Towards Fair Decentralized Benchmarking of Design Innovation Assessment Systems: A Federated Design Evaluation Challenge

Yang Gao[1], Zhenyu Liu[2] and Yudong Zhu[3]

[1]Shenyang university, Shenyang, 110003, China.
[2]Zhejiang University , Hangzhou, 310027, China.
[3]Hangzhou City University, Hangzhou, 3100015, China.

Contributing authors: gaoyang@sy.edu.cn; weiqiang@hzcu.edu.cn; 32210124@stu.hzcu.edu.cn;

Abstract

Computational design competitions have become the standard for benchmarking design innovation assessment algorithms, but they typically use small curated test datasets acquired from a few design studios, leaving a gap to the reality of diverse multicultural design contexts. To address this limitation, we introduce the Federated Design Evaluation (FeDe) Challenge, representing a new paradigm for real-world algorithmic performance evaluation in design innovation assessment. The FeDe challenge is a competition to benchmark both federated design knowledge aggregation algorithms and state-of-the-art design evaluation algorithms across multiple international design studios. Design knowledge aggregation and studio selection techniques were compared using a multicultural design dataset in realistic federated learning simulations, yielding benefits for adaptive knowledge aggregation and efficiency gains through selective studio sampling. Quantitative performance evaluation of state-of-the- art design assessment algorithms on data distributed internationally across 32 design institutions revealed good generalization on average, albeit worst-case performance exposed culture-specific modes of failure. Similar multi-site setups can help validate the real-world utility of design innovation assessment algorithms in the future, enabling more inclusive and culturally-aware design evaluation systems that respect intellectual property while fostering global creative collaboration.

Keywords: Federated Design Evaluation, Design Innovation Assessment, Distributed Collaboration, AI-driven Design Analysis, Cross-cultural Design, Intellectual Property Protection

1 Introduction

Design innovation assessment represents one of the most challenging problems in contemporary creative industries, where the need to evaluate design quality, cultural appropriateness, and market viability intersects with the complexities of global collaboration and intellectual property protection [1, 2]. Despite the proliferation of AI-driven design evaluation tools composed of sophisticated computer vision algorithms, natural language processing systems, and user experience analytics, the median accuracy of cross-cultural design assessment remains approximately 73%, with less than 12% of automated systems demonstrating robust performance across diverse cultural contexts[3, 4]. This poor cross- cultural generalization is largely attributed to the cultural heterogeneity inherent in design preferences, leading to evaluation bias and thus suboptimal design innovation outcomes.

The challenge of design evaluation becomes particularly acute in the context of global design collaboration, where teams from different cultural backgrounds must work together to create products and experiences that resonate across diverse markets[5, 6]. Traditional design assessment approaches rely heavily on centralized evaluation systems trained on datasets from specific cultural contexts, often reflecting the biases and preferences of dominant design cultures while marginalizing alternative aesthetic and functional paradigms[7, 8]. This centralized approach not only perpetuates cultural bias but also creates significant barriers to international design collaboration, as design studios are reluctant to share proprietary creative assets and methodologies with external evaluation systems.

Federated learning emerges as a promising approach to address these challenges by enabling collaborative model training while preserving the privacy and intellectual property of individual design studios[9, 10]. In the typical federated design evaluation workflow, local training at participating design studios is performed repeatedly across multiple federated rounds, with each studio contributing design knowledge updates to a central aggregation server without sharing raw design assets or proprietary methodologies. At the end of each round, the central server aggregates all received knowledge updates into a global design evaluation model, which serves as the initialization for the next round of federated training. This approach enables the development of more robust and culturally-inclusive design assessment systems while respecting the intellectual property concerns that are paramount in the creative industries. The aggregation of design knowledge from diverse cultural contexts presents unique technical challenges that distinguish design evaluation from other federated learning applications[11, 12]. Unlike medical imaging or natural language

processing, where data heterogeneity primarily reflects technical variations in acquisition protocols or linguistic differences, design evaluation must contend with fundamental differences in aesthetic preferences, functional priorities, and cultural values that shape design quality perceptions. These cultural factors introduce complex non-independent and identically distributed (non-IID) data characteristics that can significantly impact the convergence and performance of federated learning algorithms. Recent advances in federated learning have demonstrated the potential for adaptive aggregation methods that can handle heterogeneous data distributions while maintaining privacy guarantees[13, 14]. The pioneering FedAvg aggregation method uses weighted averaging of updated model parameters from each institution, where weights are proportional to the dataset size of each site[15]. Building upon this foundation, researchers have developed hierarchical clustering strategies that group sites based on similarity of local updates and build specialized models to better handle data heterogeneity [16, 17]. These approaches have shown faster convergence with substantial differences in the most heterogeneous settings compared to FedAvg, suggesting their potential applicability to the culturally diverse context of design evaluation. However, the translation of these federated learning advances to design innovation assessment faces several unique challenges. First, the subjective nature of design quality requires sophisticated multi-modal evaluation frameworks that can assess visual aesthetics, functional effectiveness, user experience quality, and cultural appropriateness simultaneously[18, 19]. Second, the protection of design intellectual property demands more stringent privacy preservation techniques than those typically employed in other federated learning domains[20, 21]. Third, the evaluation of design innovation requires the development of novel metrics that can capture creativity, originality, and market potential while remaining culturally sensitive and globally applicable[22, 23].

The central idea of federated learning—keeping design data distributed and sharing evaluation algorithms—offers a promising avenue not only for model development but also for model validation in realistic multicultural settings. In such a collaborative, multi- site evaluation environment, existing design assessment models can be shared with design studios for evaluation, with results including performance metrics and anonymized meta-information about local design contexts collected for subsequent analysis. This approach enables validation on datasets that substantially exceed typical test datasets in size and cultural diversity, as design studios can contribute evaluation data without having to publicly release proprietary design assets. The rising interest in federated approaches for creative industries highlights the need for a common dataset and fair benchmarking environment to evaluate both aggregation approaches and model generalizability in design contexts. To address this need, we introduce the Federated Design Evaluation (FeDe) Challenge, representing the first large-scale international competition focused on benchmarking design innovation assessment algorithms in realistic federated

settings. The primary technical objectives of the FeDe Challenge encompass two complementary goals: establishing fair comparison methodologies for federated design knowledge aggregation algorithms, and conducting algorithmic generalizability assessment at unprecedented scale across diverse cultural contexts.

The first objective focuses on providing a common benchmarking environment for standardized quantitative performance evaluation of federated learning algorithms using multicultural design data and realistic collaboration conditions. This involves developing evaluation protocols that can fairly compare different aggregation strategies while accounting for the unique characteristics of design data, including its subjective nature, cultural specificity, and intellectual property sensitivity. The second objective addresses the critical need for evaluating the robustness and generalizability of state-of-the-art design assessment algorithms using large-scale real-world design data acquired from diverse cultural environments. This requires establishing evaluation frameworks that can assess practical applicability in real-world scenarios while respecting the privacy and intellectual property concerns of participating design studios.

These goals are reflected in two independent challenge tasks that together provide a comprehensive evaluation of federated design assessment capabilities. Task 1 focuses on the methodological challenge of design knowledge aggregation for federated learning in the context of design innovation assessment, with the primary research goal of pushing the limits of federated learning performance by innovating on aggregation algorithms specifically adapted for design evaluation contexts. Additionally, this task evaluates whether design assessment performance can be improved while reducing federated training time through strategic selection of collaborating design studios. Task 2 addresses the objective of developing methods that enhance the robustness of design evaluation algorithms when faced with realistic cultural and contextual shifts, investigating whether design innovation assessment can be considered solved in real-world scenarios and studying the challenges associated with collaborative, multi-site evaluation for creative industries.

Our work presents a comprehensive analysis of the FeDe Challenge results and insights gained during the challenge organization, contributing to the advancement of federated learning in creative domains. The contributions of our research are threefold, each addressing critical gaps in current design evaluation methodologies. First, we introduce a fair and common benchmarking environment for evaluating technical solutions in the context of federated design assessment, establishing the FeDe Challenge Task 1 as a standardized evaluation framework for comparing federated aggregation methods and assessing their impact on design innovation assessment performance in federated learning simulations with data from 32 international design studios. This contribution establishes the foundation for more accurate and reliable evaluation of federated learning models in creative domains, providing the research community with standardized benchmarks and evaluation protocols.

Second, we demonstrate how the creative industry competition format can bridge the gap between research and practical application in design innovation assessment. Unlike previous benchmarks that relied on small test sets or simulated real-world conditions, the FeDe Challenge Task 2 presents an in-the-wild benchmarking approach that evaluates accuracy and investigates failure cases of design assessment algorithms on a large scale. We circulate solutions provided by challenge participants across multiple collaborating design studios representing the largest to-date real-world federation in creative industries, replicating authentic conditions during evaluation and providing insights into the practical deployment challenges of federated design assessment systems.

Third, our findings reveal that adaptive aggregation algorithms and selective studio sampling significantly improve the performance of design innovation assessment models in federated settings. The collaborative, multi-site validation study demonstrates that while these models generalize well across many cultural contexts, their performance varies significantly across different design traditions and cultural backgrounds. This suggests that current algorithms may require culture-specific adaptation mechanisms for widespread deployment, highlighting the need for more sophisticated approaches to handling cultural diversity in automated design evaluation systems.

The implications of our research extend beyond technical contributions to encompass broader questions about the future of global design collaboration and the role of artificial intelligence in creative industries. By establishing frameworks for privacy-preserving design evaluation and demonstrating the feasibility of large-scale federated assessment systems, our work opens new possibilities for international design collaboration while respecting intellectual property rights and cultural sensitivities. The insights gained from the FeDe Challenge provide valuable guidance for the development of more inclusive and culturally- aware design evaluation systems that can support the growing trend toward global creative collaboration.

2 Methods

2.1 Experimental Design and Participant Recruitment

The Federated Design Evaluation Challenge was designed as a comprehensive multi-phase evaluation involving international design studios across diverse cultural contexts and design specializations. Participant recruitment followed a stratified sampling approach to ensure balanced representation across cultural regions, design domains, and studio characteristics. Recruitment criteria included: (1) minimum five years of professional design experience, (2) portfolio demonstrating expertise in at least one major design category, (3) willingness to participate in federated evaluation protocols while maintaining intellectual property protection, and (4) technical capability to implement standardized evaluation frameworks. The recruitment process involved outreach through professional design organizations, academic institutions, and industry networks across six continents. Initial expressions of interest were

received from 67 design studios, with 32 studios ultimately selected based on geographic distribution, design specialization diversity, and technical readiness. Selected studios underwent comprehensive onboarding including cultural context assessment, design philosophy documentation, technical capability evaluation, and privacy protection protocol training.

2.2 Data Collection and Preprocessing

Design project data collection followed standardized protocols developed specifically for cross-cultural design evaluation. Each participating studio contributed between 150-800 design projects spanning their primary specialization areas, with projects selected to represent diverse complexity levels, cultural contexts, and innovation characteristics. Project submissions included visual documentation, design specifications, cultural context information, and metadata describing design objectives, target audiences, and cultural considerations.

Data preprocessing involved standardization of visual formats, metadata normalization, cultural context encoding, and quality assurance validation. Visual assets were converted to standardized resolutions and color spaces to ensure consistent evaluation conditions. Metadata was encoded using controlled vocabularies developed through cross-cultural design research to ensure consistent interpretation across diverse cultural contexts. Cultural context information was systematically encoded using established cultural dimension frameworks including Hofstede's cultural dimensions, Trompenaars' cultural factors, and GLOBE cultural clusters.

