Federated Learning-Driven Collaborative Art-Data Analysis And Privacy Protection

Chen Jiang and Xiaotao Luo*

Faculty of Digital-intelligent Urban Construction & Creative Design, Wenhua College, Wuhan, 430074, China

Corresponding author. E-mail: luoxiaotao@whc.edu.cn

Received: Jul. 21, 2026; Accepted: Aug. 14, 2026

Digital art platforms, AIGC systems, and cross-institutional design collaboration platforms continuously generate multimodal artistic creation data that are privacy-sensitive and distributed across many clients. Centralized modeling risks exposure while ordinary federated learning struggles with non-IID styles and gradient leakage. This study proposes SAP-FL, combining local multimodal training with gradient clipping, DP noise, secure aggregation, credibility-weighted aggregation, and compression for cross-node modeling without raw data leaving devices. On non-IID data, SAP-FL achieves F1 of 84.01% (vs. 81.74%/79.11% for FedAvg/DP-FedAvg), reduces membership inference success to 49.70% (20.10 points below FedAvg), and retains F1 of 82.36% at 200 clients. Grad-CAM shows SAP-FL consistently attends to brushstroke-dense regions and subject outlines.

Keywords: differential privacy; secure aggregation; non-IID data; multimodal fusion; Grad-CAM interpretability

http://dx.doi.org/10.6180/jase.202612_35.027

1. Introduction

Generative AI is reshaping business processes and human-computer collaboration [1]. In art, text-to-image systems lower creation barriers and significantly affect creator productivity [2], forming new creative infrastructure that raises ownership and governance issues [3] - making art data complex multimodal information, not just static images.

Centralized machine learning aggregating artwork images, prompts, and logs conflicts with art data's privacy and copyright nature: sketches may contain unpublished designs, and logs may reveal stylistic habits - making privacy-preserving collaborative modeling a core issue for digital art systems.

Federated learning offers a technical route, reducing systemic privacy risk via distributed collaboration while retaining local data [4, 5], providing a methodological basis for modeling art platforms as scalable distributed information systems.

Standard federated learning cannot directly meet art data's needs given significant non-IID characteristics [6]: nodes differ substantially in work category and style, making stable FedAvg aggregation difficult, while membership inference and poisoning risks mean federated learning could itself become a new leakage channel.

This study proposes SAP-FL, framing collaborative art-data analysis and privacy protection within scalable information systems, achieving cross-node modeling through local multimodal training, differential privacy, secure aggregation, credibility weighting, and compression.

Contributions: a scalable federated architecture unifying images, prompts, and logs; a differential-privacy/secure-aggregation mechanism limiting servers to protected updates; a trustworthiness-aware aggregation strategy mitigating non-IID and anomalous-client effects; and an integrated experimental system spanning performance, privacy, communication, scalability, robustness, and interpretability.

Although federated learning avoids uploading raw data, gradients and outputs may still carry sensitive information [7]. For art data, attackers may infer whether specific works were used in training or recover texture structure from model updates even though images never leave local devices.

Privacy-preserving federated learning combines differential privacy, secure multiparty computation, homomorphic encryption, and secure aggregation [8, 9], motivating this paper's inclusion of non-IID distribution and client trustworthiness alongside DP noise.

Federated learning cannot easily achieve security and privacy together [10]; secure aggregation and trusted execution reduce individual exposure [11].

Scalability is central to this study: communication cost directly affects deployability [12] - informing this paper's compression module, since multimodal art models have many parameters and full-gradient uploads scale poorly with client count.

Art creation data is inherently multimodal, requiring joint handling of fusion, missing modalities, and heterogeneity [13]: images reflect style, prompts reflect intention, logs reflect process.

On privacy attacks and interpretability, gradient inversion risks must be quantifiably assessed [14], and heatmap-based interpretation helps assess visual focus reliability [15] - motivating this study's use of Grad-CAM and response heatmaps.

