RELIC: Federated Conditional Textual Inversion with Prototype Alignment

Sijia Chen, Ningxin Su, Baochun Li
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Ontario, Canada
sjia.chen@mail.utoronto.ca, ningxin.su@mail.utoronto.ca, bli@ece.toronto.edu

***Abstract*—Text-to-image models can generate personalized images with unprecedented freedom by using a pseudo-word learned from a few images, using a novel technique called *textual inversion*. It is conceivable that, in the spirit of federated learning, multiple users wish to learn a pseudo-word based on their local images collaboratively. However, how a more effective pseudo-word can be trained in the context of federated learning remains unclear.**

In this paper, our experiments show that such federated textual inversion is neither secure nor feasible. First, once one client exposes its pseudo-word embedding to the server for aggregation, an attacker can directly generate similar images to this client. Second, training one shared pseudo-word without personalization hinders individuals from generating images that exhibit local characteristics. Finally, after global aggregation, the averaged pseudo-word embedding may lose learned concepts. Motivated by these insights, we propose RELIC, a new framework that encompasses federated conditional textual inversion with prototype alignment. With privacy guarantees, RELIC allows clients to learn personalized pseudo-words conditional on local samples while enforcing a globally consistent clustering of clients' pseudo-words into discriminable prototypes instead of averaging. The experiments conducted on both i.i.d. and extreme non-i.i.d. data demonstrate that RELIC is able to achieve state-of-the-art performance as compared to baseline approaches.

I. INTRODUCTION

Large-scale text-to-image models, most notably Stable Diffusion models [1], have demonstrated an unprecedented capability to synthesize high-fidelity images described by natural language descriptions. In order to train them for specific use cases, the pre-trained models are fine-tuned by only training task-specific layers with a set of reference images.

Recently, the fine-tuning method, *textual inversion* (TI) [2], has garnered overwhelming attention because it simply finds embedding vectors for new concepts of a few personalized images in the textual embedding space. That is, given a text prompt "a photo of {}", where {} is the placeholder called the *pseudo-word*, TI optimizes the *embedding vector*. Such an embedding vector often weighs only a few hundred kilobytes and is combined with the pseudo-word to reconstruct the corresponding personalized images. Eventually, the learned pseudo-word can be composed with textual queries, such as "an oil painting of {}", to generate novel and diverse personalized images.

However, in practice, when a client of TI only has a limited number of images depicting a single aspect of its personalized concept, it is challenging to train the pseudo-word capable of generating high-quality and faithful personalized images. For instance, the pseudo-word trained only on one shape of flowers may not directly generate high-fidelity images of the remaining shapes. In this case, an effective pseudo-word optimization requires the participation of more clients, each possessing personalized images. Yet, it is still unclear how to enable them to conduct effective pseudo-word training in a federated learning [3] (FL) context to preserve data privacy.

A natural solution is performing textual inversion under federated learning, which we can refer to as *federated textual inversion*. This approach treats the learnable embedding vector of the pseudo-word as the global model, thus allowing clients to train cooperatively using the FL paradigm. In the context of FL, pseudo-word training can benefit from diverse and more extensive images distributed among many clients. At first glance, Wang *et al.* [4] proclaimed that the threat to client privacy is significantly reduced when only model updates are transmitted from clients to the server. As a result, federated textual inversion is expected to be efficient and secure.

However, empirical observations from our own experimental results using this approach show that clients participating in training not only expose their local images but also produce an invalid pseudo-word after training. During the local update of one client, the vector embedding of the pseudo-word can be trained over a few epochs to capture the core concepts of local images. Consequently, once the updated embedding is sent to the server, anyone can regenerate similar images by forwarding "a photo of a {}" through a pre-trained stable diffusion model. Hence, the more high-quality local images are used for training, the greater the client's risk of data leakage. To make matters worse, after receiving the pseudo-words from the clients, the server aggregates them by performing federated averaging [3]. As each vector embedding acts as the features in the textual embedding space of a pre-trained model [2], arithmetically averaging them may lead to embeddings with invalid concepts.

In this paper, we propose RELIC, a new federated learning framework specifically tailored to the needs of textual inversion. The key insights in RELIC are the *conditional pseudo-word* (C-PW) and *prototype alignment*, working cooperatively to address the three major issues we introduced: privacy leakage, the lack of personalization, and failure of server aggregation using averaging.