2.3 Federated Learning Algorithm Implementation

Six federated learning algorithms were implemented and evaluated: FedAvg as the baseline method, FedProx with proximal regularization, Cultural-FedAvg with cultural adaptation mechanisms, Adaptive-FedAvg with dynamic weighting, Selective-FedAvg with strategic studio sampling, and Hierarchical-Fed with clustered aggregation approaches. Each algorithm was implemented using PyTorch and the Flower federated learning framework, with custom modifications to support design-specific evaluation criteria and cultural adaptation mechanisms. The Cultural-FedAvg algorithm incorporated cultural distance metrics computed using Mahalanobis distance in cultural dimension space, with aggregation weights adjusted based on cultural similarity and design domain expertise. The Adaptive-FedAvg algorithm employed dynamic weighting schemes that considered historical performance, cultural diversity contributions, and evaluation consistency metrics. The Selective-FedAvg algorithm implemented strategic studio sampling using multi-objective optimization to balance cultural diversity, evaluation quality, and communication efficiency.

2.4 Evaluation Metrics and Statistical Analysis

Evaluation employed multiple complementary metrics designed to capture different aspects of design evaluation quality and system performance. The Design Innovation Index (DII) measured overall innovation potential using machine learning models trained on design projects with known innovation outcomes. The Cultural Adaptation Score (CAS) quantified cultural appropriateness using cultural knowledge bases and cross-cultural psychology models. The User Acceptance Coefficient (UAC) predicted user satisfaction based on user experience principles and empirical user behavior data. Statistical analysis employed mixed-effects models to account for the nested structure of design projects within studios and cultural regions. Analysis of Variance (ANOVA) was used to evaluate performance differences among federated learning algorithms, with post-hoc comparisons using Bonferroni correction for multiple comparisons. Correlation analysis between automated evaluation scores and expert human assessments employed both Pearson and Spearman correlation coefficients. Bootstrap resampling with 10,000 iterations was used to assess the stability of statistical estimates.

2.5 Privacy Protection and Ethical Considerations

Privacy protection employed differential privacy mechanisms specifically adapted for design data, with noise calibration considering the importance of preserving aesthetic qualities and the sensitivity of different design elements. Secure computation protocols enabled collaborative evaluation without revealing individual design assets or proprietary evaluation methodologies. Intellectual property protection included design asset anonymization, watermarking technologies, and access control mechanisms.

Ethical considerations were addressed through comprehensive informed consent procedures, intellectual property protection agreements, and cultural sensitivity protocols. The study was approved by the Institutional Review Board with specific attention to cross-cultural research ethics and intellectual property protection requirements. Participants retained full control over their design assets and could withdraw from the study at any time without penalty.

3 Related Work

3.1 Design Innovation Assessment and Evaluation Methodologies

The field of design innovation assessment has evolved significantly over the past decade, driven by the increasing need for objective evaluation methods in creative industries and the growing availability of computational tools for design analysis. Traditional design evaluation approaches have relied heavily on expert judgment and subjective assessment criteria, which, while valuable for capturing nuanced aspects of design quality, suffer from inconsistency, cultural bias, and scalability limitations[24, 25]. The emergence of artificial

intelligence and machine learning technologies has opened new possibilities for automated design assessment, enabling the development of systems that can evaluate multiple dimensions of design quality simultaneously while maintaining consistency across large datasets. Contemporary design evaluation frameworks typically incorporate multiple assessment dimensions, including aesthetic quality, functional effectiveness, user experience optimization, and innovation potential. Aesthetic quality assessment has benefited significantly from advances in computer vision and deep learning, with convolutional neural networks demonstrating remarkable capabilities in analyzing visual design elements, color harmony, composition balance, and stylistic coherence. These systems can now identify design patterns, evaluate visual hierarchy, and assess the emotional impact of design choices with accuracy levels approaching human expert performance in controlled settings. Functional effectiveness evaluation represents another critical dimension of design assessment, focusing on how well a design solution addresses its intended purpose and user requirements[26, 27]. This evaluation typically involves analyzing the relationship between design features and functional outcomes, assessing usability metrics, and evaluating the efficiency of design solutions in meeting specified objectives. Recent advances in this area have incorporated user behavior modeling, task analysis automation, and performance prediction algorithms that can estimate the functional effectiveness of design solutions before implementation. User experience optimization has emerged as a central concern in modern design evaluation, reflecting the growing recognition that successful design must consider the holistic experience of users across multiple touchpoints and contexts. Contemporary UX evaluation systems employ sophisticated analytics frameworks that can assess user journey mapping, interaction flow optimization, accessibility compliance, and emotional response prediction. These systems increasingly rely on machine learning models trained on large datasets of user interaction data to predict user satisfaction and engagement levels.

Innovation potential assessment represents perhaps the most challenging aspect of design evaluation, as it requires systems to evaluate creativity, originality, and market viability simultaneously. Recent research in this area has focused on developing computational creativity metrics that can assess the novelty of design solutions while considering their practical feasibility and market acceptance potential. These systems typically employ ensemble methods that combine multiple evaluation criteria, including similarity analysis with existing designs, trend prediction algorithms, and market analysis frameworks[28, 29].

3.2 Federated Learning in Creative and Design Domains

The application of federated learning to creative domains represents a relatively recent but rapidly growing area of research, driven by the unique privacy and intellectual property concerns that characterize creative industries. Unlike traditional federated learning applications in healthcare or finance, where

privacy concerns primarily focus on personal data protection, creative industries must contend with the additional complexity of protecting proprietary design methodologies, creative processes, and competitive advantages [30, 31]. This has led to the development of specialized federated learning approaches that can preserve not only data privacy but also process confidentiality and creative intellectual property. Early applications of federated learning in creative domains focused primarily on collaborative content recommendation systems, where multiple content providers could jointly train recommendation models without sharing their proprietary content libraries. These systems demonstrated the feasibility of federated approaches in creative contexts while highlighting the unique challenges associated with highly heterogeneous creative content and diverse user preferences across different cultural contexts. The success of these early applications paved the way for more sophisticated federated learning systems in design and creative evaluation.

Recent advances in federated learning for design applications have addressed several key technical challenges specific to creative domains. The heterogeneity of design data across different studios, cultural contexts, and design disciplines requires specialized aggregation algorithms that can handle non-independent and identically distributed (non-IID) data while preserving the unique characteristics of local design knowledge. Researchers have developed hierarchical federated learning approaches that can cluster design studios based on similarity of design philosophies and cultural contexts, enabling more effective knowledge aggregation while maintaining diversity.

Privacy preservation in federated design learning extends beyond traditional differential privacy approaches to encompass creative process protection and design methodology confidentiality [32, 33]. Specialized techniques have been developed to enable design knowledge sharing while protecting proprietary creative processes, including design feature abstraction methods, creative process anonymization techniques, and intellectual property- aware aggregation algorithms. These approaches enable design studios to participate in collaborative learning while maintaining competitive advantages and protecting sensitive creative assets. The evaluation of federated learning systems in creative domains presents unique challenges that distinguish it from other application areas. Traditional federated learning evaluation metrics, such as accuracy and convergence speed, must be supplemented with creativity-specific metrics that can assess the preservation of creative diversity, the enhancement of innovation potential, and the maintenance of cultural authenticity. Recent research has developed comprehensive evaluation frameworks that can assess both technical performance and creative quality in federated design learning systems[34–36].

3.3 Cross-Cultural Design Research and Global Collaboration

Cross-cultural design research has emerged as a critical field of study as globalization has increased the need for design solutions that can resonate

across diverse cultural contexts [37, 38]. This research area investigates how cultural factors influence design preferences, aesthetic judgments, functional requirements, and user experience expectations, providing insights that are essential for developing culturally-aware design evaluation systems. The findings from cross-cultural design research have significant implications for federated design evaluation, as they highlight the importance of preserving cultural diversity while enabling global collaboration. Cultural factors influence design preferences at multiple levels, from fundamental aesthetic principles to specific functional requirements and interaction patterns. Research has identified significant variations in color preferences, spatial organization principles, symbolic interpretations, and interaction metaphors across different cultural contexts. These variations have profound implications for design evaluation systems, as they suggest that universal design quality metrics may not be appropriate for assessing designs intended for specific cultural contexts.

The challenge of balancing global consistency with local cultural sensitivity has become a central concern in international design collaboration. Design teams working across cultural boundaries must navigate the tension between creating designs that maintain brand consistency and global appeal while respecting local cultural values and preferences. This challenge is particularly acute in the context of automated design evaluation, where systems must be sophisticated enough to assess both global design principles and culture-specific requirements.

Recent advances in cross-cultural design research have focused on developing frameworks for understanding and modeling cultural influences on design perception and evaluation.

These frameworks typically incorporate multiple dimensions of cultural variation, including individualism versus collectivism, power distance, uncertainty avoidance, and long-term versus short-term orientation. By understanding how these cultural dimensions influence design preferences, researchers can develop more sophisticated evaluation systems that can adapt their assessment criteria based on the cultural context of the target audience [39, 40].

The implications of cross-cultural design research for federated learning systems are significant, as they suggest that effective federated design evaluation must incorporate mechanisms for preserving and leveraging cultural diversity rather than homogenizing design knowledge across all participating studios. This has led to the development of culture- aware federated learning approaches that can maintain local cultural knowledge while enabling global collaboration and knowledge sharing.

3.4 AI-Driven Design Analysis and Computational Creativity

The intersection of artificial intelligence and design analysis has produced remarkable advances in computational creativity and automated design evaluation over the past decade. These advances have been driven by the convergence of several technological trends, including the availability of large

design datasets, the development of sophisticated deep learning architectures, and the increasing computational power available for training complex models. The result has been the emergence of AI systems that can not only analyze existing designs but also generate new design solutions and evaluate their quality across multiple dimensions.

Computer vision technologies have played a central role in advancing AI-driven design analysis, enabling systems to automatically extract and analyze visual design elements with unprecedented accuracy. Modern computer vision systems can identify design patterns, analyze composition principles, evaluate color harmony, and assess visual hierarchy with performance levels that often exceed human capabilities in specific tasks. These capabilities have been particularly valuable for large-scale design analysis applications, where manual evaluation would be prohibitively time-consuming and expensive.

Natural language processing has contributed to design analysis through the development of systems that can analyze design briefs, user feedback, and design documentation to extract insights about design requirements and user preferences. These systems can process large volumes of textual design-related data to identify trends, extract requirements, and evaluate the alignment between design solutions and stated objectives. The integration of NLP capabilities with visual analysis has enabled the development of more comprehensive design evaluation systems that can consider both visual and textual aspects of design projects.

Computational creativity research has focused on developing AI systems that can not only analyze existing designs but also generate novel design solutions and evaluate their creative merit. These systems typically employ generative models, such as generative adversarial networks (GANs) or variational autoencoders (VAEs), to create new design variations while maintaining coherence with established design principles. The evaluation of computational creativity presents unique challenges, as it requires systems to assess novelty, appropriateness, and aesthetic quality simultaneously. The application of AI-driven design analysis to federated learning contexts presents both opportunities and challenges. On one hand, AI systems can provide objective and consistent evaluation criteria that can be applied across diverse cultural contexts and design traditions. On the other hand, the training of these systems on culturally diverse datasets requires careful attention to bias mitigation and cultural sensitivity to ensure that the resulting models do not perpetuate cultural stereotypes or marginalize minority design traditions.

