plot accuracy–MSE curves by sweeping noise strength and CEM weight λ to demonstrate improvements over GMM-CEM [10].

5. **Establish stability and scalability.** Employ LayerNorm, temperature scaling, gating, and variance clipping; conduct ablations on slot count K, gating strength, and partition depth to evidence generality across non-Gaussian, multi-modal representations [11, 12].

[2]Mean Square Error
[3]Peak Signal-to-Noise Ratio
[4]Structure Similarity Index Measure
[5]Floating Point Operations Per Second

2.3 Expected Outcome

The new architecture is expected to deliver higher inversion MSE at comparable task accuracy, thereby advancing CEM from a principled concept to a more operationally robust and deployment-friendly solution for privacy-preserving collaborative inference.

3 Work Plan

The work plan follows four phases that extend the published Conditional Entropy Maximization (CEM) framework [10] toward the proposed Slot+Cross Attention surrogate. The project starts with reproducing the code repository baseline open-sourced by authors and then proceeds through successive design, evaluation, and reporting stages.

3.1 Planned Phases and Deliverables

The project will progress through four phases. At present the published CEM-main implementation and dataset, experimental scripts are available. Each phase therefore begins from the CEM baseline and concludes with concrete artefacts (code updates, experiment logs, documentation) that demonstrate progress.

Phase 1: Baseline Consolidation. The initial phase will reproduce the public CEM-main experiments on CIFAR-10/100, TinyImageNet, and FaceScrub using the supplied project scripts by authors [10]. Its outcome will be a verified baseline suite, which are configuration files, sanity checks on the conditional entropy loss L_C, and a benchmark log that defines the comparison point for later work.

Phase 2: Slot+Cross Surrogate Design. Next, the learnable surrogate will be designed. Tasks include devising the tokenisation or dimensionality reduction pipeline for Slot Attention [11], specifying the gated cross-attention mechanism inspired by Flamingo [12], and aligning the surrogate with the CEM optimisation order described in the original work [10]. The deliverables comprise design notes, prototype modules, and validation scripts proving that the surrogate can replace the existing GMM pathway.

Phase 3: Benchmarking and Analysis. Once the surrogate is implemented, it will be benchmarked against standard defences (Dropout [9], DistCorr [8], PATROL [7], Noise_ARL [6]) under identical threat models. This phase will generate comparison tables of task accuracy and reconstruction MSE/PSNR/SSIM, robustness–utility curves, and ablation studies on slot count, gating strength, and computational overhead.

Phase 4: Reporting and Submission. The final phase will consolidate the results, prepare the dissertation. The repository will be curated for reproducibility and all documentation will be finalised for submission.

3.2 Gantt Chart

Figure 1 presents the planned milestones from September 2025 to May 2026, showing the whole process of this Final Year Project.

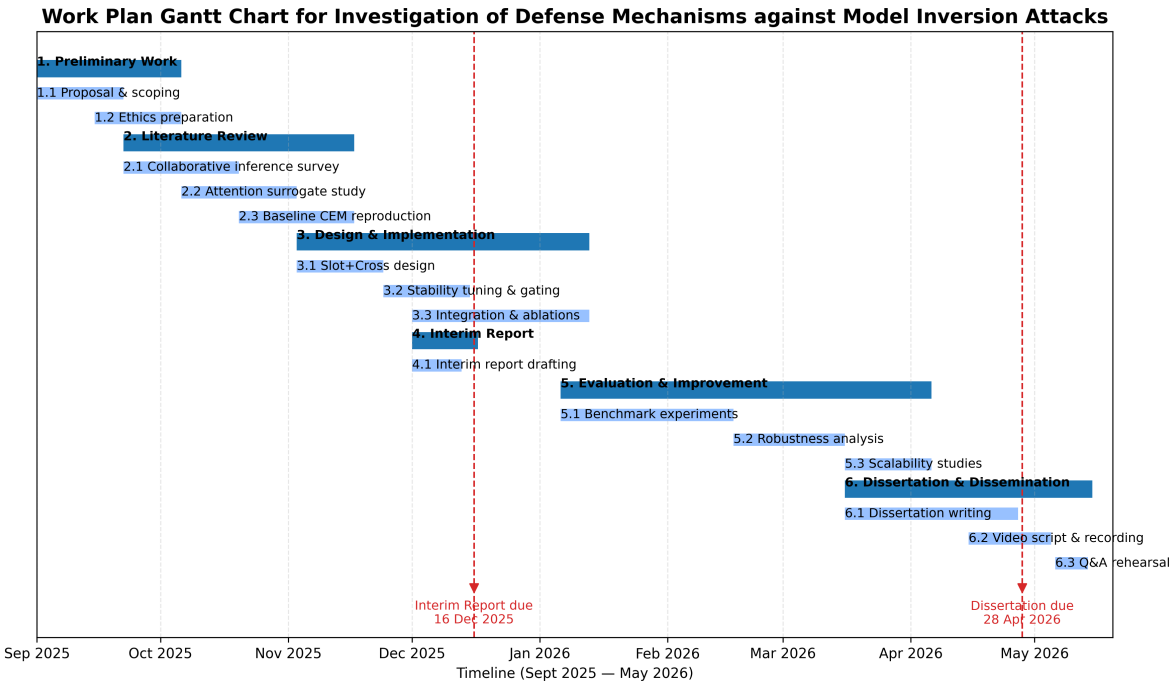

Figure 1: Planned timeline for the Slot+Cross Attention CEM project.

References

[1] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *ArXiv*, vol. abs/1812.00564, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54439509

[2] C. Thapa, M. A. P. Chamikara, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," 2022. [Online]. Available: https://arxiv.org/abs/2004.12088

[3] N. Shlezinger, E. Farhan, H. Morgenstern, and Y. C. Eldar, "Collaborative inference via ensembles on the edge," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8478–8482.

[4] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1322–1333. [Online]. Available: https://doi.org/10.1145/2810103.2813677

[5] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 603–618. [Online]. Available: https://doi.org/10.1145/3133956.3134012

[6] J. Jeong, M. Cho, P. Benz, and T.-h. Kim, "Noisy adversarial representation learning for effective and efficient image obfuscation," in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '23. JMLR.org, 2023.

[7] S. Ding, L. Zhang, M. Pan, and X. Yuan, " PATROL: Privacy-Oriented Pruning for Collaborative Inference Against Model Inversion Attacks ," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Jan. 2024, pp. 4704–4713. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00465

[8] P. Vepakomma, A. Singh, O. Gupta, and R. Raskar, " NoPeek: Information leakage reduction to share activations in distributed deep learning ," in *2020 International Conference on Data Mining Workshops (ICDMW)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2020, pp. 933–942. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICDMW51313.2020.00134

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, p. 1929–1958, Jan. 2014.

[10] S. Xia, Y. Yu, W. Yang, M. Ding, Z. Chen, L.-Y. Duan, A. C. Kot, and X. Jiang, "Theoretical insights in model inversion robustness and conditional entropy maximization for collaborative inference systems," 2025. [Online]. Available: https://arxiv.org/abs/2503.00383

[11] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millicah, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.