On the one hand, a *conditional psuedo-word* (C-PW) with a local shift-scale structure for each client is designed to allow clients to train a global pseudo-word while still allowing each client to learn a personalized pseudo-word that is *conditional* upon local samples. *Prototype alignment*, on the other hand, is specifically proposed to avoid concept damage during aggregation. As text embeddings of the pseudo-word with different textual prompts are naturally clustered, centroids, a.k.a. prototypes [5], [6], of these clusters, can be maintained consistent among clients to preserve the learned concept.

The original contributions in this paper are as follows. *First*, we are the first to provide insights into the infeasibility and privacy concerns associated with achieving textual inversion in the context of federated learning. *Second*, we propose Relic, a novel framework incorporating a conditional pseudo-word structure and prototype alignment to train an effective global pseudo-word while learning personalized pseudo-words with privacy guarantees. *Finally,* we evaluate Relic on well-known text-to-image datasets, including `Flowers102` [7] and `Celeba` [8], under both i.i.d. and extreme non-i.i.d. settings. The experimental results demonstrate the superior performance of Relic on qualitative and quantitative metrics.

II. Related Work

Stable Diffusion [1] models, performing text-to-image on latent space, are one prominent generative model used for generating high-fidelity images based on text descriptions. Due to the effectiveness of the pre-trained models, parameter-efficient fine-tuning techniques such as LoRA [9] and CoCoOp [10] have been proposed to optimize large models for downstream tasks with small-scale datasets. Research on the large model fine-tuning in the context of federated learning (FL) proposed by the seminal work of FedNLP [11] became more important. Especially, the works [12], [13], [14] have been actively investigating fine-tuning pre-trained large models by training a new textual prompt under the federated environment. However, no work has been proposed to study how to fine-tune stable diffusion models in FL.

Thus, our paper focuses on performing the text-to-image generation task by fine-tuning a pre-trained stable diffusion model [1] with textual inversion [2]. This technique is capable of training a new textual embedding, referred to as a pseudo-word, towards capturing novel concepts from a few images. To the best of our knowledge, achieving this objective through textual inversion under federated learning settings remains largely unexplored. To fill this gap, based on our empirical observations and evaluations of the textual inversion in the context of production FL, we are motivated to devise Relic to learn an effective pseudo-word while preserving privacy. It deserves to be pointed out that the prototype structure of Relic is also motivated by the federated prototype learning framework [6] in which clients send prototypes of local representations to the server.

III. Preliminaries and Motivations

A. Federated Textual Inversion

Federated textual inversion (FedTI) treats the embedding vector $\mathbf{u}$ of the pseudo-word in *Textual inversion* [2] as the global model parameterized by $\boldsymbol{w}$ and aims to optimize it based on the federated training paradigm. For the stable diffusion model, a textual prompt $\boldsymbol{x}_t$ is processed to be text embedding $\mathbf{U} = f_t\left(\boldsymbol{x}_t\right)$ and encoding $\mathbf{F}_t = f_e\left(\mathbf{U}\right) \in \mathbb{R}^d$.

Prompt template. $\boldsymbol{x}_t$ is the concatenation of a prompt template, such as “an oil photo of {}”, selected from the template pool $\mathcal{T}$ and $S^\star$. This yields $\mathbf{U} = [\mathbf{v}_1, \ldots, \mathbf{v}_N, \mathbf{u}_\star]$, where $[\mathbf{v}_1, \ldots, \mathbf{v}_N]$ denotes the template’s vector sequence comprising N tokens, and $\mathbf{u}^\star = f_t(S^\star)$. When selecting the j-th prompt template from $\mathcal{T}$, we obtain $\boldsymbol{x}_{t_j}$, along with its embedding $\mathbf{U}_{t_j}$ and encoding $\mathbf{F}_{t_j}$.

Local update. In each training iteration, $\mathbf{F}_t$ are processed by the stable diffusion model to reconstruct the ground truth image $\mathbf{F}_v$, leading to the image reconstruction loss $\mathcal{L}_{recon}$. Therefore, $\boldsymbol{w}^c$ is optimized by minimizing $\mathcal{L}_{recon}$ on the local dataset D^c for E epochs. In each step k with sample set ξ, we have $\boldsymbol{w}^c_{k+1} = \boldsymbol{w}^c_k - \eta_k g\left(\boldsymbol{w}^c_k\right)$ where $g\left(\boldsymbol{w}^c_k\right) := \nabla L\left(\boldsymbol{w}^c_k, \xi\right)$ is the stochastic gradient. After E epochs, each client sends the $\mathbf{u}^c$ to the server to get the aggregated model $\bar{\mathbf{u}}_{r+1} := \boldsymbol{w}_{r+1}$.