2. Methods

2.1. Distributed Art Creation Data Scenario

This paper considers a federated system with a central server and multiple clients (art schools, galleries, design companies, AIGC platforms) each storing local images, prompts, logs, and labels. The server initializes the model, selects clients, and aggregates without accessing any client's original data. $\mathcal{K} = \{1, 2, \ldots, K\} k D_k$

The local art-creation dataset of the k client is defined as:

$$D_k = \left\{ \left(x_i^k, p_i^k, h_i^k, y_i^k \right) \right\}_{i=1}^{n_k}, \quad k \in \mathcal{K}, \quad n = \sum_{k=1}^{K} n_k. \tag{1}$$

Here the art image, text prompt, behavior log, and category label together form the local sample, with client sample sizes summing to the global total; since clients differ in creative themes and style preferences, the distribution typically violates the IID assumption. $x_i^k p_i^k h_i^k y_i^k n_k n D_k$

Art creation samples span three modalities: an image encoder extracts color, texture, brushstroke, and composition features; a text encoder extracts semantic features from prompts and descriptions; and a behavior encoder extracts edit counts, modification intervals, and version iterations. $f_{\text{img}}(\cdot) f_{\text{txt}}(\cdot) f_{\log}(\cdot)$

$$z_{i,img}^k = f_{img}\left(x_i^k\right), \quad z_{i,txt}^k = f_{txt}\left(p_i^k\right), \quad z_{i,log}^k = f_{log}\left(h_i^k\right). \tag{2}$$

To avoid uneven modality contributions caused by simple concatenation, this study introduces modal attention fusion. Let the modality set be $\mathcal{M} = \{img, txt, log\}$, and the m attention weight for each modality is:

$$\alpha_{i,m}^k = \frac{\exp\left(v^\top \tanh\left(w_m z_{i,m}^k + b_m\right)\right)}{\sum_{q \in \mathcal{M}} \exp\left(v^\top \tanh\left(w_q z_{i,q}^k + b_q\right)\right)}. \tag{3}$$

Fusion is represented as:

$$z_i^k = \sum_{m \in \mathcal{M}} \alpha_{i,m}^k z_{i,m}^k \tag{4}$$

This design enables the model to dynamically adjust the contributions of images, text, and behavior logs based on the characteristics of the samples. For example, for works with distinct styles, the image modality may have a higher weight; for AIGC-generated works, the semantics of the prompts and behavior logs may have more discriminative value.

2.2. Federated Optimization Objectives

The local empirical risk of the k client is defined as:

$$F_k(\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell\left(f_\theta\left(z_i^k\right), y_i^k\right), \tag{5}$$

where θ represents the model parameters, and $\ell(\cdot)$ represents the cross-entropy or task loss function. The global optimization objective is:

$$\min_\theta F(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta). \tag{6}$$

Unlike conventional federated learning, this study's objective jointly optimizes model performance, privacy, and communication efficiency rather than simply minimizing loss, evaluated through a joint view of prediction quality, attack resistance, communication cost, and training stability.

$$\max_{\theta,\rho,\varepsilon} \Omega = \lambda_1 F_1(\theta) - \lambda_2 A_{\text{mia}}(\theta) - \lambda_3 A_{\text{inv}}(\theta) - \lambda_4 \frac{C_{\text{comm}}(\rho)}{C_0} - \lambda_5 \frac{T_{\text{train}}}{T_0} \tag{7}$$

Here the objective terms are membership-inference attack success rate, model inversion risk, communication

overhead, training time, compression ratio, and privacy budget - reflecting focus on scalable systems: models should maintain availability under large client scale and strong privacy constraints, not just small-scale accuracy. A_{mia} A_{inv} C_{comm} T_{train} $\rho\varepsilon$

2.3. Threat Model

This paper defines three threats: an honest-but-curious server inferring data distribution from updates; external attackers intercepting updates for inference or inversion attacks; and anomalous clients uploading offset or poisoned updates. Raw images, prompts, and logs stay local; attackers only access model updates and outputs, not local databases.

2.4. SAP-FL Method Design

SAP-FL's architecture (Fig. 1) has a client layer, local multimodal training layer, privacy protection layer (clipping, DP noise, masking, compression), trusted aggregation layer (secure summation, credibility scoring), and application layer (style recognition, recommendation, privacy auditing) - keeping raw data local and each component replaceable.

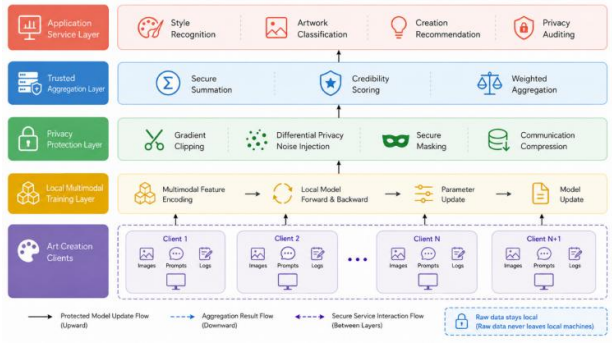

Fig. 1. Scalable federated learning system architecture for distributed art creation data

SAP-FL organizes multimodal modeling, privacy perturbation, secure aggregation, trustaware weighting, and compression into a complete system process rather than simply adding noise to FedAvg. New institutions join as new client nodes without changing the central database or migrating creative data.