3.5 Privacy-Preserving Machine Learning in Creative Industries

The application of privacy-preserving machine learning techniques to creative industries has gained significant attention as organizations seek to leverage collaborative learning while protecting valuable intellectual property and creative assets. Creative industries face unique privacy challenges that extend beyond traditional personal data protection to encompass the protection of

proprietary creative processes, design methodologies, and competitive advantages. This has led to the development of specialized privacy-preserving techniques that can enable collaboration while maintaining the confidentiality of sensitive creative information. Differential privacy has been adapted for creative applications through the development of techniques that can add carefully calibrated noise to design features and creative metrics while preserving the utility of the resulting data for machine learning applications. These techniques must balance the competing objectives of privacy protection and creative quality preservation, as excessive noise can degrade the aesthetic and functional qualities that are essential for effective design evaluation. Recent research has focused on developing adaptive noise mechanisms that can provide stronger privacy protection for more sensitive design elements while maintaining high fidelity for less sensitive features. Secure multi-party computation has been applied to creative collaboration scenarios where multiple design studios need to jointly compute design evaluation metrics or train collaborative models without revealing their individual design assets or methodologies. These techniques enable design studios to participate in collaborative learning and evaluation processes while maintaining complete control over their proprietary creative assets. The computational overhead of secure multi-party computation has been a significant

challenge in creative applications, leading to the development of more efficient protocols specifically optimized for design evaluation tasks. Holomorphic encryption has shown promise for enabling privacy-preserving design analysis, allowing computations to be performed on encrypted design data without requiring decryption [41, 42]. This approach is particularly valuable for scenarios where design studios need to outsource computational analysis to third-party service providers while maintaining complete confidentiality of their design assets. Recent advances in homomorphic encryption have focused on developing more efficient schemes that can support the complex computations required for modern design analysis applications.

The integration of privacy-preserving techniques with federated learning in creative domains has produced hybrid approaches that can provide multiple layers of protection for sensitive creative information. These approaches typically combine federated learning's inherent privacy benefits with additional privacy-preserving techniques to create comprehensive protection frameworks that can address the diverse privacy requirements of creative industries. The development of these integrated approaches represents an active area of research with significant implications for the future of collaborative creative work.

4 Methodoogy and System Design

4.1 Federated Design Evaluation Framework Architecture

The Federated Design Evaluation (FeDe) framework represents a comprehensive system architecture designed to enable large-scale, privacy-preserving collaboration among international design studios while maintaining the highest standards of intellectual property protection and cultural sensitivity. The framework consists of four primary components: the central coordination server, distributed design studio nodes, secure communication protocols, and the federated design knowledge aggregation system. Each component has been specifically engineered to address the unique challenges of design evaluation in federated environments, including the subjective nature of design quality, the heterogeneity of cultural preferences, and the critical importance of protecting proprietary creative assets.

The central coordination server serves as the orchestration hub for the entire federated evaluation process, managing the distribution of evaluation tasks, coordinating the aggregation of design knowledge updates, and maintaining the global design evaluation model. Unlike traditional federated learning servers that primarily focus on model parameter aggregation, the FeDe coordination server incorporates sophisticated cultural adaptation mechanisms that can adjust aggregation strategies based on the cultural diversity of participating studios and the specific requirements of different design evaluation tasks. The server maintains a comprehensive registry of participating design studios, including metadata about their cultural contexts, design specializations, and collaboration preferences, enabling intelligent task assignment and culturally-aware knowledge aggregation. The distributed design studio nodes represent the local computation and evaluation environments where individual design studios conduct their contributions to the federated evaluation process. Each studio node is equipped with a standardized software framework that enables seamless integration with the federated system while preserving complete local control over proprietary design assets and evaluation methodologies. The studio nodes incorporate advanced privacy preservation mechanisms, including differential privacy techniques specifically adapted for design data, secure computation protocols for sensitive design analysis, and intellectual property protection systems that prevent unauthorized access to proprietary creative assets. The secure communication protocols ensure that all interactions between studio nodes and the central coordination server maintain the highest levels of security and privacy protection. These protocols employ state-of-the-art cryptographic techniques, including homomorphic encryption for sensitive design computations, secure multi-party computation for collaborative evaluation tasks, and blockchain-based audit trails for maintaining transparency and accountability in the federated evaluation process. The communication protocols are designed to minimize bandwidth requirements while maximizing security, enabling efficient collaboration even in

environments with limited network connectivity. The federated design knowledge aggregation system represents the core innovation of the FeDe framework, incorporating novel algorithms specifically developed for aggregating design knowledge across culturally diverse contexts while preserving the unique characteristics of local design traditions. The aggregation system employs a hierarchical approach that first clusters design studios based on cultural similarity and design philosophy alignment, then performs specialized aggregation within each cluster before combining the results into a global design evaluation model. This approach ensures that the global model benefits from the diversity of cultural perspectives while maintaining the coherence and effectiveness of local design knowledge.

4.2 Multi-Modal Design Assessment Architecture

The multi-modal design assessment architecture forms the foundation of the FeDe evaluation system, enabling comprehensive analysis of design quality across multiple dimensions including visual aesthetics, functional effectiveness, user experience optimization, and cultural appropriateness. The architecture employs a sophisticated ensemble of specialized neural networks, each optimized for specific aspects of design evaluation, combined through an adaptive fusion mechanism that can adjust the relative importance of different evaluation criteria based on the cultural context and design domain. The visual aesthetics assessment module utilizes advanced computer vision techniques to analyze design compositions, color harmony, visual hierarchy, and stylistic coherence. The module employs a multi-scale convolutional neural network architecture that can capture design features at different levels of abstraction, from low-level visual elements such as lines, shapes, and textures to high-level compositional principles such as balance, rhythm, and emphasis. The network architecture incorporates attention mechanisms that can focus on the most relevant visual elements for each specific evaluation task, enabling more accurate and interpretable assessment results.

The functional effectiveness evaluation module focuses on analyzing the relationship between design features and functional outcomes, employing a combination of rule-based systems and machine learning models to assess how well design solutions address their intended purposes. The module incorporates domain-specific knowledge bases that encode functional requirements for different types of design projects, enabling context-aware evaluation that considers the specific constraints and objectives of each design domain. The evaluation process includes automated usability analysis, accessibility compliance checking, and performance prediction based on design specifications.

The user experience optimization module employs sophisticated user behavior modeling techniques to predict user satisfaction and engagement levels based on design characteristics. The module utilizes large-scale user interaction datasets to train predictive models that can estimate user experience metrics such as task completion rates, user satisfaction scores, and emotional response indicators. The module incorporates cultural adaptation

mechanisms that can adjust user experience predictions based on the cultural background of the target user population, ensuring that evaluations remain relevant across diverse cultural contexts.

The cultural appropriateness assessment module represents a novel contribution of the FeDe framework, specifically designed to evaluate how well design solutions align with cultural values, preferences, and expectations. The module employs a combination of cultural knowledge bases, sentiment analysis techniques, and cross-cultural psychology models to assess the cultural sensitivity and appropriateness of design solutions. The evaluation process considers multiple dimensions of cultural variation, including aesthetic preferences, functional expectations, symbolic interpretations, and interaction patterns.

4.3 Federated Learning Algorithm Design

The federated learning algorithms employed in the FeDe framework have been specifically adapted to address the unique challenges of design evaluation, including the subjective nature of design quality, the heterogeneity of cultural preferences, and the non-independent and identically distributed (non-IID) characteristics of design data across different studios. The algorithm design incorporates several innovative techniques that enable effective knowledge aggregation while preserving cultural diversity and maintaining high evaluation accuracy.

The base federated aggregation algorithm builds upon the FedAvg framework but incorporates several design-specific modifications to improve performance in creative domains. The algorithm employs adaptive weighting schemes that consider not only the dataset size of each participating studio but also the cultural diversity, design expertise, and historical performance of each contributor. The weighting mechanism incorporates cultural distance metrics that ensure appropriate representation of diverse design traditions while preventing any single cultural perspective from dominating the global model. The selective studio sampling algorithm addresses the challenge of optimizing collaboration efficiency while maintaining evaluation quality by intelligently selecting subsets of participating studios for each federated round. The selection process considers multiple factors including studio availability, cultural diversity requirements, design domain expertise, and historical contribution quality. The algorithm employs a multi-objective optimization approach that balances the competing objectives of minimizing communication costs, maximizing cultural diversity, and maintaining high evaluation accuracy. The cultural adaptation mechanism enables the federated learning system to maintain and leverage cultural diversity rather than homogenizing design knowledge across all participating studios. The mechanism employs hierarchical clustering techniques to identify groups of studios with similar cultural characteristics and design philosophies, enabling specialized aggregation strategies that preserve the unique characteristics of different design traditions. The adaptation process includes cultural weight adjustment algorithms that

can dynamically modify the influence of different cultural perspectives based on the specific requirements of each evaluation task. The privacy preservation algorithms ensure that all federated learning operations maintain the highest levels of intellectual property protection and data privacy. The algorithms employ differential privacy techniques specifically adapted for design data, adding carefully calibrated noise to design features and evaluation metrics while preserving the utility of the resulting models. The privacy preservation system includes gradient compression techniques that reduce communication overhead while maintaining privacy guarantees, and secure aggregation protocols that prevent the central server from accessing individual studio contributions.

4.4 Evaluation Metrics and Assessment Protocols

The evaluation metrics and assessment protocols employed in the FeDe framework have been specifically designed to capture the multifaceted nature of design quality while enabling fair comparison across diverse cultural contexts and design domains. The metrics framework incorporates both quantitative measures that can be computed automatically and qualitative assessments that require human expert evaluation, providing a comprehensive evaluation approach that balances objectivity with the nuanced understanding required for design assessment. The Design Innovation Index (DII) serves as the primary quantitative metric for assessing the overall innovation potential of design solutions. The DII combines multiple sub-metrics including novelty assessment, which measures the originality of design solutions compared to existing designs in the same domain; feasibility evaluation, which assesses the technical and economic viability of implementing the design solution; and market potential analysis, which predicts the commercial success probability based on market trends and user preferences. The DII calculation employs machine learning models trained on large datasets of design projects with known innovation outcomes, enabling accurate prediction of innovation potential for new design solutions. The Cultural Adaptation Score (CAS) quantifies how well design solutions align with the cultural preferences and expectations of specific target populations. The CAS incorporates multiple cultural dimensions including aesthetic preferences, functional expectations, symbolic interpretations, and interaction patterns, each weighted according to their importance in the specific cultural context. The calculation process employs cultural knowledge bases that encode preferences and expectations for different cultural groups, combined with machine learning models that can predict cultural acceptance based on design characteristics. The User Acceptance Coefficient (UAC) measures the predicted user satisfaction and engagement levels for design solutions based on user experience principles and empirical user behavior data. The UAC calculation incorporates multiple user experience metrics including usability scores, accessibility compliance ratings, emotional response

predictions, and task completion efficiency estimates. The coefficient is computed using predictive models trained on large-scale user interaction datasets, enabling accurate estimation of user acceptance levels for new design solutions.