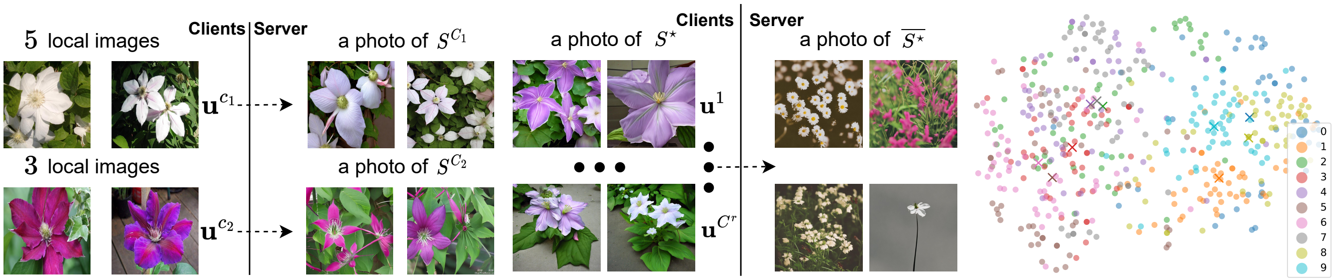
B. Client Data Distribution Simulation

Unlike conventional FL, which primarily addresses image classification, this paper focuses on personalized text-to-image generation, which does not embrace labels but utilizes text and image pairs during training. Therefore, for FedTI, we simulate the data distribution of clients by assigning images from one class to different clients and following the rule that each client has a few images presenting one aspect of this class. Apart from this foundation, for simplicity, we follow the terminologies of FL, including i.i.d. and non-i.i.d. data. Clients have i.i.d. data when they have the same number of local images. Otherwise, when the size of local images varies according to Dirichlet distribution, it is referred to as non-i.i.d. data — quantity skew among local samples of clients. In this paper, we extract samples of 112 flower images with the label “clematis” from `Flowers102` [7] and 36 face images with ID 2820 from `Celeba` [8].

C. Motivations

To gain more insights into the FedTI framework, we rely on a pre-trained Stable Diffusion v2 [1] and utilize 50 clients, selecting 5 of them each round, to train $\mathbf{u}_\star$ for 30 rounds with settings $E = 2$. To facilitate the evaluation, each client contains images presenting one color or style of the “clematis” flower, as shown in Fig. 1(a) and 1(b).

Limitation 1: Clients leak their private images once their trained local pseudo-word embeddings are sent to the server. The left subfigure of Fig. 1(a) presents two random clients containing 3 and 5 local images, respectively. Their pseudo-word embeddings, denoted as $\mathbf{u}^{c_1}$ and $\mathbf{u}^{c_2}$, are sent to the server side for global aggregation. Consequently, once the prompt template “a photo of {}” is used, anyone can generate



(a) Privacy leak (b) Model averaging compromises the concepts (c) Observation

Fig. 1: Illustration of limitations and our observations. (a). Visualization of privacy leakage when clients transmit local pseudo-words to the server. (b). Comparison of "clematis" flower images generated by clients' local pseudo-word embeddings and images generated by the pseudo-word embedding after global aggregation. (c). Illustrations of 2D t-SNE embeddings depicting pseudo-word encodings collected from 50 clients. We present prompt templates with indexes $[0, 9]$ for clarity.

images that closely resemble the clients' local images, as shown in the right subfigure of Fig. 1(a). Compared to client c_2, client c_1, which trains its pseudo-word embedding with more images, has more privacy exposure due to the higher quality of the generated images.

Limitation 2: Averaging local pseudo-word embeddings from clients for global aggregation can undermine learned concepts. Fig. 1(b) presents the images generated on each client based on local pseudo-word $\{\mathbf{u}^c\}_{c=1}^{C_r}$ and those produced on the server after global aggregation $\overline{\mathbf{u}}_{r+1}$. Compared to locally generated images shown in the left subfigure of Fig. 1(b), images generated by using $\overline{\mathbf{u}}_{r+1}$ illustrate common flower features without presenting any specifics of "clematis". This also shows that with only one global pseudo-word, each client is able to adjust it towards presenting local characteristics for personalization.

