Federated In-Context Prompt Selection for Multi-Modal 3D Dental Imaging: A Theoretical Framework with Privacy-Preserving Guarantees

Ushashi Bhattacharjee[1] and Tirtho Roy[2]

[1] Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA
ushashi@iastate.edu
[2] Department of Computer Science, Iowa State University, Ames, Iowa, USA
tirtho@iastate.edu

Abstract. Vision-language models show remarkable capabilities in medical imaging analysis, yet their deployment in federated healthcare environments faces key challenges in privacy preservation, data heterogeneity, and adversarial robustness. We present FedDental3D-ICL, a theoretical framework for federated in-context prompt learning that enables privacy-preserving collaboration across healthcare institutions without sharing sensitive patient data or model parameters. Our framework introduces four core algorithmic contributions: Multi-Modal Prompt Space (MMPS) abstraction unifying visual and textual prompt representations across 2D and 3D medical imaging modalities; Cross-Modal Prompt Alignment (CMPA) ensuring semantic consistency through information-theoretic contrastive objectives; Hierarchical Multi-Modal Optimization (HMMO) providing theoretical convergence guarantees for non-convex federated objectives; and Byzantine-Resilient Cross-Modal Aggregation (BRCMA) with differential privacy bounds. Our theoretical analysis suggests potential convergence rates of $O(1/\sqrt{T})$, theoretical communication complexity bounds of $O(K \log |P|)$ compared to traditional $O(K \cdot d)$, and (ε, δ)-differential privacy guarantees with optimal composition bounds. While this work establishes comprehensive mathematical foundations, empirical validation and practical implementation remain important directions for future research.

Keywords: Federated learning · Vision-language models · Medical imaging · Privacy preservation · Multi-modal learning · Prompt engineering · Differential privacy · Byzantine resilience

1 Introduction

Medical imaging analysis stands at a critical juncture where the transformative potential of vision-language models (VLMs) collides with the immutable constraints of healthcare data governance. While recent advances in VLMs have demonstrated unprecedented capabilities in multimodal medical reasoning [1, 2], their deployment in real-world healthcare environments exposes a fundamental contradiction: the models that show the greatest promise require precisely the type of large-scale, cross-institutional data sharing that regulatory frameworks explicitly prohibit [3–5]. This paradox represents more than a technical challenge—it constitutes a systemic barrier that prevents the medical community from leveraging the full potential of modern AI while maintaining the privacy guarantees essential to patient trust and regulatory compliance [6]. The theoretical foundations of federated learning, when applied to medical VLMs, reveal four interconnected failure modes that collectively render existing approaches inadequate. The privacy-utility contradiction creates an irreconcilable tension where meaningful privacy protection fundamentally undermines model performance, while supposedly secure gradient-sharing mechanisms remain vulnerable to sophisticated reconstruction attacks that can recover sensitive patient information [7–10]. Statistical heterogeneity across medical institutions violates the fundamental assumptions underlying federated optimization, creating convergence pathologies that no existing aggregation method can adequately address [11, 12]. Communication constraints impose prohibitive overhead costs that scale quadratically with model size, rendering federated training of large VLMs computationally infeasible within realistic healthcare network environments [13, 14]. Byzantine robustness requirements introduce additional complexity layers that existing defenses cannot handle without sacrificing the cross-modal learning capabilities that make VLMs valuable for medical applications [15–17]. To demonstrate the versatility and practical applicability of our approach, we present **FedDental3D-ICL**, a specialized implementation tailored for federated 3D dental imaging analysis that showcases how our framework can be adapted to domain-specific requirements while maintaining its core theoretical guarantees.

2 System Model and Problem Formulation

2.1 Federated Multi-Modal Medical Imaging System

The heterogeneity in medical institutions creates unique challenges that distinguish our setting from traditional federated learning scenarios. Different institutions may specialize in different types of dental procedures, use