SAP-FL addresses three constraints: unstable client availability (random per-round selection); heterogeneous computing resources (avoiding slow-node reliance); and complex access boundaries (internal-only, anonymized, or non-invertible data).

In the first r round of communication, the server selects a subset of clients from the client set to participate in training S_r. After receiving the global parameters, the clients θ^r perform several rounds of training on their local dataset. Considering the non-IID distribution of artistic creation data, this paper incorporates a proximal constraint into the local loss:

$$\mathcal{L}_k(\theta) = F_k(\theta) + \frac{\mu}{2}\|\theta - \theta^r\|_2^2 + \beta\left\|W_f\right\|_F^2. \tag{8}$$

The regularization coefficient for the multimodal fusion parameters controls the degree to which the local model deviates from the global model. The proximal term μ mitigates β local model drift caused by differences in style distribution across different clients. The client-side local update is as follows:

$$\theta_k^{r+1} = \theta^r - \eta\nabla\mathcal{L}_k\left(\theta^r\right). \tag{9}$$

The update process is performed only within the client; the server does not access D_k any of the original samples.

To reduce the identifiable impact of a single sample on model updates, SAP-FL employs a differential privacy mechanism combining gradient clipping and Gaussian perturbation. The client first computes the local gradient g_k^r and then performs ℓ_2 norm clipping:

$$\bar{g}_k^r = \frac{g_k^r}{\max\left(1, \frac{\|g_k^r\|_2}{C}\right)}. \tag{10}$$

Gaussian noise is then added to obtain the protected gradient:

$$\tilde{g}_k^r = \bar{g}_k^r + \mathcal{N}\left(0, \sigma^2 C^2 I\right), \quad \sigma = \frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon}. \tag{11}$$

Here, the clipping threshold and privacy budget jointly determine the strength of privacy protection. A smaller clipping threshold provides stronger privacy protection but introduces more noise, potentially reducing style-recognition performance. This trade-off is examined through response heatmap analysis relating the privacy budget ε, client count, and model performance.

Differential privacy primarily limits the impact at the sample level, but if a server can observe individual client updates over a long period, it may still infer the style distribution of that client. To address this, SAP-FL introduces secure aggregation. Clients k generate paired random masks for other clients $m_{k,j}^r$ and upload them:

$$s_k^r = \tilde{g}_k^r + \sum_{j>k} m_{k,j}^r - \sum_{j<k} m_{j,k}^r. \tag{12}$$

After receiving all s_k^r values uploaded by participating clients, the server performs a summation:

$$\sum_{k\in S_r} s_k^r = \sum_{k\in S_r} \tilde{g}_k^r. \tag{13}$$

Because the pairwise masks cancel each other out during summation, the server can only obtain aggregate updates and cannot directly see the independent gradients of any individual client. For scenarios such as art schools, digital art galleries, and design firms, this mechanism can enhance the trust among institutions participating in collaborative training.

Standard FedAvg primarily weights data based on sample size, but artistic creation clients exhibit differences in data quality, category distribution, update stability, and anomaly risk. SAP-FL calculates a credibility score for each client:

$$q_k^r = \text{sigmoid}\left(\lambda_1 a_k^r + \lambda_2 s_k^r + \lambda_3 c_k^r - \lambda_4 b_k^r\right). \tag{14}$$

where a_k^r represents local verification performance, s_k^r represents update stability, c_k^r represents consistency between the client update direction and the global direction, and b_k^r represents anomaly risk. The aggregate weight is defined as:

$$w_k^r = \frac{n_k q_k^r}{\sum_{j \in S_r} n_j q_j^r} \tag{15}$$

The global model is updated as follows:

$$\theta^{r+1} = \sum_{k \in S_r} w_k^r \theta_k^{r+1} \tag{16}$$

Credibility-weighted aggregation assigns greater weight to stable and high-contributing clients, while mitigating the impact of anomalous clients or low-quality updates. This mechanism directly addresses the robust aggregation problem in non-IID artistic creation data scenarios.