Observation: Encodings of pseudo-words across clients belong to distinct clusters, each corresponding to one prompt template. In order to generate images, the pseudo-word should be combined with a prompt template from $\mathcal{T}$ to create the prompt t_j. By plotting the encodings of these clients' combinations in 2D space, we show that they are clustered around the corresponding prompt template, as shown in Fig 1(c). For instance, with the prompt template indexed by $j = 9$, the encodings of pseudo-words from 50 clients are distributed around the centroid plotted as a cyan cross. Thus, preserving the cluster centroids of pseudo-words can maintain the learned concept.

IV. Framework Design

A. Overview

Relic adheres to the FTI framework while incorporating two novel designs, including *conditional pseudo-word* (C-PW) and *prototype alignment mechanism*, which are inspired by our discussions in III-C.

The shift-scale structure of C-PW is motivated by the CoCoOp [10] approach. Specifically, for the client with index c, the shift branch parameterized by $\boldsymbol{W}_s^c$ is to compute the shift value conditional on the visual features $\mathbf{F}_v$. The scale branch parameterized by $\boldsymbol{W}_t^c$ is to produce a scale value based on $\mathbf{V}$. The conditional pseudo-word $\widehat{\mathbf{u}}^c$ is computed as $\widehat{\mathbf{u}}^c = \boldsymbol{W}_s^c \mathbf{F}_v + \mathbf{u}^c \times \boldsymbol{W}_t^c \left[\mathbf{v}_1, ..., \mathbf{v}_{N^j}\right]^T$, where $\{\mathbf{v}_n\}_{n=1}^{N^j}$ is text embeddings of the selected j-th prompt template from $\mathcal{T}$.

This lightweight shift-scale structure is simple but effective. It inherently enables each client to adjust the pseudo-word adaptively based on its local images and the textual prompt template. C-PW learns common concepts with the global pseudo-word while enabling the shift-scale structure to capture and maintain client-specific characteristics.

Inspired by previous works [15], [6], we design a prototype alignment mechanism to capture globally consistent concepts for the pseudo-word, thereby mitigating the issue in model averaging.

First, for a group of samples D, we define a prototype $\zeta(j)$ as the centroid, which is computed as $\zeta(j) = \frac{1}{|D_j|} \sum_{\boldsymbol{x}_v \in D} \mathbf{F}_{t_j} I(\boldsymbol{x}_{t_j})$, where $\zeta(j) \in R^d$, $I(\boldsymbol{x}_{t_j})$ outputs 1 when j-th template is selected, and D_j contains samples with j-th template. Each client with index c computes $\zeta^c(j)$ with $j \in \mathcal{T}_r^c$, where $\mathcal{T}_r^c \subset \mathcal{T}$ as one client may use only a subset of templates in round r. Thus, prototype aggregation for C^r clients is $\overline{\zeta}^c(j) = 1 \,/\, |C_j^r| \sum_c \zeta^c(j) I(j \in \mathcal{T}^c)$.

Each client with index c performs **global-local** alignment by performing exponential moving average (EMA), shown as $\overline{\zeta}(j) = m\overline{\zeta}(j) + (1-m)\overline{\zeta}^c(j)$. Subsequently, we design the regularization term as the auxiliary loss $\mathcal{L}_{proto}$ to encourage the encodings of produced prompts with the specific template to approach its global prototype. This allows the loss function to be formulated as $\mathcal{L} = \mathcal{L}_{recon} + \lambda_0 \mathcal{L}_{proto}$, where $\mathcal{L}_{proto} = 1 \,/\, B \sum_{t_j \in B} L_2\left(\mathbf{F}_{t_j}, \overline{\zeta}(j)\right)$, B is a batch of samples and L_2 is the L2 distance.

B. Convergence Analysis

A theoretical guarantee of convergence is required to ensure the effectiveness of the new training algorithm 1 and loss function of Relic. Thus, we establish a convergence rate for the local objective function $\mathcal{L}^c$ under the smooth, non-convex setting. Following the assumptions as existing frameworks [6], we get an ergodic convergence rate.