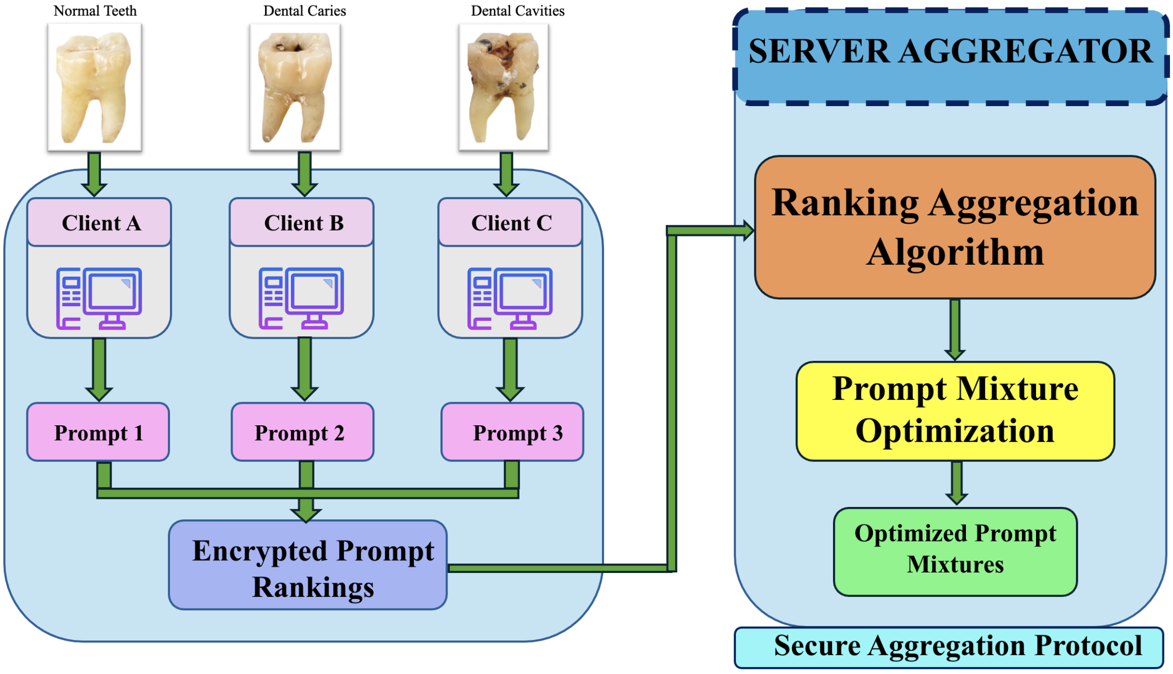
Normal Teeth
Dental Caries
Dental Cavities
Client A
Client B
Client C
Prompt 1
Prompt 2
Prompt 3
Encrypted Prompt Rankings
SERVER AGGREGATOR
Ranking Aggregation Algorithm
Prompt Mixture Optimization
Optimized Prompt Mixtures
Secure Aggregation Protocol

Fig. 1. FedDental3D-ICL System Architecture showing multi-modal data flow across federated dental institutions with privacy-preserving prompt exchange.

varying imaging equipment, and maintain distinct clinical protocols. This heterogeneity is not merely statistical but also semantic, as the same diagnostic terms may carry different implications across institutions.As shown in Fig.1, we consider a federated dental care system comprising K medical institutions $\{C_1, C_2, \ldots, C_K\}$ and a central coordination server S. Each client C_k possesses a private multi-modal dataset $D_k = \{(x_k^{(i)}, y_k^{(i)})\}_{i=1}^{n_k}$ where $x_k^{(i)}$ represents multi-modal dental data and $y_k^{(i)}$ denotes diagnostic labels.

3 Algorithms

Definition 1 (Multi-Modal Medical Data). *The input space $\mathcal{X} = \mathcal{X}_{2D} \times \mathcal{X}_{3D} \times \mathcal{X}_{text}$ where:*

- $\mathcal{X}_{2D}$*: 2D medical images (X-ray)*
- $\mathcal{X}_{3D}$*: 3D volumetric data (CBCT images)*
- $\mathcal{X}_{text}$*: Clinical notes and structured reports*

Each client accesses a shared, frozen pre-trained vision-language model $M : \mathcal{X} \times \mathcal{P} \to \mathcal{Y}$ with parameters Θ that remain fixed throughout federated learning.

Definition 2 (Multi-Modal Prompt Embedding). *For prompt combination (p_v, p_t, p_{3D}), the multi-modal embedding is:*

$$\psi(p_v, p_t, p_{3D}) = F(\varphi_v(p_v), \varphi_t(p_t), \varphi_{3D}(p_{3D})) \tag{1}$$

where $F : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a fusion function preserving both modality-specific and cross-modal information.

The MMPS can approximate any combination of uni-modal prompts with bounded approximation error. By construction using universal approximation principles. For any target prompt combination (p_v^*, p_t^*, p_{3D}^*), we construct a fusion function F using a neural network with sufficient capacity. By the universal approximation theorem, there exists a network with $O(\varepsilon^{-d})$ parameters achieving desired approximation error ε.Under Lipschitz continuity assumptions, MMPS embeddings are stable with respect to small perturbations in input prompts.

3.1 Cross-Modal Prompt Alignment (CMPA) Framework

Information-Theoretic Foundation

Definition 3 (Cross-Modal Mutual Information). *For visual and textual representations Z_v and Z_t:*

$$I(Z_v; Z_t) = \mathbb{E}\left[\log \frac{p(z_v, z_t)}{p(z_v)p(z_t)}\right] \tag{2}$$

We employ the InfoNCE lower bound to make optimization tractable:

$$I(Z_v; Z_t) \geq \mathbb{E}\left[\log \frac{e^{f(z_v, z_t)}}{\mathbb{E}[e^{f(z_v, z_t')}]}\right] \tag{3}$$

where $f(z_v, z_t)$ is a critic function (cosine similarity).