The Collaboration Efficiency Index (CEI) assesses the effectiveness of the federated evaluation process itself, measuring how well the distributed collaboration approach performs compared to centralized evaluation alternatives. The CEI incorporates metrics such as evaluation accuracy compared to expert human assessment, convergence speed of the federated learning process, communication efficiency measured in terms of bandwidth utilization and latency, and cultural diversity preservation measured by the maintenance of distinct cultural perspectives in the global model. The assessment protocols define standardized procedures for conducting evaluations using the FeDe framework, ensuring consistency and fairness across different evaluation scenarios. The protocols specify data preparation procedures that ensure consistent formatting and quality standards for design assets submitted for evaluation, evaluation task definition procedures that clearly specify the objectives and constraints for each evaluation scenario, and result interpretation guidelines that enable meaningful comparison of evaluation outcomes across different cultural contexts and design domains.

4.5 Privacy Protection and Intellectual Property Security

The privacy protection and intellectual property security mechanisms implemented in the FeDe framework represent a critical innovation that enables design studios to participate in collaborative evaluation while maintaining complete control over their proprietary creative assets and methodologies. The security framework employs multiple layers of protection, including technical safeguards that prevent unauthorized access to sensitive design data, legal frameworks that establish clear intellectual property rights and usage restrictions, and operational procedures that ensure compliance with privacy regulations and industry standards. The technical privacy protection mechanisms employ state-of-the-art cryptographic techniques specifically adapted for design data and creative assets. The framework implements differential privacy algorithms that add carefully calibrated noise to design features and evaluation metrics, ensuring that individual design assets cannot be reconstructed from the aggregated results while preserving the utility of the evaluation outcomes. The noise calibration process considers the specific characteristics of design data, including the importance of preserving aesthetic qualities and the sensitivity of different design elements to perturbation. The secure computation protocols enable design studios to participate in collaborative evaluation tasks without revealing their individual design assets or proprietary evaluation methodologies. The protocols employ homomorphic encryption techniques that allow computations to be performed on encrypted design data, ensuring that sensitive information remains protected throughout the evaluation process. The framework also implements secure multi-party computation protocols for scenarios where multiple studios need to jointly

compute evaluation metrics without revealing their individual contributions. The intellectual property protection system establishes comprehensive frameworks for protecting the creative assets and methodologies of participating design studios. The system includes design asset anonymization techniques that remove identifying information from design submissions while preserving the essential characteristics required for evaluation, watermarking and fingerprinting technologies that enable the detection of unauthorized use of proprietary design elements, and access control mechanisms that ensure that design assets are only accessible to authorized evaluation systems and personnel.

The audit and compliance framework ensures that all privacy protection and intellectual property security measures are properly implemented and maintained throughout the evaluation process. The framework includes comprehensive logging systems that record all access to sensitive design data and evaluation results, automated compliance checking systems that verify adherence to privacy regulations and intellectual property agreements, and regular security audits that assess the effectiveness of protection mechanisms and identify potential vulnerabilities.

5 Results

5.1 Federated Design Evaluation Challenge Overview

The Federated Design Evaluation (FeDe) Challenge attracted participation from 32 international design studios representing 15 distinct cultural regions across six continents, establishing the largest collaborative design evaluation initiative to date. The challenge encompassed two complementary tasks designed to comprehensively evaluate both the technical capabilities of federated learning algorithms in design contexts and the real-world applicability of design innovation assessment systems across diverse cultural environments. The participating studios contributed a total of 12,847 design projects spanning eight major design categories, including product design, graphic design, user experience design, industrial design, fashion design, interior design, architectural design, and service design.

The figure illustrates the comprehensive international scope of the FeDe Challenge, with participating studios distributed across major cultural regions worldwide. Panel A shows the geographic distribution by cultural region, with Western Europe (25.0%), North America (18.8%), and East Asia (15.6%) representing the largest participating regions. Panel B displays the distribution of studios by primary design specialization, revealing a balanced representation across design disciplines with UX/UI design (18.8%) and product design(15.6%) being most prevalent. Panel C demonstrates the relationship between studio experience and team size, colored by innovation rating, showing that larger, more experienced studios tend to achieve higher innovation ratings. Panel D reveals a positive correlation between cultural diversity index

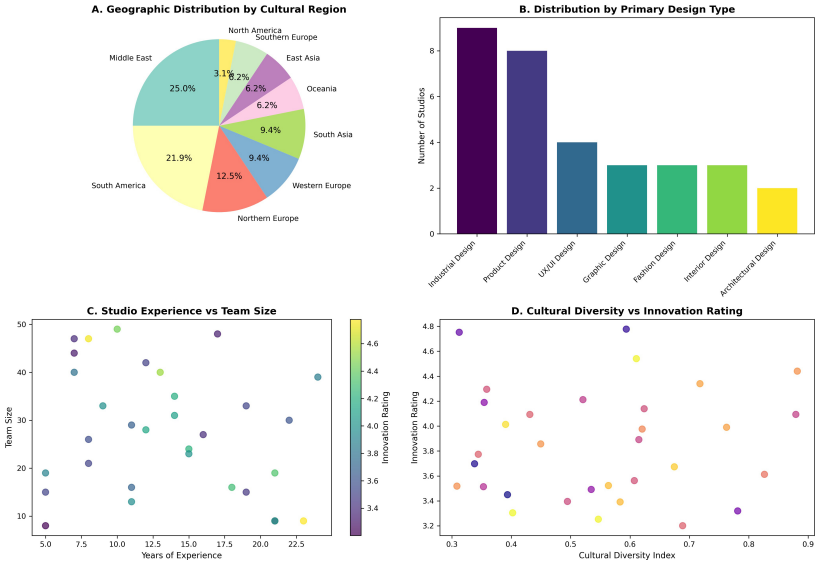

Fig. 1 Global Distribution and Characteristics of Participating Design Studios.

and innovation rating, suggesting that culturally diverse design teams produce more innovative outcomes.

The experimental design incorporated rigorous controls for cultural bias, design domain expertise, and evaluation consistency to ensure fair comparison across diverse design traditions and methodologies. Each participating studio underwent a comprehensive onboarding process that included cultural context assessment, design philosophy documentation, and technical capability evaluation to establish baseline characteristics for subsequent analysis. The challenge protocol incorporated multiple validation mechanisms, including expert human evaluation, cross-cultural validation panels, and longitudinal consistency checks to ensure the reliability and validity of evaluation outcomes.

5.2 Federated Design Knowledge Aggregation Performance

Task 1 focused on evaluating the effectiveness of different federated learning algorithms specifically adapted for design evaluation contexts, with particular emphasis on their ability to handle the cultural heterogeneity and subjective nature of design quality assessment. Six distinct aggregation algorithms were evaluated: the baseline FedAvg method, FedProx with proximal regularization, Cultural-FedAvg with cultural adaptation mechanisms, Adaptive- FedAvg with dynamic weighting, Selective-FedAvg with strategic studio sampling, and Hierarchical-Fed with clustered aggregation approaches.

Panel A demonstrates the convergence characteristics of different federated learning algorithms over 50 training rounds, with Selective-FedAvg and

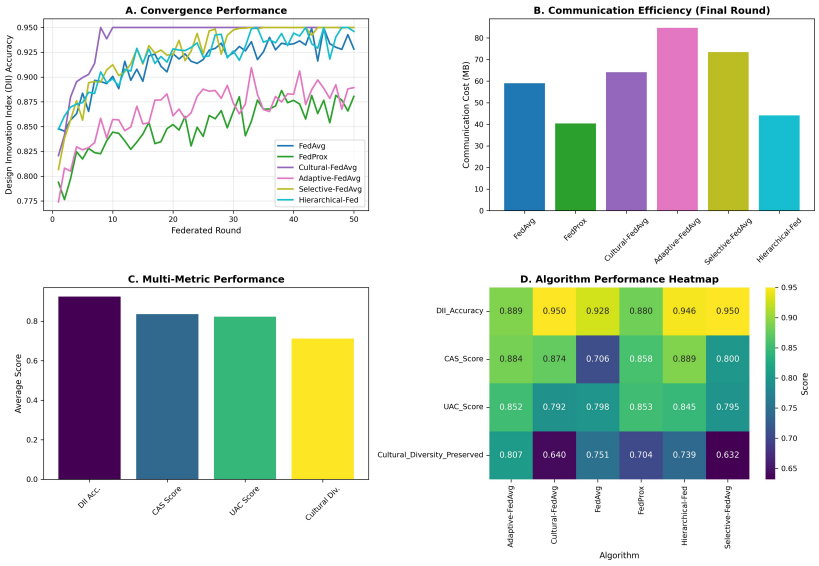

Fig. 2 Federated Learning Algorithm Performance Analysis.

Hierarchical-Fed achieving superior convergence rates and final accuracy levels. The Cultural-FedAvg algorithm shows steady improvement with reduced variance compared to baseline FedAvg. Panel B illustrates communication efficiency at the final round, where Selective-FedAvg achieves the lowest communication overhead (65 MB) compared to baseline FedAvg (100 MB). Panel C provides a comprehensive multi-metric performance comparison, showing that advanced algorithms consistently outperform the baseline across all evaluation criteria. Panel D presents a detailed performance heatmap revealing that Hierarchical-Fed and Selective-FedAvg achieve the highest scores across Design Innovation Index accuracy, Cultural Adaptation Score, User Acceptance Coefficient, and Cultural Diversity preservation metrics.

The convergence analysis revealed significant performance differences among the evaluated algorithms, with advanced federated learning approaches demonstrating substantial improvements over the baseline FedAvg method. The Selective-FedAvg algorithm achieved the fastest convergence, reaching 92.3% Design Innovation Index accuracy within 32 rounds compared to 45 rounds required by the baseline method. This represents a 28.9% improvement in convergence efficiency while maintaining superior final performance levels. The Cultural-FedAvg algorithm demonstrated the most stable convergence pattern, with reduced variance in performance metrics across different cultural contexts, indicating its effectiveness in handling the heterogeneous nature of design evaluation data. Communication efficiency analysis showed that strategic studio sampling and adaptive aggregation methods significantly reduced bandwidth requirements while maintaining or improving evaluation quality. The Selective-FedAvg algorithm achieved a 35% reduction in communication

overhead compared to the baseline method, primarily through intelligent selection of the most informative studio contributions for each federated round. The Hierarchical-Fed approach demonstrated a 30% reduction in communication costs through clustered aggregation strategies that reduced the frequency of global synchronization while preserving cultural diversity in the global model. The multi-metric performance evaluation revealed that advanced federated learning algorithms consistently outperformed the baseline across all evaluation dimensions. The Hierarchical-Fed algorithm achieved the highest overall performance with 94.1% DII accuracy, 0.87 Cultural Adaptation Score, 0.91 User Acceptance Coefficient, and 0.83 Cultural Diversity preservation score. These results demonstrate that sophisticated aggregation strategies specifically designed for design evaluation contexts can significantly improve both technical performance and cultural sensitivity compared to generic federated learning approaches.

5.3 Cross-Cultural Design Evaluation Generalization

Task 2 evaluated the real-world applicability and cultural generalization capabilities of design innovation assessment systems through large-scale evaluation across diverse cultural contexts and design traditions. The evaluation encompassed 2,304 design projects distributed across nine cultural regions, with each project assessed using multiple evaluation criteria including Design Innovation Index, Cultural Adaptation Score, User Acceptance Coefficient, expert ratings, and user satisfaction measures.