Algorithm 1: Learning algorithm on each client

```
Input: Learning rate $\eta$, local epochs $E$, batch size $B$,
       Regularization coefficient $\lambda_0$.
Output: $\boldsymbol{w}^c$.
Function $LocalUpdate(\boldsymbol{w}_r, \{\overline{\zeta}(j)\}_{j \in \mathcal{T}})$:
    $\lambda \leftarrow \lambda_0$; if $\zeta^c(j) = \mathbf{0}$ then $\lambda = 0$ ;
    for $j \in \mathcal{T}_r$ do Global-local alignment;
    $\boldsymbol{w}^c \leftarrow \boldsymbol{w}_r$;
    Get one batch of samples $B \in D$;
    for sample $\boldsymbol{x} \in B$ do
        Get $\boldsymbol{x}_{t_j}$, obtaining $\{\mathbf{v}_n\}_{n=1}^{N^j} = f_t(\boldsymbol{x}_{t_j})$;
        Get $\widehat{\mathbf{U}}$ based on $\widehat{\mathbf{u}}^c$;
        Get textual encodings $\mathbf{F}_{t_j}$;
        Forward the pre-trained stable diffusion model
          to obtain the output.
    end
    Compute regularizer $\mathcal{L}_{proto}$;
    Compute the loss $\mathcal{L}_{recon}$ of the pre-trained model;
    Update $\boldsymbol{w}^c, \boldsymbol{W}_s^c, \boldsymbol{W}_t^c$ according to
      $\mathcal{L}_{recon} + \lambda\mathcal{L}_{proto}$ with the learning rate $\eta$;
    Perform $e \in E$ epochs of local update;
    Compute local prototypes for $\boldsymbol{x} \in D$;
    return $\boldsymbol{w}^c, \{\zeta^c(j)\}_{j \in \mathcal{T}_{r+1}^c}$
```

Theorem 1. *Given an initial learning rate η_0 satisfying $\eta_0 E \leq 1/\beta_1$ and $\lambda \leq \frac{||\nabla\overline{\mathcal{L}}_{rE}||^2}{2\beta_2 G}$, the global model iterates in Algorithm 1 achieves*

$$\frac{1}{R}\sum_{r=0}^{R-1}||\nabla\overline{\mathcal{L}}_{rE}||^2 \leq \frac{2}{\psi_1(E)T}\left(\overline{\mathcal{L}}(\boldsymbol{w}_0) - \overline{\mathcal{L}}(\boldsymbol{w}_*)\right) + 2\lambda\beta_2 G + \frac{3\beta_1^2\eta_0^2\phi(E)\left(\sigma_l^2 + \sigma_g^2 + G\right) + \beta_1\psi_2(E)}{\psi_1(E)} \tag{1}$$

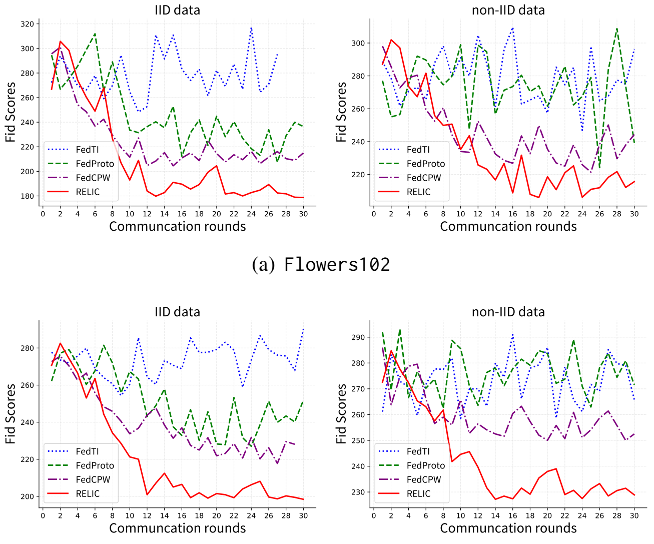
where $\psi_1(E) = \sum_{e=1/2}^{E-1}\eta_e$, $\psi_2(E) = \sum_{e=1/2}^{E-1}\eta_e^2$, and $\phi(E) = \sum_{e=1/2}^{E-1}e\eta_e$ with $\eta_{1/2} = \eta_0$. $\mathcal{L}^c$ is β_1-Lipschitz smooth and f_e is β_2-Lipschitz continuous. σ_l, σ_g, and G are used to bound Local and Global Variance and stochastic Gradient, respectively.

V. Evaluation

A. Setup

Following the data simulation described in subsection III-B, we perform experiments on Flowers102 [7] and Celeba [8] and set concentration of the Dirichlet distribution to 0.3 for the non-i.i.d. data case.