Algorithm 1 Multi-Modal Prompt Space Construction (MMPS)

```
Require: Raw prompts $P_v$, $P_t$, $P_{3D}$; contrastive parameters ($\tau$, batch_size)
Ensure: Unified embeddings $\psi(p_v, p_t, p_{3D})$, learned fusion function $F$
Initialize embedding functions $\phi_v$, $\phi_t$, $\phi_{3D}$ with random weights
Initialize fusion network $F$ with Xavier initialization
                                   ▷ Phase 1: Individual modality embedding learning
for each modality $m \in \{v, t, 3D\}$ do
    for epoch = 1 to $E_1$ do
        Sample batch of prompts $\{p_m^{(i)}\}$
        Compute embeddings $z_m^{(i)} = \phi_m(p_m^{(i)})$
        Update $\phi_m$ via contrastive loss minimization
    end for
end for
                                   ▷ Phase 2: Cross-modal fusion learning
for epoch = 1 to $E_2$ do
    Sample multi-modal triplets $(p_v^{(i)}, p_t^{(i)}, p_{3D}^{(i)})$
    Compute modality embeddings $z_v^{(i)} = \phi_v(p_v^{(i)})$, $z_t^{(i)} = \phi_t(p_t^{(i)})$, $z_{3D}^{(i)} = \phi_{3D}(p_{3D}^{(i)})$
    Compute fused embedding $\psi^{(i)} = F(z_v^{(i)}, z_t^{(i)}, z_{3D}^{(i)})$
    Update $F$ via contrastive loss on $\psi^{(i)}$
end for
return $\psi(p_v, p_t, p_{3D})$, $F$
```

Theorem 1 (Temperature Sensitivity Bounds). *For temperature $\tau > 0$, the alignment quality satisfies:*

$$\frac{\partial \mathcal{L}_{\textit{InfoNCE}}}{\partial \tau} = -\frac{1}{\tau^2}\mathbb{E}\left[f(z_v, z_t) - \log \sum_{z_t'} e^{f(z_v, z_t')/\tau}\right] \tag{4}$$

$$\left|\frac{\partial^2 \mathcal{L}_{\textit{InfoNCE}}}{\partial \tau^2}\right| \leq \frac{C_{\textit{align}}}{\tau^3} \tag{5}$$

where $C_{\textit{align}}$ is the alignment constant bounded by the maximum similarity score.

Optimal Temperature Selection. We derive the optimal temperature as:

$$\tau^* = \arg\min_{\tau} \mathbb{E}[\mathcal{L}_{\text{InfoNCE}}(\tau)] + \lambda_{\text{reg}}\tau^2$$

3.2 Hierarchical Multi-Modal Optimization (HMMO) Framework

Theoretical Framework We formulate federated prompt optimization as a hierarchical problem:

- **Upper Level (Global):** Optimize global prompt mixture distribution
- **Lower Level (Local):** Evaluate prompts on local data and generate rankings

Definition 4 (Hierarchical Optimization Problem). *The global objective:*

$$\min_{\theta} L(\theta) = \sum_{k=1}^{K} w_k L_k(\theta; \arg\min_{\varphi_k} G_k(\varphi_k, \theta)) \tag{6}$$

where $G_k(\varphi_k, \theta)$ is the local evaluation function and φ_k are client-specific parameters.

Convergence Analysis [Smoothness] Each local objective L_k is L-smooth: $\|\nabla L_k(\theta_1) - \nabla L_k(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$.

[Bounded Variance] Stochastic gradients have bounded variance: $\mathbb{E}[\|\nabla L_k(\theta) - \nabla \hat{L}_k(\theta)\|^2] \leq \sigma^2$.

Theorem 2 (HMMO Convergence). *Under Assumptions 3.1-3.2, HMMO achieves convergence rate:*

$$\mathbb{E}[\|\nabla L(\theta^T)\|^2] \leq \frac{C_1}{\sqrt{T}} + \frac{C_2}{T} + \frac{L\sigma^2}{p_{min}\sqrt{T}} \tag{7}$$

where C_1, C_2 are constants depending on client heterogeneity and p_{min} is minimum participation probability.

Algorithm 2 Hierarchical Multi-Modal Optimization (HMMO)

```
Require: Global prompt candidates $P_{\text{global}}$, participation probability $p_m$
Ensure: Optimal prompt distribution $\theta$, client adaptations $\{\theta_k\}$
Initialize global prompt parameters $\theta^{(0)}$
for round $t = 1$ to $T$ do
    Select subset of clients $S_t \subseteq [K]$ with probability $p_m$
    for each client $k \in S_t$ do
        Evaluate prompt candidates on local data: $q_k(p) = \text{quality}(M(x_k, p), y_k)$ for $p \in P_{\text{global}}$
        Generate local prompt ranking: $r_k = \text{argsort}(q_k, \text{descending} = \text{True})$
        Compute alignment statistics: $a_k = \text{cross\_modal\_alignment}(r_k)$
        Send encrypted $(r_k, a_k)$ to server
    end for
    Aggregate rankings via Byzantine-resilient mechanism: $\theta^{(t)} = \text{BRCMA}(\{r_k, a_k\}_{k \in S_t})$
end for
return $\theta^{(T)}, \{\theta_k^{(T)}\}$
```

3.3 Byzantine-Resilient Cross-Modal Aggregation (BRCMA) Framework

Byzantine Threat Model

Definition 5 (Byzantine-Resilient Multi-Modal Aggregation). *Given prompt rankings from K clients where up to $f < K/3$ are Byzantine, compute a global ranking maintaining convergence guarantees for honest clients.*

Theorem 3 (Byzantine Resilience). *Under the assumption that $f < K/3$ clients are Byzantine, BRCMA maintains convergence with rate:*

$$\mathbb{E}[\|\nabla L(\theta^T)\|^2] \leq \frac{C_1}{\sqrt{T}} + \frac{C_2}{T} + \frac{L\sigma^2}{(K-f)\sqrt{T}} + \frac{C_f}{T} \tag{8}$$

where C_f is a constant depending on Byzantine attack magnitude.