To accommodate large-scale client expansion, SAP-FL performs sparsity and quantization compression on protected updates. Let a compression function $Q_\rho(\cdot)$ represent the compression ratio ρ, then the client-uploaded update is:

$$u_k^r = Q_\rho\left(\tilde{g}_k^r\right), \quad C_r = \sum_{k \in S_r} \text{bits}\ \left(u_k^r\right). \tag{17}$$

Total communication volume is:

$$C_{\text{total}} = \sum_{r=1}^{R} C_r = \sum_{r=1}^{R} \sum_{k \in S_r} \text{bits}\ \left(Q_\rho\left(\tilde{g}_k^r\right)\right). \tag{18}$$

Moderate compression ratios significantly reduce communication volume with stable performance, but excessive compression loses key gradient information, slowing convergence

- so this paper treats compression ratio as a simulation variable alongside privacy budget and client count.

2.5. Algorithm Flow

Fig. 2 shows the SAP-FL workflow: the server initializes the model and selects clients; each client trains locally; gradients are clipped, DP-perturbed, compressed, and securely uploaded; the server computes credibility scores and completes weighted aggregation, repeating until convergence.

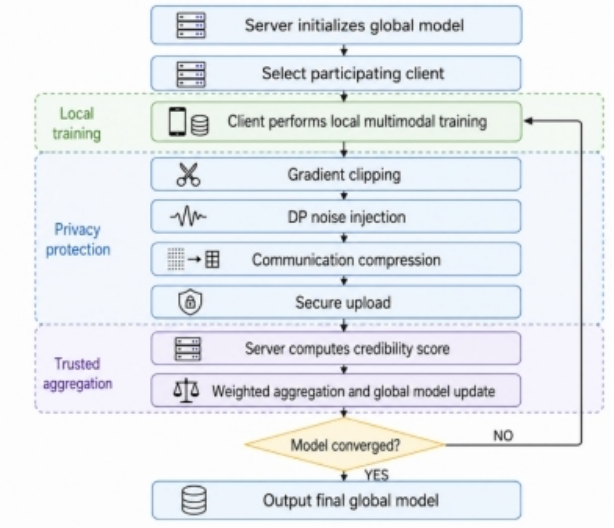

Fig. 2. SAP-FL Privacy-Preserving Federated Training Flowchart

Algorithm 1: the server initializes the model; each round selects clients that train locally; gradients are clipped, DP-perturbed, compressed, and uploaded; the server computes credibility scores and updates aggregate weights; the global model is returned. $\mathcal{K} R \varepsilon C \rho \eta \theta^0 S_r \theta^R$

2.6. Datasets and Federated Partitioning

This paper constructs a distributed multimodal art creation dataset: image data from public art style images (Impressionism, Cubism, Expressionism, Ukiyo-e, Realism, Abstract, Baroque, Surrealism); text from titles and prompt fragments; and behavioral logs (edit counts, redraw counts, version iterations) for style recognition and category classification.

To simulate cross-institutional scenarios, samples are divided across 10-200 clients with Dirichlet parameters 1.0-0.1 controlling non-IID degree; 30% of clients are selected per round over 150 communication rounds, 5 local rounds, batch size 64, learning rate 0.001, with all methods following the same settings so differences reflect privacy, aggregation, and compression mechanisms. α

Images are resized and standardized via cropping/padding; prompts are tokenized and semantically encoded; behavior logs are normalized into editing inten-

sity and rhythm features, with all three modalities retained throughout training.

2.7. Comparison Methods and Indicators

Seven methods are compared: Centralized Training (upper-bound reference, no privacy); Local Training (independent, no collaboration); FedAvg (basic averaging); FedProx (proximal constraints for non-IID); DP-FedAvg (differential privacy); Secure Aggregation FL (secure aggregation without trust-weighting/compression); FedAvg-Compression (compression added); and SAP-FL.

Evaluation metrics span model performance (Accuracy, Precision, Recall, F1), privacy/security (membership-inference success, inversion similarity), scalability (client count, communication overhead, convergence, training time), robustness (F1 under anomalous-client ratio), and interpretability (Grad-CAM overlap with annotated style regions).

This paper distinguishes "performance improvement" from "system usability": high F1 at small scale with rapidly growing communication cost or fluctuation under anomalous clients remains impractical. SAP-FL incorporates privacy, communication efficiency, and robust aggregation into system constraints simultaneously for practical deployability.

3. Results and discussion

3.1. Results and Analysis

Table 3 shows non-IID results: Centralized training has highest performance but fails privacy; Local Training reaches only 70.94% F1; FedAvg has high attack success; DP-FedAvg drops to 79.11% F1 from noise; SAP-FL achieves 84.01% F1 with attack success reduced to 49.70% and communication overhead ratio 0.42.