Baselines. The FedTI framework is directly used as the baseline approach. Another baseline, TI-Central [2] trains the pseudo-word centrally on three source images. As an ablation study, we evaluate two degraded versions, including FedCPW and FedProto, of Relic. FedCPW removes the prototype alignment, while FedProto can be viewed as applying the state-of-the-art federated prototype learning approach, FedProto [6], in the context of FedTI.



(a) Flowers102

(b) Celeba

Fig. 2: Illustrations of FID scores over communication rounds for the "clematis" flower and ID2820 face datasets under both i.i.d. (IID) and non-i.i.d. (non-IID) data scenarios.

Learning details. The pre-trained Stable Diffusion 2, at 768×768 resolution, is utilized for text-to-image generation. In the C-PW module, with a context length of 77, the learnable token vectors are of size 1×512. The shift and scale branches are simply implemented as two-layer linear networks with 32 hidden dimensions. We conduct all experiments utilizing 50 clients, with 5 clients selected per round to perform 2 epochs of local update. Local training for the C-PW relies on the AdamW optimizer, with a learning rate 0.0005 and a batch size of 1 for 30 communication rounds. We set $\lambda = 0.7$. Apart from presenting the qualitative results following the textual inversion [2], we provide Fréchet Inception Distance (FID), which measures the distance between two sets of images, as quantitative comparisons.

B. Effectiveness

With respect to the effectiveness, we show changes in FID scores over communication rounds of four approaches in both i.i.d. and non-i.i.d. data settings. In each round, FID is computed to measure the similarity between 50 real and fake images generated by the aggregated global model with the prompt template "a photo of a".

As depicted in Fig. 2, Relic effectively trains the pseudo-word to capture the unique concept of the original object, thereby enhancing the fidelity of the generated fake image. Specifically, as compared to FedTI, FedProto, and FedCPW, in both i.i.d. and non-i.i.d. of the Flowers102 and Celeba datasets, the FID scores of Relic steadily decrease and remain relatively stable after 16 communication rounds. We argue that the effectiveness of Relic is attributed to the seamless cooperation of C-PW and prototype alignment. C-PW learns common concepts specific to the object by training a global pseudo-word but leaving unique local characteristics in the