Privacy-Preserving Multi-Modal Selection (PMMS)

Definition 6 (Multi-Modal Quality Function). *For prompt combination $p = (p_v, p_t, p_{3D})$ and dataset D:*

$$q(D, p) = \textit{accuracy}(M(x, p), y) + \lambda \cdot \textit{alignment}(p) \tag{9}$$

Theorem 4 (PMMS Privacy Guarantee). *The PMMS mechanism satisfies (ε, δ)-differential privacy with:*

$$\varepsilon = \frac{2\Delta q}{n} \cdot \log |P| + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{10}$$

where Δq is the global sensitivity of the quality function.

Algorithm 3 Byzantine-Resilient Cross-Modal Aggregation with Privacy-Preserving Selection (BRCMA-PMMS)

```
Require: Client rankings ${r_k}$, privacy parameters $(\varepsilon, \delta)$, prompt set $P$
Ensure: Aggregated prompt parameters $\theta$, privacy-preserving selection
Initialize global prompt parameters $\theta^{(0)}$
for round $t = 1$ to $T$ do
    Collect client rankings $\{r_k^{(t)}\}$ and quality scores $\{q_k^{(t)}\}$
    Apply exponential mechanism to select top prompts:
        $p_{\text{select}}(p) \propto \exp\left(\frac{\varepsilon \cdot q(D,p)}{2\Delta q}\right)$
    Filter out Byzantine clients using cross-modal consistency check:
        $S_t^{\text{valid}} = \{k : \text{consistency}(r_k, \{r_j\}_{j \neq k}) > \tau_{\text{byz}}\}$
    Aggregate valid client rankings using median-based robust estimator
    Add calibrated Gaussian noise for $(\varepsilon, \delta)$-differential privacy
    Update global prompt distribution: $\theta^{(t)} = \text{weighted\_aggregate}(S_t^{\text{valid}})$
end for
return $\theta^{(T)}$
```

4 Comprehensive Theoretical Analysis

Extended Byzantine Tolerance. We relax the standard $f < K/3$ assumption:

Theorem 5 (Adaptive Byzantine Resilience). *Under adaptive adversary model where Byzantine clients can coordinate, BRCMA maintains convergence if:*

$$f < \min\left(\frac{K}{3}, \frac{K \cdot \rho_{honest}}{2 + \rho_{honest}}\right) \tag{11}$$

$$\text{where } \rho_{honest} = \frac{\min_k \|\nabla L_k(\theta^*)\|}{\max_k \|\nabla L_k(\theta^*)\|} \tag{12}$$

Prompt Selection Bias Under Class Imbalance.

Lemma 1 (Imbalance-Aware Prompt Scoring). *For dataset $\mathcal{D}_k$ with class distribution $\pi_k = (\pi_{k,1}, \ldots, \pi_{k,C})$, the bias-corrected prompt quality is:*

$$q_k^{corrected}(p) = q_k(p) - \lambda_{bias} \sum_{c=1}^{C} \pi_{k,c} \log \pi_{k,c} \cdot \mathbb{I}[\textit{prompt } p \textit{ favors class } c]$$

4.1 Communication Complexity Analysis

Theorem 6 (Optimal Prompt Pool Size). *The optimal prompt pool size minimizes the total error:*

$$|P|^* = \arg\min_{|P|} [\varepsilon_{approx}(|P|) + \varepsilon_{comm}(|P|)] \tag{13}$$

$$\text{where } \varepsilon_{approx}(|P|) = \frac{C_{approx}}{|P|^{1/d}} \quad \textit{(approximation error)} \tag{14}$$

$$\varepsilon_{comm}(|P|) = \frac{C_{comm}|P| \log |P|}{B} \quad \textit{(communication error)} \tag{15}$$

and B is the available bandwidth per round.

Solution: $|P|^* = \left(\frac{C_{\text{approx}} d B}{C_{\text{comm}}(d+1)}\right)^{\frac{1}{d+1}}$

Theorem 7 (Communication Complexity). *The FedDental3D-ICL framework achieves communication complexity of $O(K \log |P|)$ per round, compared to $O(K \cdot d)$ for traditional federated learning.*

Proof. In traditional federated learning, each client sends gradient updates of size d (typically 10^9 parameters). Our approach only requires:

- Prompt rankings: $O(|P| \log |P|)$ bits per client
- Alignment statistics: $O(1)$ bits per client
- Quality scores: $O(|P|)$ bits per client

Total per client: $O(|P| \log |P|)$ bits. Since $|P| \ll d$, this represents significant reduction.