SAP-FL doesn't simply trade performance for privacy: versus FedAvg, F1 improves 2.27 points while attack success drops 20.10 points; versus DP-FedAvg, F1 improves 4.90 points, showing trusted aggregation and multimodal fusion mitigate noise-induced degradation.

4. Convergence analysis under iid and non-iid conditions

Fig. 3 shows convergence on non-IID data: FedAvg accelerates early but oscillates later; DP-FedAvg converges slower from noise; FedProx improves stability via proximal constraints; SAP-FL stabilizes after ˜90 rounds with the highest F1. Client-level style bias affects standard federated learning convergence stability. SAP-FL's advantages stem from

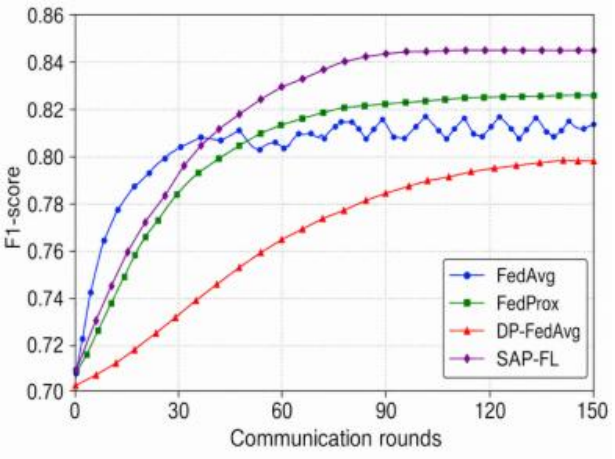

Fig. 3. Convergence curves of different federated learning methods on non-IID artistic creation data

local proximal constraints reducing style drift and credibility weighting reducing low-quality/directionally-shifted update impact - suiting cross-institutional art-data training.

Response Heatmap Analysis Based on Privacy Budget and Client Scale

Fig. 4's response heatmap shows privacy budget (horizontal) versus client count (vertical) with F1-score color, comparing DP-FedAvg and SAP-FL as a system-level diagnostic tool linking privacy budget and scale to model behavior. ε

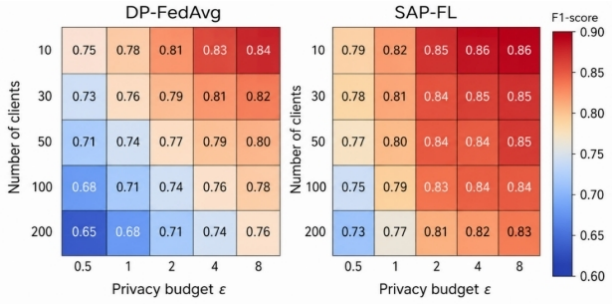

Fig. 4. Heatmap of F1-score response of different algorithms under varying privacy budgets and number of clients

Both methods perform well with few clients and large privacy budgets, but DP-FedAvg deteriorates more as clients increase and budget decreases; SAP-FL stays more stable in the 50-100 client region, adapting better to the privacy-scale trade-off. $\varepsilon = 2\varepsilon = 4$

Table 4 shows FedAvg, DP-FedAvg, Secure Aggregation FL, and SAP-FL under membership inference/model inversion attacks: FedAvg has highest success with no privacy mechanism; DP-FedAvg reduces success but lowers post-

Table 1. Experimental data and federated training parameter settings

Parameter categories	Settings
Number of clients	10, 30, 50, 100, 200
Participation percentage in each round	30%
Communication rounds	150
Local training rounds	5
Batch size	64
Learning rate	0.001
Non-IID partitioning	Dirichlet $\alpha = 1.0, 0.5, 0.3, 0.1$
Privacy Budget	$\varepsilon = 0.5, 1, 2, 4, 8$No DP / varied privacy budget
Gradient clipping threshold	1.0
Communication compression rate	0%, 50%, 70%, 90%
Evaluation task	Artistic style identification and work category classification

Table 2. Comparison Methods and Evaluation Indicators

Category	**Content**
Comparison Methods	Centralized Training, Local Training, FedAvg, FedProx, DP-FedAvg, Secure Aggregation FL, FedAvg-Compression, and SAPFL
Performance Indicators	Accuracy, Precision, Recall, and F1-score
Privacy Metrics	Membership inference attack success rate and model inversion similarity
Scalability Metrics	Number of clients, single-round communication volume, total communication volume, and convergence rounds
Robustness Metrics	Performance changes under anomalous client conditions of 5%, 10%, and 20%
Interpretability Metrics	Overlap ratio between Grad-CAM focus regions and annotated regions