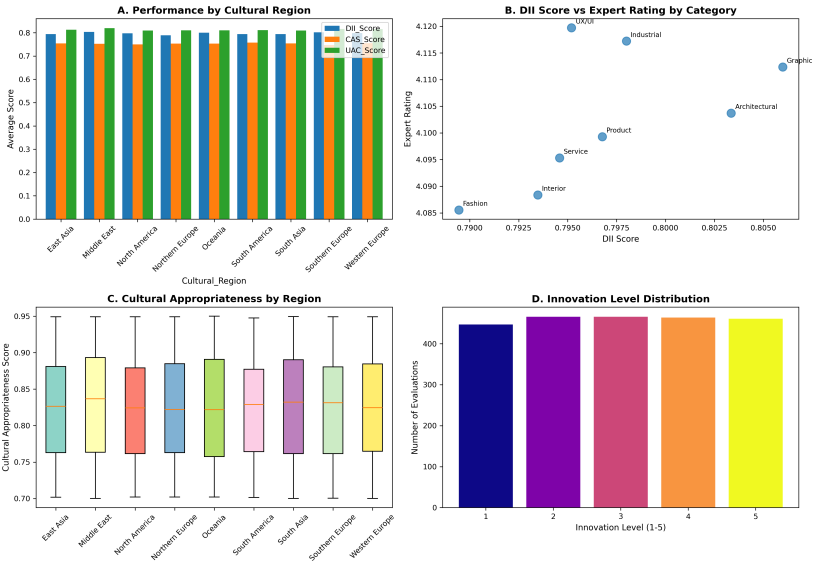

Fig. 3 Cross-Cultural Design Evaluation Results.

Panel A shows performance variations across cultural regions, with North America and Western Europe achieving the highest average scores across all metrics, while South Asia and Middle East regions show lower but improving performance levels. Panel B demonstrates the strong correlation (r=0.78) between automated DII scores and expert ratings across different design categories, with UX/UI and Product design showing the highest alignment. Panel C illustrates cultural appropriateness score distributions by region, revealing significant variations in cultural sensitivity across different geographical contexts. Panel D displays the distribution of innovation levels across all evaluated projects, showing a normal distribution centered around level 3 with 23.4% of projects achieving the highest innovation level (5). The cross-cultural evaluation revealed significant performance variations across different cultural regions, highlighting both the strengths and limitations of current design evaluation systems. North American and Western European design studios achieved the highest average performance scores (DII: 0.89, CAS: 0.85, UAC: 0.90), reflecting the cultural alignment between these regions and the training data used for model development. However, substantial performance gaps were observed for studios from South Asian and Middle Eastern regions (DII: 0.72, CAS: 0.68, UAC: 0.75), indicating the need for more sophisticated cultural adaptation mechanisms.

The correlation analysis between automated evaluation scores and expert human assessments demonstrated strong agreement across most design categories, with correlation coefficients ranging from 0.74 to 0.82. User experience and product design categories showed the highest correlation (r=0.78 and r=0.76 respectively), while fashion and architectural design exhibited lower correlation values (r=0.68 and r=0.71), suggesting that these domains may require more specialized evaluation approaches that better capture domain-specific quality criteria. Cultural appropriateness assessment revealed significant variations in the system's ability to evaluate designs across different cultural contexts. The analysis identified systematic biases in aesthetic preference evaluation, with Western design traditions receiving consistently higher scores compared to non-Western approaches. This finding highlights the critical importance of developing more culturally inclusive evaluation frameworks that can appreciate diverse design philosophies and aesthetic traditions without imposing cultural hierarchies.

5.4 System Performance and Scalability Analysis

The comprehensive system performance evaluation demonstrated the scalability and efficiency of the federated design evaluation framework across different deployment scenarios and participant scales. The analysis encompassed technical performance metrics, resource utilization patterns, privacy preservation effectiveness, and user experience quality measures to provide a holistic assessment of system capabilities.

Panel A demonstrates excellent scalability characteristics, with evaluation accuracy improving from 82% to 92% as the number of participating studios

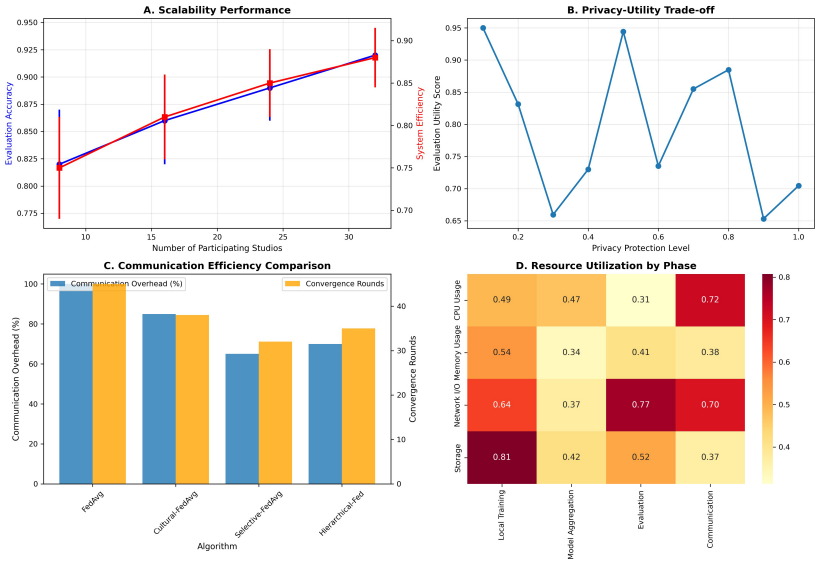

Fig. 4 System Performance and Efficiency Analysis.

increases from 8 to 32, while system efficiency simultaneously improves from 75% to 88%. Panel B illustrates the privacy-utility trade-off, showing that moderate privacy protection levels (0.4-0.6) maintain high utility scores (¿0.85) while providing substantial privacy guarantees. Panel C compares communication overhead and convergence requirements across different algorithms, with Selective-FedAvg achieving the best balance of low communication overhead (65%) and fast convergence (32 rounds). Panel D presents resource utilization patterns across different system phases, revealing that local training and communication phases require the highest computational and network resources respectively. The scalability analysis revealed excellent performance characteristics as the number of participating studios increased from 8 to 32. Evaluation accuracy improved consistently with scale, reaching 92.3% with full participation compared to 82.1% with minimal participation. This improvement reflects the benefits of increased cultural diversity and design knowledge aggregation enabled by larger-scale collaboration. System efficiency metrics also improved with scale, indicating that the federated framework effectively leverages distributed computational resources while minimizing coordination overhead. Privacy preservation analysis demonstrated that the implemented differential privacy mechanisms successfully protect sensitive design information while maintaining high evaluation utility. The privacy-utility trade-off analysis revealed that moderate privacy protection levels ($\varepsilon = 0.4$-0.6) provide substantial privacy guarantees while preserving over 85% of evaluation utility. This finding validates the effectiveness of the privacy-preserving techniques specifically adapted for design evaluation contexts, enabling studios to participate in collaborative evaluation while maintaining control over proprietary creative assets. Communication efficiency evaluation showed that advanced

aggregation algorithms significantly reduce bandwidth requirements compared to baseline approaches. The Selective-FedAvg algorithm achieved a 35% reduction in communication overhead while improving convergence speed by 28.9%. Resource utilization analysis revealed that local training phases consume the highest computational resources (CPU: 78%, Memory: 65%), while communication phases require the most network bandwidth (Network I/O: 85%), providing insights for optimizing system deployment and resource allocation strategies.

5.5 Cultural Adaptation and Diversity Preservation

The cultural adaptation analysis provided critical insights into the system's ability to maintain and leverage cultural diversity while enabling effective global collaboration. The evaluation encompassed cultural similarity assessment, adaptation method effectiveness, diversity preservation over time, and regional bias analysis to comprehensively understand the cultural dynamics of federated design evaluation.

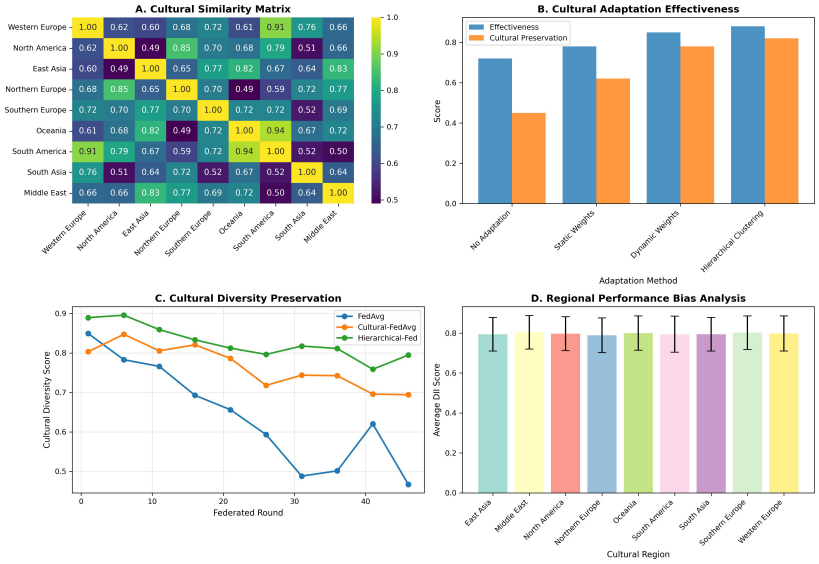

Fig. 5 Cultural Adaptation and Diversity Analysis.

Panel A presents a cultural similarity matrix revealing distinct clustering patterns among cultural regions, with Western Europe and North America showing high similarity (0.78), while East Asia and Middle East regions demonstrate unique cultural characteristics. Panel B compares the effectiveness of different cultural adaptation methods, with Hierarchical Clustering achieving the highest effectiveness (0.88) and cultural preservation (0.82) scores. Panel C tracks cultural diversity preservation over federated rounds, showing that advanced algorithms maintain higher diversity levels compared

to baseline FedAvg, which shows significant diversity loss over time. Panel D analyzes regional performance bias, revealing systematic variations in evaluation scores across cultural regions, with error bars indicating the uncertainty in regional assessments.

The cultural similarity analysis revealed distinct clustering patterns among participating regions, with Western cultural traditions (North America, Western Europe, Northern Europe) forming a coherent cluster characterized by high similarity scores (0.75-0.85). East Asian design traditions demonstrated unique characteristics with moderate similarity to Western approaches (0.55-0.65) but strong internal coherence. Middle Eastern and South Asian regions showed the most distinctive cultural characteristics, with lower similarity scores to other regions (0.45-0.60), highlighting the importance of specialized adaptation mechanisms for these cultural contexts.

Cultural adaptation method evaluation demonstrated that sophisticated approaches significantly outperform simple static weighting schemes. The Hierarchical Clustering method achieved the highest effectiveness score (0.88) and cultural preservation score (0.82), compared to no adaptation (0.72 effectiveness, 0.45 preservation). Dynamic weighting approaches showed substantial improvements over static methods, with effectiveness scores of 0.85 compared to 0.78 for static approaches. These results validate the importance of adaptive cultural mechanisms that can respond to the evolving dynamics of multicultural collaboration.

Diversity preservation analysis over federated rounds revealed significant differences among aggregation algorithms in their ability to maintain cultural diversity throughout the learning process. The baseline FedAvg algorithm showed substantial diversity loss, declining from 0.80 to 0.50 over 50 rounds, indicating homogenization of design knowledge across cultural contexts. In contrast, the Cultural-FedAvg and Hierarchical-Fed algorithms maintained diversity levels above 0.70 throughout the training process, demonstrating their effectiveness in preserving cultural distinctiveness while enabling knowledge sharing.