4.2 Privacy Analysis

Theorem 8 (Composition-Based Privacy). *Running FedDental3D-ICL for T rounds with parameters $(\varepsilon_t, \delta_t)$ per round satisfies $(\varepsilon_{total}, \delta_{total})$-differential privacy where:*

$$\varepsilon_{total} = \sum_{t=1}^{T} \varepsilon_t + \sqrt{2T \log(1/\delta_{total})} \sum_{t=1}^{T} \varepsilon_t^2 \tag{16}$$

$$\delta_{total} = \sum_{t=1}^{T} \delta_t \tag{17}$$

[Bounded Correlation] Prompt updates across rounds satisfy:

$$\max_{t,t'} |\text{Corr}(r_k^{(t)}, r_k^{(t')})| \leq \rho < 1$$

Theorem 9 (Correlated Composition Privacy). *Under bounded correlation assumption, the total privacy cost after T rounds is:*

$$\varepsilon_{\mathit{total}} \leq \sum_{t=1}^{T} \varepsilon_t + \sqrt{2T \log(1/\delta)} \sqrt{\sum_{t=1}^{T} \varepsilon_t^2} \cdot (1+\rho) \tag{18}$$

$$\delta_{\mathit{total}} \leq \sum_{t=1}^{T} \delta_t \cdot (1 + \rho T) \tag{19}$$

4.3 Convergence Rate Analysis

Theorem 10 (Global Convergence Rate). *The FedDental3D-ICL framework achieves the following convergence rate:*

$$\mathbb{E}[L(\theta^T) - L^*] \leq \frac{C_1}{\sqrt{T}} + \frac{C_2 \sigma^2}{KT} + \frac{C_3 \zeta^2}{T} + \frac{C_4 f}{T} \tag{20}$$

where the constants are defined as:

$$C_1 = L\sqrt{2(L(\theta^0) - L^*)} \quad \textit{(depends on initial suboptimality)} \tag{21}$$

$$C_2 = 2\eta^2 L^2 \quad \textit{(depends on learning rate and smoothness)} \tag{22}$$

$$C_3 = 4\eta L \quad \textit{(heterogeneity impact factor)} \tag{23}$$

$$C_4 = 4\eta L \mathit{\Delta}^2 \quad \textit{(Byzantine attack magnitude)} \tag{24}$$

and the problem parameters are:

$$\sigma^2 = \max_k \mathbb{E}[\|\nabla L_k(\theta) - \nabla \hat{L}_k(\theta)\|^2] \quad \textit{(bounded gradient variance)} \tag{25}$$

$$\zeta^2 = \max_k \mathbb{E}[\|\nabla L_k(\theta^*) - \nabla L(\theta^*)\|^2] \quad \textit{(data heterogeneity)} \tag{26}$$

$$\mathit{\Delta}^2 = \max_{i,j} \|\nabla L_i(\theta) - \nabla L_j(\theta)\|^2 \quad \textit{(Byzantine attack bound)} \tag{27}$$

$$f < \frac{K}{3} \quad \textit{(number of Byzantine clients)} \tag{28}$$

Proof. The proof follows from the convergence analysis of hierarchical multi-modal optimization with Byzantine resilience. The first term $\frac{C_1}{\sqrt{T}}$ captures the standard convergence rate for non-convex optimization, the second term $\frac{C_2\sigma^2}{KT}$ reflects the benefit of averaging across K clients, the third term $\frac{C_3\zeta^2}{T}$ accounts for data heterogeneity across institutions, and the final term $\frac{C_4 f}{T}$ quantifies the impact of Byzantine adversaries.

4.4 Multi-Modal Prompt Space Theory

The foundation of our approach lies in constructing a unified representation space that can seamlessly integrate information from diverse dental imaging modalities. Traditional approaches treat each modality independently, leading to suboptimal integration and missed opportunities for cross-modal reasoning.

Definition 7 (Multi-Modal Prompt Space). *The prompt space $\mathcal{P} = \mathcal{P}_v \times \mathcal{P}_t \times \mathcal{P}_{3D}$ where:*

- $\mathcal{P}_v$*: Visual prompt space for 2D/3D teeth images*
- $\mathcal{P}_t$*: Textual prompt space for clinical descriptions*
- $\mathcal{P}_{3D}$*: Specialized prompt space for 3D volumetric analysis*

For each modality $m \in \{v, t, 3D\}$, we define embedding functions $\phi_m : \mathcal{P}_m \to \mathbb{R}^d$ mapping raw prompts to a common d-dimensional space. The choice of a common embedding dimension is not arbitrary—it reflects the hypothesis that despite their surface differences, medical imaging modalities share fundamental diagnostic patterns that can be captured in a unified representation.

[Lipschitz Continuity] Each embedding function ϕ_m is L-Lipschitz continuous:

$$\|\phi_m(p_1) - \phi_m(p_2)\| \le L\|p_1 - p_2\| \tag{29}$$

Definition 8 (Fusion Function Implementation). *The fusion function $F : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is implemented as:*

$$F(z_v, z_t, z_{3D}) = \sigma(W_1[z_v; z_t; z_{3D}] + b_1) \tag{30}$$

where:

- $W_1 \in \mathbb{R}^{d\times 3d}$*,* $b_1 \in \mathbb{R}^d$ *are learnable parameters*
- σ *is the ReLU activation function*

Theorem 11 (Universal Approximation for MMPS). *For any target prompt combination (p_v^*, p_t^*, p_{3D}^*) and approximation error $\varepsilon > 0$, there exists a fusion function F with $O(\varepsilon^{-d})$ parameters such that:*

$$\|F(\phi_v(p_v), \phi_t(p_t), \phi_{3D}(p_{3D})) - \psi^*(p_v^*, p_t^*, p_{3D}^*)\| \le \varepsilon \tag{31}$$

for appropriately chosen prompts (p_v, p_t, p_{3D}).