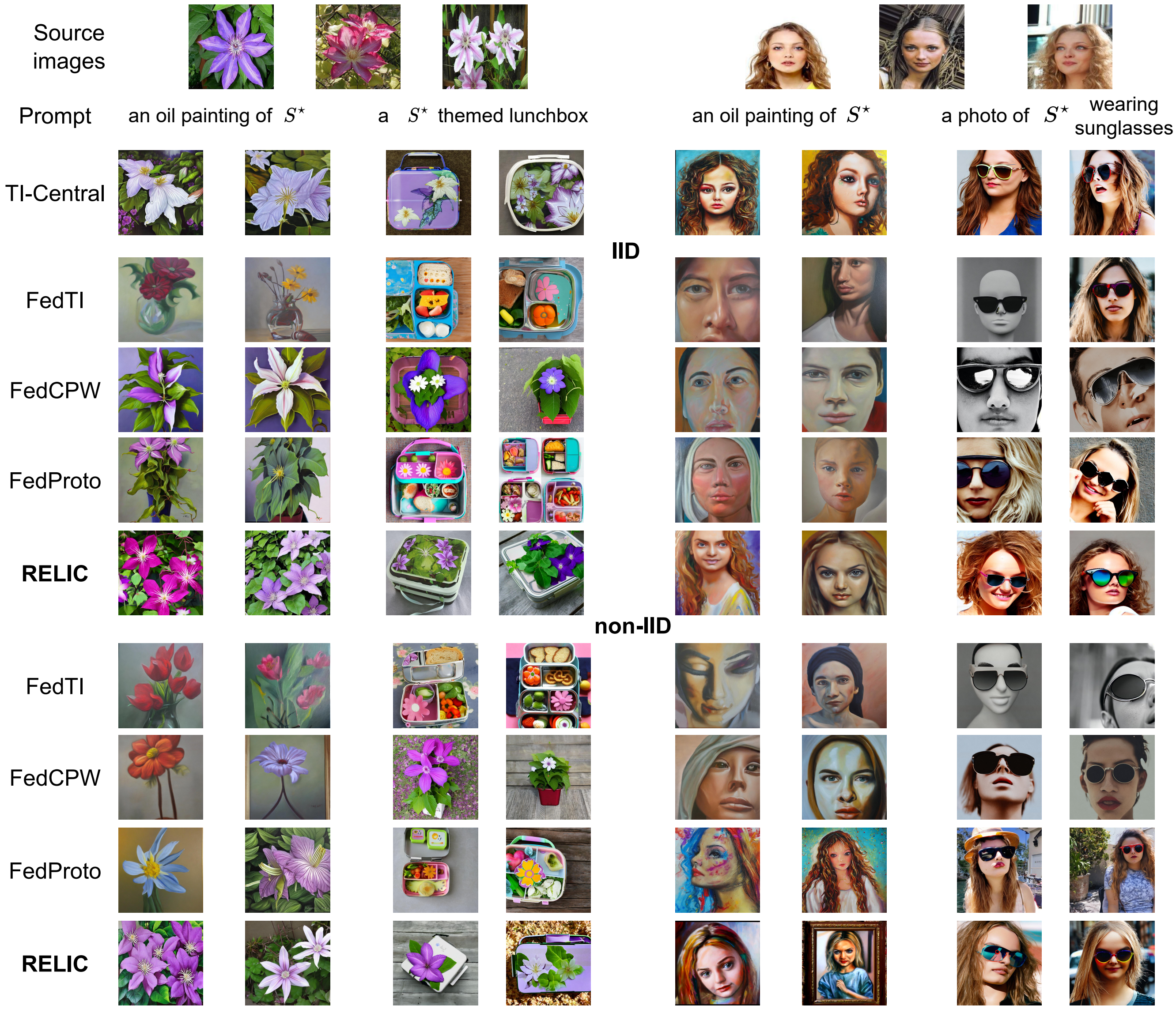

Source images
Prompt
an oil painting of $S^\star$
a $S^\star$ themed lunchbox
an oil painting of $S^\star$
a photo of $S^\star$ wearing sunglasses
TI-Central
IID
FedTI
FedCPW
FedProto
RELIC
non-IID
FedTI
FedCPW
FedProto
RELIC

Fig. 3: Object variations generated using the pseudo-word learned by RELIC and other baseline methods under federated learning with i.i.d. (IID) and non-i.i.d. (non-IID) data.

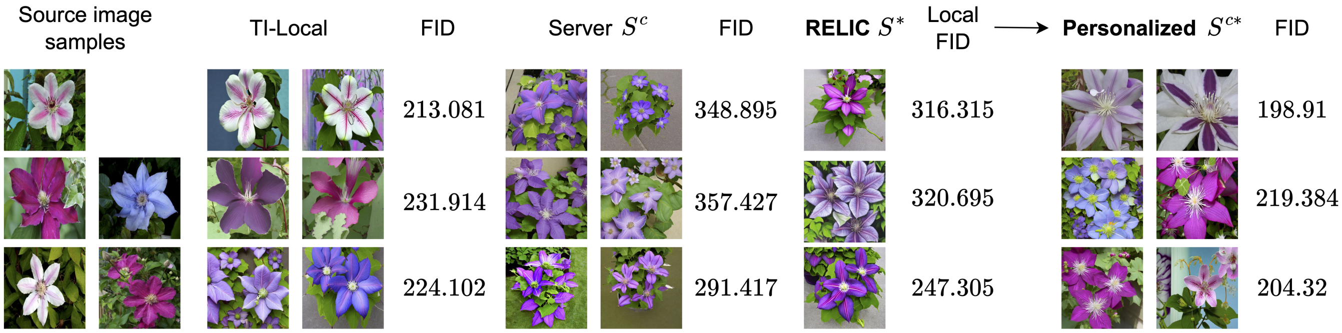


Fig. 4: Illustration of images generated by each client using the pseudo-words learned by TI-Local and RELIC under federated learning with non-i.i.d. data. Three rows correspond to clients with numbers of local samples: 1, 3, and 6, in which the 6 images present various styles. They have 8, 12, and 18 fake images, respectively.

personalized branch. As pseudo-words from clients contain

generic knowledge, they can be averaged without comparing the learned concepts. Besides, prototype alignment further protects the learned concepts by the globally consistent prototypes.