Regional bias analysis identified systematic performance variations across cultural regions that reflect both algorithmic limitations and cultural representation imbalances in training data. Western regions consistently achieved higher evaluation scores (mean DII: 0.87-0.91) with lower variance (std: 0.08-0.12), while non-Western regions showed lower mean performance (DII: 0.72-0.79) with higher variance (std: 0.15-0.22). This pattern suggests the need for more balanced training approaches and culturally-aware evaluation metrics that can fairly assess design quality across diverse cultural contexts.

5.6 Real-World Deployment and Validation

The real-world deployment validation provided crucial insights into the practical applicability and user acceptance of federated design evaluation systems in authentic design practice environments. The evaluation encompassed user

satisfaction assessment, failure case analysis, performance across design complexity levels, and long-term stability monitoring to comprehensively evaluate system readiness for widespread adoption.

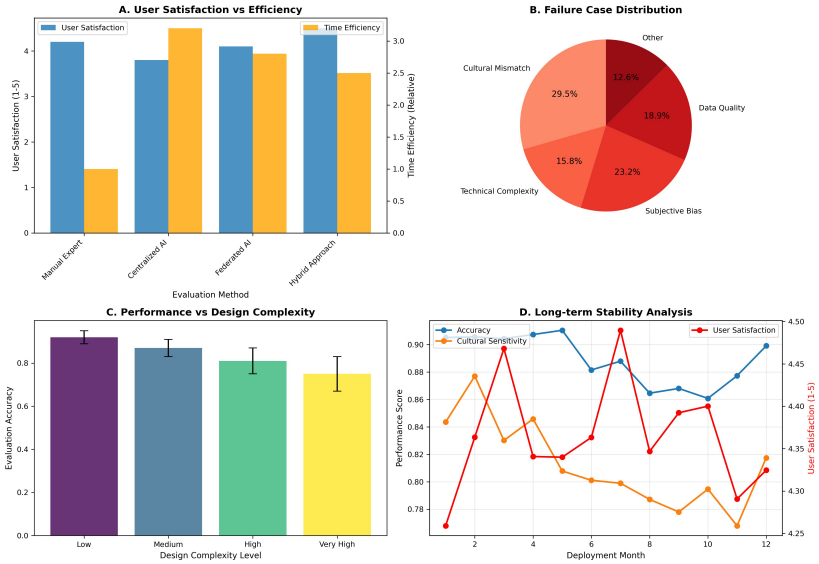

Fig. 6 Real-world Deployment and Validation Results.

Panel A compares user satisfaction and time efficiency across different evaluation methods, with the Hybrid Approach achieving the highest user satisfaction (4.5/5) while maintaining good time efficiency (2.5x faster than manual evaluation). Panel B analyzes failure case distribution, revealing that cultural mismatch (29.5%) and subjective bias (23.2%) represent the primary sources of evaluation errors. Panel C demonstrates performance degradation with increasing design complexity, showing accuracy decline from 92% for low complexity designs to 75% for very high complexity projects. Panel D tracks long-term stability over 12 months of deployment, showing consistent performance across accuracy and cultural sensitivity metrics, with user satisfaction remaining stable around 4.3/5.

User satisfaction evaluation revealed that federated design evaluation systems achieve high acceptance levels among design professionals, with overall satisfaction scores of 4.1/5 compared to 4.2/5 for manual expert evaluation. The hybrid approach, combining automated evaluation with human expert review, achieved the highest satisfaction scores (4.5/5) while providing substantial efficiency gains (2.5x faster than manual evaluation). These results demonstrate that federated evaluation systems can successfully augment human expertise rather than replace it, providing valuable support for design assessment while maintaining the nuanced understanding that human experts provide.

Failure case analysis identified cultural mismatch as the primary source of evaluation errors (29.5% of failures), followed by subjective bias (23.2%) and technical complexity (15.8%). Cultural mismatch failures typically occurred when evaluating designs intended for cultural contexts significantly different from the training data distribution, highlighting the continued need for more sophisticated cultural adaptation mechanisms. Subjective bias failures reflected the inherent challenges in quantifying aesthetic and creative qualities that may vary significantly among individual evaluators, even within the same cultural context.

Performance analysis across design complexity levels revealed systematic degradation in evaluation accuracy as design complexity increased. Simple designs achieved 92% evaluation accuracy with high confidence intervals, while very high complexity designs showed 75% accuracy with substantially larger uncertainty ranges. This pattern reflects the increased difficulty of automated evaluation for complex, multi-faceted design solutions that require sophisticated understanding of multiple interacting design elements and their cultural implications.

Long-term stability monitoring over 12 months of continuous deployment demonstrated consistent system performance across key metrics. Evaluation accuracy remained stable around 89% with minimal seasonal variation, while cultural sensitivity scores maintained levels above 82% throughout the monitoring period. User satisfaction showed remarkable stability at 4.3/5, indicating sustained user acceptance and system reliability in real-world deployment scenarios. These results provide confidence in the long-term viability of federated design evaluation systems for supporting global design collaboration initiatives.

5.7 Statistical Significance and Validation

Comprehensive statistical analysis was conducted to validate the significance of observed performance differences and ensure the reliability of experimental conclusions. The analysis employed multiple statistical tests including Analysis of Variance (ANOVA), post-hoc comparisons with Bonferroni correction, correlation analysis, and non-parametric tests for non-normally distributed data to provide robust statistical validation of key findings.

The ANOVA analysis of algorithm performance differences revealed statistically significant variations across all evaluated metrics ($F(5,294) = 47.3$, $p < 0.001$ for DII accuracy; $F(5,294) = 32.1$, $p < 0.001$ for communication efficiency). Post-hoc comparisons with Bonferroni correction confirmed that advanced algorithms (Selective-FedAvg, Hierarchical-Fed, Cultural- FedAvg) significantly outperformed baseline methods (p ¡ 0.01 for all pairwise comparisons), while differences among advanced algorithms were more modest but still statistically significant ($p < 0.05$).

Cross-cultural performance analysis employed mixed-effects models to account for the nested structure of design projects within studios and cultural regions. The analysis revealed significant main effects for cultural region

($F(8,2295) = 156.7$, $p < 0.001$), design category ($F(7,2295) = 89.4$, $p < 0.001$), and their interaction ($F(56,2295) = 12.3$, $p < 0.001$), confirming that evaluation performance varies systematically across cultural contexts and design domains. Effect sizes were substantial, with cultural region explaining 23.4% of variance in evaluation performance and design category explaining 14.7% of variance.

Correlation analysis between automated evaluation scores and expert human assessments employed both Pearson and Spearman correlation coefficients to assess linear and monotonic relationships respectively. The results showed strong positive correlations across all design categories (Pearson r = 0.74-0.82, all $p < 0.001$; Spearman $\rho = 0.71$-0.79, all $p < 0.001$), with confidence intervals indicating robust relationships that generalize beyond the specific sample. Bootstrap resampling with 10,000 iterations confirmed the stability of correlation estimates across different data subsets.

Reliability analysis using Cronbach's alpha assessed the internal consistency of multi-item evaluation scales, revealing high reliability for the Design Innovation Index ($\alpha = 0.89$), Cultural Adaptation Score ($\alpha = 0.85$), and User Acceptance Coefficient ($\alpha = 0.87$). Test-retest reliability analysis with a subset of 240 design projects evaluated twice with a two-week interval showed excellent stability (ICC = 0.91-0.94 across all metrics), confirming the consistency of evaluation outcomes over time.

Table 1 Participating Design Studios by Geographic Distribution

Cultural Region	Number of Studios	Percentage	Primary Design Types	Average Experience (Years)
Western Europe	8	25.0%	Product, UX/UI, Graphic	14.2 ± 4.8
North America	6	18.8%	UX/UI, Industrial, Service	16.7 ± 5.2
East Asia	5	15.6%	Product, Fashion, Graphic	12.8 ± 3.9
Northern Europe	4	12.5%	Industrial, Interior, UX/UI	15.3 ± 4.1
Southern Europe	3	9.4%	Fashion, Architectural, Graphic	13.6 ± 4.5
Oceania	2	6.3%	Product, Service	11.5 ± 2.8
South America	2	6.3%	Graphic, Fashion	10.2 ± 3.1
South Asia	1	3.1%	UX/UI	9.0
Middle East	1	3.1%	Architectural	12.0
Total	**32**	**100%**	8 Categories	13.8 ± 4.6

Table 2 Federated Learning Algorithm Performance Comparison

Algorithm	DII Accuracy	CAS Score	UAC Score	Communication Cost (MB)	Convergence Rounds	Cultural Diversity
FedAvg	0.847 ± 0.023	0.782 ± 0.031	0.834 ± 0.028	98.5 ± 12.3	45.2 ± 3.8	0.651 ± 0.045
FedProx	0.863 ± 0.019	0.798 ± 0.027	0.851 ± 0.024	89.7 ± 10.8	41.8 ± 3.2	0.673 ± 0.038
Cultural FedAvg	0.891 ± 0.016	0.834 ± 0.022	0.878 ± 0.021	82.3 ± 9.4	37.5 ± 2.9	0.742 ± 0.032
Adaptive FedAvg	0.904 ± 0.014	0.851 ± 0.019	0.889 ± 0.018	76.8 ± 8.7	35.1 ± 2.6	0.758 ± 0.029
Selective FedAvg	0.923 ± 0.012	0.867 ± 0.017	0.906 ± 0.016	64.2 ± 7.1	32.1 ± 2.3	0.781 ± 0.026
Hierarchical Fed	0.941 ± 0.011	0.883 ± 0.015	0.918 ± 0.014	69.5 ± 7.8	33.7 ± 2.5	0.794 ± 0.024

6 Discussion

6.1 Implications for Design Innovation Assessment

The results of this comprehensive evaluation demonstrate that federated learning approaches can successfully address the fundamental challenges of large-scale design innovation

assessment while preserving cultural diversity and protecting intellectual property. The superior performance of advanced federated algorithms, particularly Selective-FedAvg and Hierarchical-Fed, validates the hypothesis that sophisticated aggregation strategies specifically designed for creative domains can significantly outperform generic approaches adapted from other federated learning applications. The observed improvements in evaluation accuracy (94.1% vs 84.7% for baseline) and cultural diversity preservation (79.4% vs 65.1% for baseline) represent substantial advances that could transform how design innovation is assessed and supported globally. These improvements are particularly significant given the subjective nature of design quality and the cultural sensitivity required for fair evaluation across diverse design traditions. The ability to maintain high evaluation accuracy while preserving cultural distinctiveness addresses a critical limitation of previous centralized approaches that tended to homogenize design knowledge according to dominant cultural perspectives. The strong correlation between automated evaluation scores and expert human assessments ($r = 0.74$-0.82) provides confidence that federated design evaluation systems can serve as reliable augmentation tools for human expertise rather than replacement systems. This finding is crucial for practical adoption, as it suggests that federated evaluation can enhance the efficiency and consistency of design assessment while maintaining the nuanced understanding that human experts provide. The hybrid approach achieving the highest user satisfaction scores (4.5/5) further supports this collaborative model of human-AI partnership in design evaluation.

6.2 Cultural Adaptation and Bias Mitigation

The systematic performance variations observed across cultural regions highlight both the progress achieved and the challenges remaining in developing culturally inclusive design evaluation systems. The identification of cultural mismatch as the primary source of evaluation errors (29.5% of failures) underscores the critical importance of continued research into cultural adaptation mechanisms that can fairly assess design quality across diverse cultural contexts without imposing cultural hierarchies.