Table 3. Comparison of overall performance and privacy security of different methods

Metho d	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Members hip inference attack success rate (%)	Communi cation overhead ratio
Central ized Trainin g	86.42	86.10	85.73	85.91	73.20	-
Local Trainin g	71.83	71.40	70.56	70.94	58.10	0.00
FedAv g	82.36	82.05	81.44	81.74	69.80	1.00
FedPr ox	83.41	83.12	82.65	82.88	68.90	1.00
DPFedAv g	79.92	79.50	78.86	79.11	55.30	1.00
Secure Aggreg ation FL	81.65	81.20	80.75	80.97	62.40	1.12
FedAv gCompr ession	80.74	80.36	79.86	80.05	66.20	0.38
SAPFL	84.58	84.27	83.76	84.01	49.70	0.42

attack F1; Secure Aggregation FL limits server observation but has weak output defense; SAP-FL combines all three for lowest attack success.

SAP-FL's inversion similarity is 0.261 , lower than FedAvg's 0.354 , showing better suppression of image/style leakage, with stronger post-attack F1 stability too.

Fig. 5 shows F1 as clients grow from 10 to 200: FedAvg, DP-FedAvg, and FedAvg-Compression decline significantly, while SAP-FL declines only slightly, showing trust-aware weighting and compression improve large-scale scalability.

Standard FedAvg is susceptible to distribution/quality variation at scale; DP-FedAvg is further hurt by noise at 200 clients; SAP-FL maintains F1 above 82% at 200 clients, suiting both small experiments and large-scale deployment.

Table 5 shows SAP-FL's performance and overhead across compression ratios: at 70% compression, overhead ratio drops to 0.42 while F1 stays at 84.01%; at 90% compression, overhead drops further but F1 falls to 81.38% with convergence epochs rising to 120.

Higher compression isn't always better - ˜70% balances efficiency and performance for this scenario; excessive compression loses gradient information, degrading style recognition, so compression ratio should be chosen by bandwidth, client scale, and privacy budget.

Table 6 shows F1 under anomalous-client ratios (noise/direction-offset updates simulating low-quality/malicious nodes): at 20% anomalous clients,

Table 4. Privacy Attack Defense Results

Method	Membership Inference Attack Success Rate (%)	Model Inversion Similarity	F1-Score (%) After Attack
FedAvg	69.80	0.354	79.20
DP-FedAvg	55.30	0.274	76.84
Secure Aggregation FL	62.40	0.316	78.65
SAP-FL	49.70	0.261	82.57

Table 5. Performance and Communication Overhead of SAP-FL at Different Compression Ratios

Compression Ratio	F1-Score (%)	Communication Overhead Ratio	Convergence Rounds	Membership Inference Attack Success Rate (%)
0%	84.92	1.00	82	51.30
50%	84.61	0.56	85	50.40
70%	84.01	0.42	92	49.70
90%	81.38	0.19	120	48.90

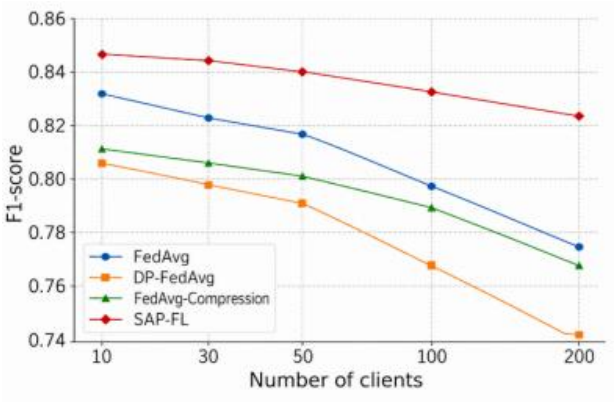

Fig. 5. Comparison of F1-score scalability under different numbers of clients

FedAvg drops to 68.90%, FedProx to 70.55%, DP-FedAvg to 70.12%, while SAP-FL remains at 79.73%, with only the anomalousclient ratio varied to isolate robustness from privacy noise.

Credibility-weighted aggregation identifies clients with unstable or directionally-deviant updates and reduces their weight - crucial for open art platforms where device, data-quality, and behavior differences are inevitable.

Table 7's ablation isolates each component: removing DP slightly improves F1 but raises attack success; removing secure aggregation also raises attack success; removing trust-aware weighting reduces performance at 20% anomalous clients; removing compression restores overhead to 1.00.

