FLAIR: Federated Learning for Augmented Industrial Retrieval

Diletta Chiaro, Pian Qi, Valeria Mele, and Francesco Piccialli

***Abstract*—Deep learning (DL) has significantly advanced Industry 4.0 by leveraging data from the Industrial Internet of Things (IIoT) to enable smart manufacturing, predictive maintenance, and data-driven product marketing. However, multimodal industrial data presents challenges for traditional frameworks, including scalability, data privacy, and integration efficiency. This article introduces an efficient product retrieval framework for e-commerce systems, addressing privacy and performance challenges through federated learning (FL). Specifically, we propose FL for augmented industrial retrieval (FLAIR), a novel part retrieval system where distributed warehouses collaboratively train a multimodal foundation model, contrastive language-image pretraining (CLIP), by fine-tuning only the Adapter module via FL, ensuring data privacy and efficiency. To address the limited availability of multimodal industrial data, our framework incorporates effective data augmentation strategies to enhance the diversity and quality of the training dataset. Comprehensive experiments on the industrial language-image dataset (ILID) highlight that FLAIR holds effective privacy safeguards and strong retrieval capabilities. Additionally, an advanced e-commerce recommendation system built on FLAIR showcases its practical effectiveness. FLAIR represents the first application of FL for industrial product retrieval, optimizing part searches, inventory management, and customer experience while maintaining data security. The complete code is available at https://github.com/MODAL-UNINA/FLAIR.**

***Index Terms*—Contrastive language-image pretraining (CLIP) model, federated learning (FL), multimodal foundation model, retrieval system.**

I. Introduction

Over recent decades, deep learning (DL) has achieved remarkable success across various domains, revolutionizing numerous industries. Industry 4.0 serves as a prime example of this transformation. In this era, vast amounts of data are collected from various sources through the Industrial Internet of Things (IIoT), which is then processed using DL can power applications, like smart manufacturing, fault detection, and product marketing. These advancements have significantly enhanced production efficiency and informed enterprise decision-making.

However, single-modal DL frameworks often fall short when addressing complex industrial data, which are typically multimodal and information-rich, such as vibration signals from machines, product images, text records from warehouses, etc. To address this challenge, the emergence of large-scale pretrained foundational models has marked a great advancement in DL. For instance, Li et al. [1] integrated foundational models into urban property management systems. By combining large-scale vision-language models (VLMs), they offer innovative solutions to challenges in corporate operations, including energy infrastructure security, residential community management, and urban operational oversight. The application of such large-scale pretrained models in industrial contexts, though not entirely new, has yielded profound benefits [2]. Models such as InstructGPT [3], Gorilla [4], GPT-4 [5], and Qwen2 [6] exemplify this progress, driving considerable advancements across the sector.

Building on the success of Large Language Models, visual foundation models (VFMs) have garnered significant attention in recent years. These models extend beyond language to incorporate visual modalities, enabling them to process and integrate knowledge from both text and images. Multimodal foundation models, such as contrastive language-image pretraining (CLIP) [7], OFA [8], BLIP-2 [9], and BEIT-3 [10], enrich their knowledge systems by learning from diverse modalities through self-supervised or unsupervised learning approaches. For vision-language tasks, these models leverage text and image encoders to learn from batches of positive and negative sample pairs. This contrastive learning approach effectively aligns visual and textual information, creating robust joint representations. VFMs exhibit robust knowledge representation capabilities and superior generalization compared to traditional models. With effective fine-tuning (e.g., Prompts [11] and Adapters [12]), they adapt well to various downstream tasks, excelling in applications such as object detection, semantic segmentation, and cross-modal retrieval [13].

Another critical challenge is the private and inaccessible nature of industrial datasets, as major companies treat them as valuable assets and prioritize data security within industrial systems. In this context, federated learning (FL) has emerged as a transformative distributed machine learning paradigm [14], [15]. At its core, FL enables multiple participants to collaborate on model training without sharing their raw data. Each participant trains a model locally on

Received 13 December 2024; revised 26 February 2025 and 23 April 2025; accepted 30 April 2025. Date of publication 8 May 2025; date of current version 25 September 2025. This work was supported in part by the Infrastructure for Big Data and Scientific Computing Project (IBiSco) through Call 424-2018—Action II.1 and the HPC Cluster under Grant PON R&I 2014-2020, and in part by PNRR Project FAIR—Future AI Research, Spoke 3, through the NRRP MUR Program funded by the NextGenerationEU under Grant PE00000013. *(Corresponding author: Francesco Piccialli.)*

The authors are with the Department of Mathematics and Applications R. Caccioppoli, Mathematical Modelling and Data AnaLysis Research Group (M.O.D.A.L.), University of Naples Federico II, 80138 Naples, Italy (e-mail: diletta.chiaro@unina.it; pian.qi@unina.it; valeria.mele@unina.it; francesco.piccialli@unina.it).