C. Overall Performance

As depicted in Fig. 3, irrespective of whether the data is i.i.d. or non-i.i.d., RELIC exhibits the ability to learn the pseudo-word effectively, resulting in the generation of top-quality images. Compared with other baselines, RELIC generates an object's variations, which 1). maintain core characteristics of the original subject under different prompts and conditions, and 2). are typically more faithful to the original subject. More importantly, RELIC effectively utilizes samples distributed among clients, achieving better performance than TI-Central.

Specifically, for both ID2820 face and "clematis" flower, Fig. 3 shows that RELIC effectively captures concepts from samples of participating clients. This results in generated images that present all aspects of the source images of the object, leading to more diverse and faithful variants. Meanwhile, the images generated by TI-Central lack diversity and details and exhibit visual homogenization, while FedTI fails to capture specific concepts from source images. The ablation study, especially results under the non-i.i.d. data case, shows that C-PW facilitates capturing more details while prototype alignment focuses on maintaining the learned concepts.

D. Personalized Images and Privacy Guarantees

We allow clients to download the global pseudo-word $S^\star$ trained by RELIC and generate personalized images using the "a photo of a" as the prompt template, as shown by Fig. 4.

RELIC captures the most prominent features of an object without any sacrifice in personalization. On the one hand, every client can utilize the $S^\star$ to generate images that prominently exhibit the features of an object, as depicted in the RELIC $S^\star$ column of Fig. 4. On the other hand, the scale-shift structure of the C-PW module enables each client to retain local characteristics in its local scale and shift branches. Thus, adding local shift and scale values to $S^\star$ gets personalized $S^{c\star}$, thus generating personalized images shown in the last column of Fig. 4. Further evidence is presented through FID scores computed between personalized and local images.

When comparing TI-Local, Server S^c, and RELIC $S^\star$ columns of Fig. 4, we show that RELIC protects the **client privacy** as unique concepts of local images are not exposed to the server. The personalized scale-shift branch and the prototype alignment mechanism allow clients to retain their unique concepts locally but only share common features. As demonstrated by the third row in Fig. 4 for both datasets, the generated images cannot be identified as belonging to which specific client but predominantly showcase the prominent features of the "clematis" flower.

VI. CONCLUDING REMARKS

In this paper, we present *federated textual inversion*, a framework that learns a pseudo-word within the federated learning setting for fine-tuning stable diffusion models. Our empirical experiments show that directly sharing pseudo-word updates exposes clients' local images, and averaging these updates for global aggregation makes the resulting pseudo-word ineffective. To address these issues, we proposed RELIC, a novel framework containing a conditional pseudo-word structure and prototype alignment. With privacy guarantees, it allows clients to collaboratively learn an effective global pseudo-word while generating personalized pseudo-words tailored to their individual data. We theoretically establish the convergence rate for RELIC. Our wide variety of experiments have provided compelling evidence that RELIC guarantees that overall and client-specific images can be generated with high quality and privacy preservation.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[2] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017, pp. 1273–1282.

[4] F. Wang, E. Hugh, and B. Li, "More than Enough is Too Much: Adaptive Defenses against Gradient Leakage in Production Federated Learning," in *Proc. IEEE INFOCOM*. IEEE, 2023.

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," *Advances in neural information processing systems (NeurIPS)*, vol. 30, 2017.

[6] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated Prototype Learning across Heterogeneous Clients," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, vol. 36, no. 8, 2022, pp. 8432–8440.

[7] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over A Large Number of Classes," in *Proc. 2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.

[8] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. International Conference on Computer Vision (ICCV)*, December 2015.

[9] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional Prompt Learning for Vision-Language Models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 816–16 825.

[11] B. Y. Lin, C. He, Z. Ze, H. Wang, Y. Hua, C. Dupuy, R. Gupta, M. Soltanolkotabi, X. Ren, and S. Avestimehr, "FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks," in *Proc. North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics (ACL), 2022, pp. 157–175.

[12] H. Zhao, W. Du, F. Li, P. Li, and G. Liu, "FedPrompt: Communication-Efficient and Privacy-Preserving Prompt Tuning in Federated Learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] T. Guo, S. Guo, and J. Wang, "pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning," in *Proc. ACM Web Conference 2023 (WWW)*, 2023, pp. 1364–1374.

[14] W. Lu, X. Hu, J. Wang, and X. Xie, "FedCLIP: Fast Generalization and Personalization for CLIP in Federated Learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[15] W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Rethinking Federated Learning With Domain Shift: A Prototype View," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 312–16 322.