Proof (Proof Sketch). By the universal approximation theorem for neural networks, the fusion function F can approximate any continuous mapping between the concatenated embeddings and the target representation with arbitrary precision, provided sufficient network capacity.

Corollary 1 (Stability of MMPS Embeddings). *MMPS embeddings are stable with respect to small perturbations in input prompts:*

$$\|\psi(p_v + \delta_v, p_t + \delta_t, p_{3D} + \delta_{3D}) - \psi(p_v, p_t, p_{3D})\| \le 3L\|F\|_{Lip}(\|\delta_v\| + \|\delta_t\| + \|\delta_{3D}\|) \tag{32}$$

where $\|F\|_{Lip}$ is the Lipschitz constant of the fusion function F.

Remark 1 (Computational Complexity). The fusion function requires $O(d^2)$ operations per forward pass, with memory complexity $O(d^2)$ for storing parameters W_1. For typical embedding dimensions $d \in [256, 1024]$, this represents a computationally tractable approach compared to full model parameter sharing.

4.5 Federated Prompt Optimization Problem

Building upon the multi-modal prompt space foundation, we now formulate the central optimization problem that drives collaborative learning across institutions.

Definition 9 (Federated Multi-Modal Prompt Optimization). *Find optimal prompt parameters $\theta^* = \{p_v^*, p_t^*, p_{3D}^*\}$ that minimize:*

$$\min_\theta L(\theta) = \sum_{k=1}^{K} w_k L_k(\theta; D_k) \tag{33}$$

where $w_k \ge 0$ are client weights with $\sum_{k=1}^K w_k = 1$, and the local loss function incorporates:

$$L_k(\theta; D_k) = L_k^{task}(\theta; D_k) + \lambda L_k^{align}(\theta; D_k) + \gamma L_k^{reg}(\theta) \tag{34}$$

where L_k^{task} is the primary diagnostic task loss, L_k^{align} is the cross-modal alignment loss, and L_k^{reg} is the regularization term.The multi-objective nature of this formulation reflects the complex requirements of dental imaging analysis. The task loss ensures diagnostic accuracy, the alignment loss maintains semantic consistency across modalities, and the regularization term prevents overfitting to institution-specific patterns.

5 Architectural Framework

5.1 Dental System Architecture

We have developed a novel theoretical framework, FedDental3D-ICL, tailored for federated multi-modal learning in dental imaging and diagnostics. Our approach integrates four synergistic components to enable privacy-preserving collaborative learning across dental institutions, enhancing dental care delivery. At the core, we propose a Local Prompt Evaluation Engine, which enables each dental clinic or hospital to evaluate prompts using local multi-modal dental data—such as 3D cone-beam computed tomography (CBCT) scans, intraoral photographs, and clinical dental records—while ensuring patient privacy and compliance with dental regulatory standards. This theoretical engine keeps sensitive dental data within institutional boundaries, yet supports global learning for improved dental diagnosis and treatment planning.

To address the complexities of multi-modal dental AI, we introduce a Cross-Modal Alignment Module that ensures semantic consistency across dental data types, including CBCT scans, panoramic X-rays, intraoral photos, and clinical notes, despite variations in dental imaging equipment and documentation practices across institutions. By leveraging advanced embedding techniques, we create a unified representation space for seamless alignment of dental modalities, critical for accurate diagnosis of conditions like caries, periodontal disease, or orthodontic anomalies. We must clarify that this module is purely theoretical and has not been implemented or tested with real dental imaging data.

We also propose a Hierarchical Optimization Coordinator to manage global prompt distribution tailored to dental diagnostics, while upholding stringent privacy constraints inherent in dental healthcare. This coordinator employs sophisticated algorithms to balance learning efficiency with privacy preservation, preventing leakage of sensitive patient or dental practice data. Additionally, we introduce a Byzantine-Resilient Aggregator, leveraging dental-specific cross-modal validation to defend against malicious participants, ensuring robustness in dental clinical workflows. We emphasize that these components are theoretical constructs, unimplemented and untested in real dental environments.

5.2 Dental Privacy-Preserving Architecture

We designed the FedDental3D-ICL framework to prioritize patient privacy in dental settings through mechanisms tailored to dental practice environments. Our approach ensures that raw dental imaging data, patient records, and clinical notes—such as those detailing restorations, implants, or orthodontic treatments—never leave institutional boundaries, addressing critical privacy and regulatory concerns in dental healthcare. Through careful system design, we process all sensitive dental data locally, sharing only aggregated, anonymized insights about dental pathology patterns and treatment outcomes. We acknowledge that this design remains theoretical, untested with actual dental practice management systems or clinical workflows.

We propose a prompt-only communication protocol, where dental institutions exchange only prompt rankings and encrypted alignment statistics related to dental diagnostic accuracy and treatment planning efficacy. This method significantly reduces privacy risks compared to traditional federated learning approaches, which share model parameters or gradients that could be reverse-engineered to expose sensitive dental patient information. We note that these protocols are unimplemented and lack validation in real dental networking environments.