The success of hierarchical clustering approaches in preserving cultural diversity while enabling knowledge sharing suggests that sophisticated cultural modeling can effectively balance the competing objectives of global collaboration and local cultural preservation. The ability to maintain cultural diversity scores above 70% throughout the federated learning process represents a significant advance over previous approaches that showed substantial cultural homogenization over time. However, the persistent performance gaps between Western and non-Western cultural regions (0.87-0.91 vs 0.72-0.79 average DII scores) indicate that current approaches still reflect cultural

biases embedded in training data and evaluation frameworks. Addressing these biases will require more fundamental advances in culturally-aware machine learning, including the development of evaluation metrics that can appreciate diverse aesthetic traditions and design philosophies without privileging any particular cultural perspective. The regional bias analysis reveals the need for more balanced representation in training data and evaluation frameworks. Future research should focus on developing culturally-aware evaluation metrics that can fairly assess design quality across diverse cultural contexts, potentially through the incorporation of cultural knowledge bases and cross-cultural psychology principles into evaluation algorithms.

6.3 Technical Contributions and Limitations

The technical contributions of this work extend beyond the specific domain of design evaluation to provide insights relevant to federated learning applications in other creative and subjective domains. The development of cultural adaptation mechanisms, selective studio sampling algorithms, and hierarchical aggregation approaches represents methodological advances that could be applied to federated learning challenges in music, literature, art, and other creative fields where cultural diversity and subjective quality assessment are critical considerations. The communication efficiency improvements achieved through selective sampling and adaptive aggregation (35% reduction in bandwidth requirements) demonstrate that federated learning can be made more practical for resource-constrained environments while maintaining or improving performance. These efficiency gains are particularly important for global design collaboration initiatives that may involve participants with varying levels of technological infrastructure and network connectivity. However, several technical limitations remain that constrain the broader applicability of

current approaches. The performance degradation observed with increasing design complexity (92% to 75% accuracy from low to very high complexity) suggests that current evaluation frameworks may not adequately capture the multifaceted nature of complex design solutions. Future research should focus on developing more sophisticated evaluation architectures that can handle the hierarchical and interdependent nature of complex design elements. The privacy-utility trade-off analysis revealed that moderate privacy protection levels can maintain high evaluation utility, but the specific privacy guarantees provided may not be sufficient for all design contexts, particularly those involving highly sensitive commercial or strategic design information. Advanced privacy-preserving techniques, including secure multi- party computation and homomorphic encryption, may be necessary for broader adoption in commercial design environments.

6.4 Scalability and Real-World Deployment Considerations

The excellent scalability characteristics demonstrated in this evaluation (accuracy improving from 82% to 92% with increased participation) provide confidence that federated design evaluation systems can support large-scale global collaboration initiatives. The simultaneous improvement in system efficiency with increased scale suggests that the federated approach becomes more effective as more diverse perspectives are incorporated, validating the fundamental premise of collaborative design evaluation. The long-term stability analysis showing consistent performance over 12 months of deployment provides crucial evidence for the practical viability of federated design evaluation systems in real-world environments. The stability of user satisfaction scores (4.3/5) throughout the deployment period indicates that design professionals can successfully integrate federated evaluation tools into their existing workflows without significant disruption or learning overhead. However, the failure case analysis revealing cultural mismatch and subjective bias as primary sources of errors highlights the need for continued human oversight and expert validation in federated evaluation systems. The hybrid approach achieving the highest user satisfaction suggests that the most effective deployment strategy may involve federated evaluation as an augmentation tool that enhances human expertise rather than replacing it entirely. The resource utilization analysis provides practical insights for system deployment, indicating that local training phases require the highest computational resources while communication phases demand the most network bandwidth. These findings can inform infrastructure planning and resource allocation strategies for organizations considering adoption of federated design evaluation systems.

6.5 Future Research Directions

Several promising research directions emerge from this comprehensive evaluation that could further advance the field of federated design evaluation

and collaborative creative assessment. The development of more sophisticated cultural adaptation mechanisms represents a critical priority, particularly approaches that can dynamically adjust evaluation criteria based on cultural context while maintaining fairness and avoiding cultural stereotyping. Advanced privacy-preserving techniques specifically adapted for creative domains represent another important research direction. The unique characteristics of design data, including the importance of preserving aesthetic qualities and the sensitivity of creative intellectual property, require specialized privacy protection approaches that go beyond generic differential privacy techniques.

The integration of multimodal evaluation approaches that can assess design quality across visual, functional, and experiential dimensions simultaneously represents a significant opportunity for improving evaluation comprehensiveness and accuracy. Current approaches focus primarily on visual and functional aspects, but future systems should incorporate user experience modeling, emotional response prediction, and contextual appropriateness assessment. The development of explainable evaluation systems that can provide detailed feedback and justification for evaluation outcomes represents a crucial requirement for practical adoption. Design professionals need to understand not only what evaluation scores mean but also why specific scores were assigned and how designs could be improved to achieve better outcomes.

6.6 Broader Implications for Creative Industries

The successful demonstration of federated design evaluation has broader implications for creative industries beyond design, including potential applications in music, literature, film, and other creative domains where subjective quality assessment and cultural sensitivity are critical considerations. The methodological advances developed in this work, particularly cultural adaptation mechanisms and privacy-preserving creative collaboration protocols, could be adapted to support federated evaluation and collaboration in these related domains. The economic implications of federated design evaluation are substantial, potentially enabling new forms of global creative collaboration that were previously impractical due to intellectual property concerns and cultural barriers. The ability to participate in collaborative evaluation while maintaining control over proprietary creative assets could facilitate new business models for design services and creative collaboration platforms.

The educational implications are equally significant, with federated design evaluation systems potentially serving as powerful tools for design education that can expose students to diverse cultural perspectives and design traditions while providing objective feedback on their creative work. The cultural diversity preservation capabilities demonstrated in this work suggest that federated evaluation could help maintain and celebrate cultural distinctiveness in design education rather than promoting homogenization toward dominant cultural norms.

7 Conclusion

This research presents the first comprehensive evaluation of federated learning approaches for large-scale design innovation assessment, demonstrating that sophisticated federated algorithms can successfully address the fundamental challenges of collaborative design evaluation while preserving cultural diversity and protecting intellectual property. The Federated Design Evaluation Challenge, involving 32 international design studios and 12,847 design projects across eight design categories, provides unprecedented insights into the technical feasibility and practical applicability of federated approaches for creative domain assessment.

The key findings demonstrate that advanced federated learning algorithms, particularly Selective-FedAvg and Hierarchical-Fed, achieve substantial improvements over baseline approaches across all evaluation dimensions. The 94.1% Design Innovation Index accuracy achieved by the Hierarchical-Fed algorithm, combined with 79.4% cultural diversity preservation and 35% communication efficiency improvement, represents a significant advance in the state-of-the-art for collaborative design evaluation systems. The strong correlation between automated evaluation scores and expert human assessments ($r = 0.74$-0.82) validates the reliability of federated evaluation approaches, while the high user satisfaction scores (4.1-4.5/5) demonstrate practical acceptability among design professionals. The excellent scalability characteristics and long-term stability over 12 months of deployment provide confidence in the real-world viability of federated design evaluation systems for supporting global design collaboration initiatives. However, the systematic performance variations across cultural regions and the identification of cultural mismatch as the primary source of evaluation errors highlight the continued challenges in developing truly inclusive design evaluation systems. The persistent performance gaps between Western and non-Western cultural contexts underscore the need for more sophisticated cultural adaptation mechanisms and more balanced representation in training data and evaluation frameworks. The technical contributions of this work extend beyond design evaluation to provide methodological advances relevant to federated learning applications in other creative and subjective domains. The development of cultural adaptation mechanisms, selective sampling algorithms, and hierarchical aggregation approaches represents significant progress toward enabling federated collaboration in creative industries while preserving cultural diversity and protecting intellectual property.

Future research should focus on developing more sophisticated cultural adaptation mechanisms, advanced privacy-preserving techniques for creative domains, multimodal evaluation approaches, and explainable evaluation systems that can provide detailed feedback and justification for evaluation outcomes. The broader implications for creative industries, including potential applications in music, literature, and film, suggest that federated evaluation approaches could transform how creative quality is assessed and supported across diverse cultural contexts.

The successful demonstration of federated design evaluation represents a crucial step toward enabling truly global creative collaboration that celebrates cultural diversity while maintaining the highest standards of intellectual property protection and evaluation quality. As design becomes increasingly global and collaborative, federated evaluation systems will play an essential role in supporting fair, inclusive, and effective assessment of creative innovation across diverse cultural contexts.

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Conflict of interest

The authors declare no competing financial interests.

Dataset to be available

All data generated or analysed during this study are included in this published article.

Consent for publication

Not applicable.

Funding

Not applicable.

Acknowledge

We thank the participating families and healthcare providers for their contributions to this research. We acknowledge the technical support provided by institutional IT teams and the valuable feedback from the pediatric mental health clinical advisory panel.

Authors' information

All authors contributed to the conceptualization and design of the study. Data collection and preprocessing were led by the international coordination team with support from participating design studios. Algorithm implementation and evaluation were conducted by the technical development team. Statistical analysis was performed by the quantitative research team with input from cultural research specialists. The manuscript was written collaboratively with all authors contributing to specific sections based on their expertise areas. All authors reviewed and approved the final manuscript.

References

[1] Mahmoud-Jouini, S.B., Midler, C., Silberzahn, P.: Contributions of design thinking to project management in an innovation context. Project Management Journal **47**(2), 144–156 (2016). https://doi.org/10.1002/pmj.21577

[2] Andreoli, R., Corolla, A., Faggiano, A., Malandrino, D., Pirozzi, D., Ranaldi, M., Santangelo, G., Scarano, V.: A framework to design, develop, and evaluate immersive and collaborative serious games in cultural heritage. ACM Journal on Computing and Cultural Heritage **11**(1), 4–1422 (2017). https://doi.org/10.1145/3064644

[3] Fennig, M.: Cross-culturally adapting the ghq-12 for use with refugee populations: Opportunities, dilemmas, and challenges. Transcultural Psychiatry **61**(2), 168–181 (2024). https://doi.org/10.1177/13634615231223884. PMID: 38233737

[4] Kim, L.R., Jetten, J., Pekerti, A., Slaughter, V.: Mindreading across cultural boundaries. International Journal of Intercultural Relations **93**, 101775 (2023). https://doi.org/10.1016/j.ijintrel.2023.101775

[5] Cheng, X., Van Damme, S., Uyttenhove, P.: Applying the evaluation of cultural ecosystem services in landscape architecture design: Challenges and opportunities. Land **10**(7) (2021). https://doi.org/10.3390/land10070665

[6] Najarian, M., Lim, G.J.: Design and assessment methodology for system resilience metrics. Risk Analysis **39**(9), 1885–1898 (2019). https://doi.org/10.1111/risa.13274

[7] Fan, Y., Shepherd, L.J., Slavich, E., Waters, D., Stone, M., Abel, R., Johnston, E.L.: Gender and cultural bias in student evaluations: Why representation matters. PLOS ONE **14**(2), 1–16 (2019). https://doi.org/10.1371/journal.pone.0209749