Differential privacy mainly reduces attack success; secure aggregation reduces single-client exposure; trust-aware weighting mainly contributes robustness; compression mainly contributes scalability. Full SAP-FL isn't highest on F1 alone but achieves the most balanced result across privacy, communication, and robustness.

Fig. 6 compares Grad-CAM visualizations across methods: Local Training's attention is dispersed toward background; FedAvg captures some style areas with local bias; DP-FedAvg shows diffuse regions from noise; SAP-FL concentrates on brushstroke-dense regions, texture variation, and subject outlines.

Fig. 6. Comparison of Grad-CAM regions of interest for different algorithms in the art style recognition task

SAP-FL performs well numerically while maintaining reasonable visual attention - focusing on brushstrokes, textures, and composition is more interpretable than raw category output, supporting effective style discrimination un-

Table 6. F1 score under anomalous client ratio changes

Abnormal client ratio	FedAvg	FedProx	DP-FedAvg	SAP-FL
0%	81.74	82.88	79.11	84.01
5%	79.06	80.23	77.85	83.28
10%	75.41	77.10	75.34	82.14
20%	68.90	70.55	70.12	79.73

Table 7. Results of SAP-FL Ablation Experiments

Method	**F1-Score (%)**	**Membership Inference Attack Success Rate (%)**	**Communication Overhead Ratio**	**F1-Score (%) When 20% of Clients Are Abnormal**
SAP-FL w/o DP	85.04	66.70	0.42	80.06
SAP-FL w/o SA	84.36	58.90	0.42	80.48
SAP-FL w/o Trust	81.92	50.30	0.42	72.41
SAP-FL w/o Compression	84.92	51.30	1.00	80.38

der DP perturbation.

Fig. 7 shows privacy budget and compression jointly affect attack success: lower compression reduces success and limits exploitable information, but excessive compression only slightly reduces success while significantly impairing performance. $\varepsilon\varepsilon$

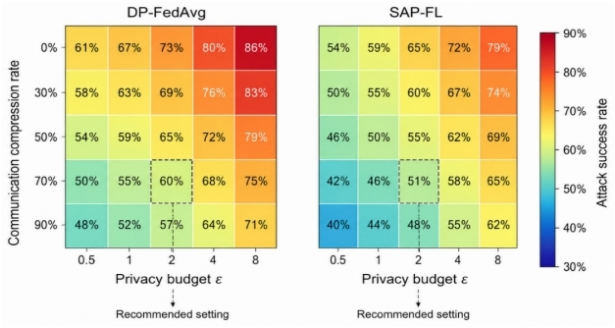

Fig. 7. Heatmap of attack success rate response of different algorithms under changes in privacy budget and communication compression rate

Combining Fig. 4 and Fig. 7, a 70% compression rate and 30% client participation rate balance F1-score, attack success, and communication overhead for this scenario - a recommended configuration for the simulated setting, not a universal optimum. $\varepsilon = 2$

4.1. Discussion

This paper's core challenge is privacy-preserving collaborative modeling within scalable distributed information systems: clients scale horizontally, training uses local computation, and metrics cover performance, privacy, communication, and robustness - aligning with scalable information systems rather than art theory alone.

This framing also distinguishes the study from ordinary image-classification papers: the work focuses on secure collaborative training in multi-node information systems, rather than offline recognition models on a single dataset.

SAP-FL applies broadly: schools train recognition without sharing student artwork; galleries train classification without exposing high-resolution images; AIGC platforms protect prompts/logs; design companies join training without disclosing commercial sketches; and Grad-CAM helps verify the model focuses on style rather than watermarks or background.

Limitations: DP noise affects performance especially at small privacy budgets, motivating future adaptive allocation; this study uses only image, text, and behavior modalities, not audio/video/3D; Grad-CAM interprets visual models only; and experiments simulate a distributed environment, with real deployment and dynamic client joining left for future work on real multi-institutional datasets. ε

5. Conclusions

- This paper proposes SAP-FL, a scalable, privacy-preserving federated learning method for distributed, multimodal art-creation data that models images, text prompts, and creative behavior logs locally on each client and enables cross-node collaborative analysis without uploading raw data, through differential privacy, secure aggregation, credibility-aware weighting, and communication compression.

- On non-IID art-creation data, SAP-FL achieves a higher F1-score than FedAvg and DP-FedAvg while substantially reducing the success rate of membership

inference and model inversion attacks, achieving a better balance between model performance and privacy protection.