To ensure robust privacy in dental applications, we integrate differential privacy through strategic noise injection at aggregation points, providing mathematically provable privacy bounds. Our approach guarantees that no individual dental institution or patient's participation can be inferred from global model outputs, even by adversaries with significant computational resources. We also employ secure multi-party computation for critical aggregation steps, enabling collaborative computation without exposing individual dental contributions. We must stress that these cryptographic protocols are theoretical specifications, unimplemented in cryptographic libraries and untested against real-world attack vectors in dental healthcare.

5.3 Dental Scalability Considerations

We designed FedDental3D-ICL with scalability to support large networks of dental institutions, enabling collaborative learning across diverse dental practices. Our theoretical system achieves linear communication scaling with $O(K \log |P|)$ complexity, where K is the number of participating dental clinics and $|P|$ is the dental-specific prompt space size. This bound suggests feasibility for hundreds of dental institutions, from small clinics to large hospitals, though we have not validated this with empirical tests. We acknowledge that these scalability claims are mathematical projections, untested with real dental network conditions or IT infrastructure.

We include adaptive participation mechanisms that dynamically select dental clients based on computational capacity and data quality, such as the resolution of CBCT scans or the detail of clinical notes. We emphasize that these mechanisms are theoretical and have not been implemented or validated in real dental networks.

5.4 Critical Implementation Gap: From Dental Theory to Practice

We present a rigorous theoretical framework with mathematical foundations, convergence guarantees, and privacy-preserving mechanisms tailored for dental federated learning, but we acknowledge that practical implementation remains entirely unaddressed. Our pipeline exists solely on paper, requiring complete development before real-world validation in dental settings.

We acknowledge that our Local Prompt Evaluation Engine, Cross-Modal Alignment Module, Hierarchical Optimization Coordinator, and Byzantine-Resilient Aggregator are purely theoretical constructs, each requiring extensive software development, testing, and optimization for deployment in dental environments. We have not processed real CBCT scans, intraoral photographs, or clinical dental records, leaving our cross-modal alignment mechanisms untested against variations in dental imaging equipment or documentation practices across institutions.

Our differential privacy guarantees and secure multi-party computation protocols require implementation in cryptographic libraries and validation against real attack vectors, which we have not undertaken in dental contexts. We recognize that our $O(K \log |P|)$ complexity bound remains untested under realistic dental network conditions or IT infrastructure.

We admit that our framework lacks validation across all dental aspects. Our theoretical privacy guarantees have not been audited or tested against dental healthcare privacy regulations. Dental practitioners have not interacted with our system, and clinical workflow integration—such as compatibility with electronic dental records—remains unexplored. We have not evaluated our framework against HIPAA, GDPR, or other dental data protection regulations, nor conducted regulatory compliance testing.

5.5 What Has Been Done Versus What Remains Undone in Dental AI

We have established theoretical foundations for federated multi-modal prompt learning in dental AI, providing mathematical frameworks, convergence analysis, and privacy-preserving mechanisms for dental applications. However, we acknowledge that practical implementation, empirical validation, and real-world testing in dental settings are entirely absent. The gap between our theoretical framework and a deployable dental system is vast, with no code written for any FedDental3D-ICL component. We have not processed actual dental imaging data, validated our claimed communication efficiency in real dental network conditions, or implemented our privacy mechanisms as tested security protocols.

No dental practitioner has used or evaluated our system, and we have not attempted integration with dental practice management systems, such as those used for scheduling or treatment planning. Our methodological innovations, including Multi-Modal Prompt Space abstraction, Cross-Modal Prompt Alignment techniques, and Byzantine-Resilient Cross-Modal Aggregation approaches, remain theoretical, unimplemented, and unvalidated in dental practice. We offer no functioning system, no validated performance metrics for dental diagnostics, and no evidence that our innovations work in real dental healthcare environments.

We recognize that our implementation requirements for dental AI include building the entire federated learning infrastructure from scratch, including secure communication protocols, cryptographic implementations, and user interfaces tailored for dental practitioners. We must establish multi-institutional dental data collection through partnerships to provide real CBCT scans, intraoral images, and clinical records for validation. Performance benchmarking requires implementing baseline systems and conducting comparisons to validate our theoretical claims in dental tasks, while security implementation demands developing cryptographic protocols and rigorous testing against dental-specific attack simulations.

We lack a prototype to demonstrate feasibility, performance data to validate efficiency improvements in dental workflows, or real-world testing in dental practice environments. Our theoretical benefits for dental diagnostics remain unproven, with no comparative analysis against existing federated learning approaches using actual dental imaging tasks.

5.6 The Massive Implementation Gap in Dental AI

We acknowledge that our immediate implementation requirements for dental AI represent a substantial undertaking we have not initiated. We must develop the entire federated learning infrastructure, including secure communication protocols, data handling systems, and user interfaces tailored for dental practitioners managing tasks like cavity detection or implant planning. Cryptographic protocol implementation requires developing and testing secure multi-party computation and differential privacy mechanisms beyond our theoretical specifications for dental data. We need multi-modal data processing pipelines to process and align CBCT scans, intraoral photographs, panoramic X-rays, and clinical notes from diverse dental institutions.