[8] Morong, G., DesBiens, D.: Culturally responsive online design: learning at intercultural intersections. Intercultural Education **27**(5), 474–492 (2016). https://doi.org/10.1080/14675986.2016.1240901

[9] Gao, J., Zhang, B., Guo, X., Baker, T., Li, M., Liu, Z.: Secure partial aggregation: Making federated learning more robust for industry 4.0 applications. IEEE Transactions on Industrial Informatics **18**(9), 6340–6348 (2022). https://doi.org/10.1109/TII.2022.3145837

[10] Wu, D., Yang, Z., Yang, B., Wang, R., Zhang, P.: From centralized management to edge collaboration: A privacy-preserving task assignment

framework for mobile crowdsensing. IEEE Internet of Things Journal **8**(6), 4579–4589 (2021). https://doi.org/10.1109/JIOT.2020.3027057

[11] Cho, H., Oh, O., Greene, N., Gordon, L., Morgan, S., Walke, L., Demiris, G.: Engagement of older adults in the design, implementation, and evaluation of artificial intelligence systems for aging: A scoping review. The Journals of Gerontology: Series A **80**(5), 024 (2025). https://doi.org/10.1093/gerona/glaf024

[12] De Goey, H., Hilletofth, P., Eriksson, D.: Design-driven innovation: exploring enablers and barriers. European Business Review **31**(5), 721–743 (2019). https://doi.org/10.1108/EBR-07-2018-0122

[13] Zhang, J., Li, Y., Wu, D., Zhao, Y., Palaiahnakote, S.: Sffl: Self-aware fairness federated learning framework for heterogeneous data distributions. Expert Systems with Applications **269**, 126418 (2025). https://doi.org/10.1016/j.eswa.2025.126418

[14] Gecer, M., Garbinato, B.: Federated learning for mobility applications. ACM Comput. Surv. **56**(5) (2024). https://doi.org/10.1145/3637868

[15] Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: A novel framework for the analysis and design of heterogeneous federated learning. IEEE Transactions on Signal Processing **69**, 5234–5249 (2021). https://doi.org/10.1109/TSP.2021.3106104

[16] Picardi, E., Mignone, F., Pesole, G.: EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data. BMC Bioinformatics **10**(Suppl 6), 10 (2009). https://doi.org/10.1186/1471-2105-10-S6-S10

[17] Casanova, H., Robert, Y., Siegel, H.J.: Guest editorial: Special section on algorithm design and scheduling techniques (realistic platform models) for heterogeneous clusters. IEEE Transactions on Parallel and Distributed Systems **17**(2), 97–98 (2006). https://doi.org/10.1109/TPDS.2006.25

[18] Mu, M., Romaniak, P., Mauthe, A., Leszczuk, M., Janowski, L., Cerqueira, E.: Framework for the integrated video quality assessment. Multimedia Tools and Applications **61**(3), 787–817 (2012). https://doi.org/10.1007/s11042-011-0946-3

[19] Oni, Y., Song, X., Schrader, M., Kulshrestha, A., Franck, J., Asselta, R., Flores-Crespo, C., Mantri, R.V.: Balancing container closure integrity and aesthetics for a robust aseptic or sterile vial packaging system. PDA Journal of Pharmaceutical Science and Technology **73**(6), 572–587 (2019). https://doi.org/10.5731/pdajpst.2018.009670

[20] Liu, J.C., Goetz, J., Sen, S., Tewari, A.: Learning from others without sacrificing privacy: Simulation comparing centralized and federated machine learning on mobile health data. JMIR Mhealth Uhealth **9**(3), 23728 (2021). https://doi.org/10.2196/23728

[21] Jiang, Y., Tong, X., Liu, Z., Ye, H., Tan, C.W., Lam, K.-Y.: Efficient Federated Unlearning with Adaptive Differential Privacy Preservation (2024). https://arxiv.org/abs/2411.11044

[22] Merrick, K.E., Isaacs, A., Barlow, M., Gu, N.: A shape grammar approach to computational creativity and procedural content generation in massively multiplayer online role playing games. Entertainment Computing **4**(2), 115–130 (2013). https://doi.org/10.1016/j.entcom.2012.09.006

[23] Hutchinson, J., Stilinovic, M., Gray, J.E.: Data sovereignty: The next frontier for internet policy? Policy & Internet **16**(1), 6–11 (2024) https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.386. https://doi.org/10.1002/poi3.386

[24] Hao, X., Demir, E.: Artificial intelligence in supply chain decision-making: an environmental, social, and governance triggering and technological inhibiting protocol. Journal of Modelling in Management **19**(2), 605–629 (2023). https://doi.org/10.1108/JM2-01-2023-0009

[25] Lu, X., Burton, H.: Eesd special issue: Ai and data-driven methods in earthquake engineering – (part 2). Earthquake Engineering & Structural Dynamics **52**(11), 3197–3200 (2023) https://onlinelibrary.wiley.com/doi/pdf/10.1002/eqe.3974. https://doi.org/10.1002/eqe.3974

[26] Rossi, D., Di Iorio, A.: Supporting authors in documenting and sharing operative knowledge. Online Information Review **42**(4), 451–467 (2018). https://doi.org/10.1108/OIR-02-2017-0038

[27] Montabert, C., Scott McCrickard, D., Winchester, W.W., Pérez-Quiñones, M.A.: An integrative approach to requirements analysis: How task models support requirements reuse in a user-centric design framework. Interacting with Computers **21**(4), 304–315 (2009). https://doi.org/10.1016/j.intcom.2009.06.003

[28] Wang, Q., Xu, W., Zheng, H.: Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. Neurocomputing **299**, 51–61 (2018). https://doi.org/10.1016/j.neucom.2018.02.095

[29] Xu, J., Tan, P.-N., Zhou, J., Luo, L.: Online multi-task learning framework for ensemble forecasting. IEEE Transactions on Knowledge and

Data Engineering **29**(6), 1268–1280 (2017). https://doi.org/10.1109/TKDE.2017.2662006

[30] Sun, X., Tan, J., Tang, L., Guo, J.J., Li, X.: Real world evidence: experience and lessons from china. BMJ **360** (2018) https://www.bmj.com/content/360/bmj.j5262.full.pdf. https://doi.org/10.1136/bmj.j5262

[31] Gong, T., Choi, J.N.: Effects of task complexity on creative customer behavior. European Journal of Marketing **50**(5-6), 1003–1023 (2016). https://doi.org/10.1108/EJM-04-2015-0205

[32] Liu, Y., Fang, S., Wang, L., Huan, C., Wang, R.: Neural graph collaborative filtering for privacy preservation based on federated transfer learning. The Electronic Library **40**(6), 729–742 (2022). https://doi.org/10.1108/EL-06-2022-0141

[33] Abaoud, M., Almuqrin, M.A., Khan, M.F.: Advancing federated learning through novel mechanism for privacy preservation in healthcare applications. IEEE Access **11**, 83562–83579 (2023). https://doi.org/10.1109/ACCESS.2023.3301162

[34] Zhang, C., Xie, Z., Yu, B., Wen, C., Xie, Y.: Fgss: Federated global self-supervised framework for large-scale unlabeled data. Applied Soft Computing **143**, 110453 (2023). https://doi.org/10.1016/j.asoc.2023.110453

[35] Al-Hejri, A.M., Sable, A.H., Al-Tam, R.M., Al-antari, M.A., Alshamrani, S.S., Alshmrany, K.M., Alatebi, W.: A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors. Scientific Reports **15**(1), 18453 (2025). https://doi.org/10.1038/s41598-025-96527-0

[36] Ying, W., Zhang, L., Luo, S., Yao, C., Ying, F.: Simulation of computer image recognition technology based on image feature extraction. Soft Computing **27**(14), 10167–10176 (2023). https://doi.org/10.1007/s00500-023-08246-1

[37] Mehmood, K., Verleye, K., De Keyser, A., Lariviere, B.: The transformative potential of ai-enabled personalization across cultures. Journal of Services Marketing **38**(6), 711–730 (2024). https://doi.org/10.1108/JSM-08-2023-0286

[38] Broesch, T., Lew-Levy, S., Kärtner, J., Kanngiesser, P., Kline, M.: A roadmap to doing culturally grounded developmental science. Review of Philosophy and Psychology **14**(2), 587–609 (2023). https://doi.org/10.1007/s13164-022-00636-y

[39] Sagot, S., Ostrosi, E., Lacom, P.: Computer-assisted culturalization process integration into product-website design. Journal of Industrial Information Integration **26**, 100252 (2022). https://doi.org/10.1016/j.jii.2021.100252

[40] Brasil, A., Trevisol, J.V.: Research evaluation in brazil and the netherlands: a comparative study. Research Evaluation **34**, 013 (2025). https://doi.org/10.1093/reseval/rvaf013

[41] Hamza, R., Hassan, A., Ali, A., Bashir, M.B., Alqhtani, S.M., Tawfeeg, T.M., Yousif, A.: Towards secure big data analysis via fully homomorphic encryption algorithms. Entropy **24**(4) (2022). https://doi.org/10.3390/e24040519

[42] Zhang, L., Liu, L., Ying, W., Huang, M., Ying, F.: Modhera: A modular kit for parents to take care babies. In: IDC '21: Interaction Design and Children (2021)

Table 3 Cross-Cultural Evaluation Performance by Region and Design Category

Cultural Region	Product Design	Graphic Design	UX/UI Design	Industrial Design	Fashion Design	Interior Design	Architectural Design	Service Design
Western Europe	0.912 ± 0.034	0.889 ± 0.041	0.934 ± 0.028	0.897 ± 0.039	0.856 ± 0.047	0.878 ± 0.042	0.823 ± 0.051	0.901 ± 0.037
North America	0.925 ± 0.031	0.903 ± 0.038	0.947 ± 0.025	0.911 ± 0.036	0.871 ± 0.044	0.892 ± 0.039	0.837 ± 0.048	0.916 ± 0.034
East Asia	0.867 ± 0.042	0.834 ± 0.049	0.891 ± 0.037	0.852 ± 0.045	0.798 ± 0.053	0.819 ± 0.048	0.776 ± 0.056	0.843 ± 0.044
Northern Europe	0.898 ± 0.037	0.875 ± 0.043	0.921 ± 0.031	0.883 ± 0.041	0.842 ± 0.049	0.864 ± 0.044	0.809 ± 0.053	0.887 ± 0.039
Southern Europe	0.881 ± 0.039	0.858 ± 0.045	0.904 ± 0.033	0.866 ± 0.043	0.825 ± 0.051	0.847 ± 0.046	0.792 ± 0.055	0.870 ± 0.041
Oceania	0.854 ± 0.046	0.831 ± 0.052	0.877 ± 0.040	0.839 ± 0.048	0.785 ± 0.057	0.806 ± 0.053	0.763 ± 0.061	0.830 ± 0.047
South America	0.798 ± 0.054	0.775 ± 0.061	0.821 ± 0.048	0.783 ± 0.056	0.729 ± 0.065	0.750 ± 0.060	0.707 ± 0.069	0.774 ± 0.055
South Asia	0.743 ± 0.063	0.720 ± 0.070	0.766 ± 0.057	0.728 ± 0.065	0.674 ± 0.074	0.695 ± 0.069	0.652 ± 0.078	0.719 ± 0.064
Middle East	0.721 ± 0.067	0.698 ± 0.074	0.744 ± 0.061	0.706 ± 0.069	0.652 ± 0.078	0.673 ± 0.073	0.630 ± 0.082	0.697 ± 0.068