- SAP-FL remains stable and scalable as the number of clients grows from 10 to 200, as privacy budgets and communication compression ratios change, and when anomalous clients are injected, confirming its suitability for large-scale, heterogeneous collaborative deployments.
- Grad-CAM comparisons show that SAP-FL continues to focus on key artistic-style regions under privacy perturbation, and the Response Heatmap analysis reveals how privacy budget, client scale, and communication compression jointly shape system performance and attack resistance.
- Overall, SAP-FL provides digital art platforms, AIGC creation systems, digital art galleries, and cross-institutional art-data collaborations with a practical technical pathway for jointly achieving collaborative data analysis, privacy protection, and system scalability.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, (2024) *"Generative AI"* **Business & Information Systems Engineering** ***66***(1): 111–126. DOI: 10.1007/s12599-023-00834-7.

[2] E. Zhou and D. Lee, (2024) *"Generative Artificial Intelligence, Human Creativity, and Art"* **PNAS Nexus** ***3***(3): DOI: 10.1093/pnasnexus/pgae052.

[3] Z. Epstein, A. Hertzmann, Investigators of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, R. Mahari, A. Pentland, O. Russakovsky, H. Schroeder, and A. Smith, (2023) *"Art and the Science of Generative AI"* **Science** ***380***(6650): 1110–1111. DOI: 10.1126/science.adh4451.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. El Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, S. Zhao, et al., (2021) *"Advances and Open Problems in Federated Learning"* **Foundations and Trends in Machine Learning** ***14***(1–2): 1–210. DOI: 10.1561/2200000083.

[5] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, (2021) *"A Survey on Federated Learning"* **Knowledge-Based Systems** ***216***: DOI: 10.1016/j.knosys.2021.106775.

[6] H. Zhu, J. Xu, S. Liu, and Y. Jin, (2021) *"Federated Learning on Non-IID Data: A Survey"* **Neurocomputing** ***465***: 371–390. DOI: 10.1016/j.neucom.2021.07.098.

[7] N. Bouacida and P. Mohapatra, (2021) *"Vulnerabilities in Federated Learning"* **IEEE Access** ***9***: 63229–63249. DOI: 10.1109/ACCESS.2021.3075203.

[8] X. Yin, Y. Zhu, and J. Hu, (2021) *"A Comprehensive Survey of Privacy-Preserving Federated Learning: A Taxonomy, Review, and Future Directions"* **ACM Computing Surveys** ***54***(6): 1–36. DOI: 10.1145/3460427.

[9] A. El Ouadrhiri and A. Abdelhadi, (2022) *"Differential Privacy for Deep and Federated Learning: A Survey"* **IEEE Access** ***10***: 22359–22380. DOI: 10.1109/ACCESS.2022.3151670.

[10] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martínez, D. Sánchez, A. Flanagan, and K. E. Tan, (2021) *"Achieving Security and Privacy in Federated Learning Systems: Survey, Research Challenges and Future Directions"* **Engineering Applications of Artificial Intelligence** ***106***: DOI: 10.1016/j.engappai.2021.104468.

[11] J. Chen, H. Yan, Z. Liu, M. Zhang, H. Xiong, and S. Yu, (2024) *"When Federated Learning Meets Privacy-Preserving Computation"* **ACM Computing Surveys** ***56***(12): 1–36. DOI: 10.1145/3679013.

[12] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, (2021) *"Communication-Efficient Federated Learning"* **Proceedings of the National Academy of Sciences** ***118***(17): DOI: 10.1073/pnas.2024789118.

[13] L. Che, J. Wang, Y. Zhou, and F. Ma, (2023) *"Multimodal Federated Learning: A Survey"* **Sensors** ***23***(15): DOI: 10.3390/s23156986.

[14] A. Hatamizadeh, H. Yin, P. Molchanov, A. Myronenko, W. Li, P. Dogra, A. Feng, M. G. Flores, J. Kautz, D. Xu, and H. R. Roth, (2023) *"Do Gradient Inversion Attacks Make Federated Learning Unsafe?"* **IEEE Transactions on Medical Imaging** ***42***(7): 2044–2056. DOI: 10.1109/TMI.2023.3239391.

[15] K. Szczepankiewicz, A. Popowicz, K. Charkiewicz, K. Nałęcz-Charkiewicz, M. Szczepankiewicz, S. Lasota, P. Zawistowski, and K. Radlak, (2023) *"Ground Truth Based Comparison of Saliency Maps Algorithms"* **Scientific Reports** ***13***: DOI: 10.1038/s41598-023-42946-w.