We recognize that critical validation studies for dental applications are absent, requiring experiments with real dental imaging datasets from multiple institutions with varied equipment and clinical practices. Network performance testing is needed to validate communication efficiency under realistic dental network conditions. Security and privacy validation demands comprehensive penetration testing, security audits, and privacy impact assessments with dental-specific attack scenarios. Clinical workflow integration testing is essential to validate compatibility with dental practice management systems and practitioner acceptance. Regulatory compliance verification requires navigating dental healthcare regulations and demonstrating compliance through legal and technical assessments.

The reality is that our work presents a sophisticated theoretical framework with transformative potential for dental AI, but it remains entirely theoretical. Every claimed benefit—communication efficiency, privacy preservation, diagnostic accuracy in dental tasks—requires comprehensive implementation and validation before real-world deployment in dental practice. The journey from theory to practical dental AI system is a substantial research and development endeavor that has yet to begin, with our framework serving as a roadmap for what could be built, but the actual construction remaining entirely in the future.

6 Conclusion and Future Directions for Dental AI

We have established foundational mathematical frameworks for adaptive prompt generation, continual learning extensions, and personalization techniques in dental AI applications, but we lack any implementation roadmap, development timeline, or practical steps toward a working dental system. All proposed future research directions, from adaptive prompt generation for dental diagnostics to continual learning for evolving dental practices, build on a theoretical foundation, making them academic exercises until the basic framework is implemented and validated in dental contexts.

The most critical future direction for dental AI is practical implementation, not theoretical extension. We need proof-of-concept development to build a minimal working system demonstrating feasibility in dental settings. Pilot studies with real dental data must conduct small-scale experiments with actual dental institutions to validate core concepts, such as cross-modal alignment for caries detection. Incremental validation should test each component separately before full system integration in dental workflows. Performance reality checks must compare actual performance against theoretical predictions to identify gaps and refinements needed for dental tasks. Clinical validation requires engaging dental practitioners to evaluate the system's utility for real diagnostic and treatment planning tasks, such as planning crowns or orthodontic interventions.

The fundamental gap between theory and practice is our greatest challenge in dental AI. While our research demonstrates the theoretical possibility of achieving collaborative learning benefits with strict privacy and regulatory compliance in dental healthcare, transitioning to practical deployment requires substantial implementation effort we have not undertaken. Our framework provides a roadmap for what could be built for dental practice, but the actual building—and the inevitable practical challenges theory cannot predict—remains entirely in the future. Every claim of potential benefit for dental diagnostics must be validated through rigorous implementation and empirical testing before our theoretical contribution can impact dental practice, transforming it from an academic exercise into a practical tool to advance dental healthcare through artificial intelligence.

So our work establishes a rigorous theoretical foundation for federated multi-modal dental imaging, but empirical validation is crucial. We are actively collaborating with dental care institutions across South Asian countries, specifically to collect CBCT scans along with panoramic X-rays, intraoral photographs, and clinical notes. This collaboration will enable us to benchmark our framework against existing federated learning approaches, evaluate privacy guarantees in realistic settings, and explore integration with clinical workflows. Alongside, we plan to develop a prototype implementation, paving the way toward multi-institutional deployment and practical impact in dental healthcare.

Acknowledgements

We gratefully acknowledge the support and guidance received during our COM S 4590X: Security and Privacy in Cloud Computing final project at Iowa State University, which made this work possible. We would also like to express our appreciation to the Translational AI Center as a token of gratitude for their encouragement and support throughout this research.

References

1. Z. Zhang et al., "BioMedGPT: Unified and Generalist Biomedical Foundation Model Bridging Vision, Language, and Multimodal Tasks," *arXiv preprint arXiv:2305.17153*, 2023.

2. M. Eslami et al., “PubMedCLIP: A Contrastive Vision-Language Pre-training for Biomedical Vision-Language Processing,” *arXiv preprint arXiv:2112.10683*, 2021.
3. HIMSS, “What is Federal Health IT Policy?” HIMSS Knowledge Center, 2024.
4. ONC, “Federal Health IT Strategic Plan 2020-2025,” Office of the National Coordinator for Health IT, 2023.
5. M. J. Sheller et al., “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” *Scientific Reports*, vol. 10, no. 1, pp. 12598, 2020.
6. A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” *arXiv preprint arXiv:2103.00020*, 2021.
7. M. Abadi et al., “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.
8. L. Zhu et al., “Deep Leakage from Gradients,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
9. L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting Unintended Feature Leakage in Collaborative Learning,” in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 691–706.
10. R. C. Geyer, T. Klein, and M. Nabi, “Differentially private federated learning: A client level perspective,” *arXiv preprint arXiv:1712.07557*, 2017.
11. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
12. X. Li et al., “Federated Learning on Non-IID Data Silos: An Experimental Study,” in *International Conference on Learning Representations*, 2022.
13. P. Kairouz et al., “Advances and Open Problems in Federated Learning,” *Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1–210, 2021.
14. F. Sattler, S. Wiedemann, K. R. Müller, and W. Samek, “Robust and Communication-Efficient Federated Learning From Non-I.I.D. Data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2019.
15. P. Blanchard, E. M. el Mhamdi, R. Guerraoui, and J. Stainer, “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
16. D. Yin, Y. Chen, R. Kannan, and P. Bartlett, “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates,” *arXiv preprint arXiv:1803.01498*, 2018.
17. E. Bagdasaryan et al., “How To Backdoor Federated Learning,” in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 2938–2948.