Digital Object Identifier 10.1109/JIOT.2025.3567194

their private dataset and exchanges only the resulting model updates [16]. These updates are then aggregated to form a global model with robust knowledge representation. This approach ensures that sensitive data remains secure while achieving effective collaborative training [17], [18].

Industrial applications face challenges in handling complex multimodal data, particularly when privacy and data scarcity are concerns [19]. While pretrained models like CLIP show promise, their use in industrial settings, especially with FL, remains underexplored. This research aims to fill this gap by proposing FL for augmented industrial retrieval (FLAIR), a framework that combines FL and CLIP for industrial product retrieval. We address key challenges in multimodal data integration, privacy protection, and limited datasets, ultimately enhancing e-commerce systems' performance while ensuring privacy. This article introduces FLAIR, a novel industrial product retrieval framework that utilizes FL to address privacy and performance challenges in distributed e-commerce platforms. By fine-tuning only the Adapter module of CLIP through FL, our system avoids the need to exchange base model parameters, thus preserving privacy. To counter the scarcity of industrial data, we propose a data augmentation strategy to enrich the dataset and improve federated training effectiveness. Extensive experiments on the industrial language-image dataset (ILID) validate FLAIR's robust privacy protection and superior image–text retrieval performance. We further extend FLAIR's capabilities by incorporating an advanced e-commerce recommendation system for efficient retrieval and recommendations from distributed warehouses, optimizing inventory management and enhancing customer experience.

To the best of our knowledge, this is the first work to apply FL to design a retrieval system tailored for industrial products. Our framework not only enables intuitive and efficient part searches but also optimizes inventory management and enhances the customer experience. Our contributions are summarized as follows.

1) We introduce FLAIR, an advanced e-commerce retrieval system designed to optimize the retrieval of industrial multimodal data.
2) We present a data enhancement strategy tailored for industrial datasets, addressing the challenge of domain data scarcity.
3) We leverage parameter-efficient fine-tuning (PEFT) within the FL framework for cooperative training, to strike an optimal balance between privacy protection and operational efficiency.

II. Related Works

A. Multimodal Foundation Models for IIoT

Although multimodal foundation models perform well in daily environments, their adoption in IIoT remains limited, with CNN-based neural networks still dominant [20]. This is due to the diversity of data and the dynamic nature of industrial environments, which challenge the deployment of large foundation models [21]. Moreover, IIoT devices often have limited capacity, computing power, and memory, hindering the use of resource-intensive multimodal models [22]. To address these challenges and adapt large models to industrial needs while enhancing reasoning and decision-making, research efforts in this area are emerging [23]. Wang et al. [24] proposed an industrial foundation model leveraging the Metaverse to manage resources and provide integrated services. The framework consists of three core models—visual, language, and operational—supporting decision-making at different organizational levels. It is illustrated through three scenarios: 1) clothing manufacturing; 2) oil extraction; and 3) optical inspection. Although still conceptual, it demonstrates the potential of large-scale models in industry. Picard et al. [25] explored multimodal VLMs in engineering design tasks such as sketch similarity analysis, topology optimization, and manufacturability assessment. They conducted extensive experiments with over 1000 input queries to evaluate GPT-4V and LLaVA 1.6 34B in these applications. Gong et al. [26] proposed a cross-modal zero-shot attribute value generation framework, ViOC-AG, which leverages the CLIP model to extract product attribute values, removing the need for buyers to manually enter product descriptions. Hua et al. [27] also leveraged CLIP, introducing a hierarchical alignment and cropping framework called HieClip to align industrial images and text. However, its primary focus is on anomaly detection rather than contrastive retrieval. While the study provides valuable insights into industrial applications, it primarily relies on existing models, limiting their effectiveness for specialized tasks.While multimodal foundation models show potential, their application in IIoT is limited due to challenges like data diversity, resource constraints, and lack of task-specific adaptation. Existing research often relies on pre-existing models or conceptual frameworks without fully addressing the unique needs of industrial settings. To overcome these limitations, related research should focus on fine-tuning models, optimizing resource use, and tailoring them to specific industrial tasks, which would enhance their practical applicability in IIoT systems.

B. Federated Learning for Retrieval System

Recent research work [28], [29] have explored the application of FL to cross-modal retrieval, where data is distributed among multiple parties, with each party retaining the privacy and inaccessibility of its local data. Cross-modal retrieval aims to construct a shared subspace for diverse data modalities, enabling direct similarity measurements by projecting the data into this common representation. For example, Li et al. [30] proposed FedSRMR, a federated supervised cross-modal retrieval method. This approach leverages adaptive aggregation of linear layer parameters from multiple clients, enabling dynamic global aggregation and the learning of a shared subspace. Despite its contributions, the method has some limitations that only focuses exclusively on a scenario with data partitioned across three clients, offering no insights into data partitioning strategies in FL and their effects on model performance. Especially, the heterogeneous nature of different client data can critically influence model outcomes. To tackle the challenge of data heterogeneity in FL, Feng et al. [31] introduced a method called FedPAM. This approach identifies the top k most similar text-image pairs from a private database,

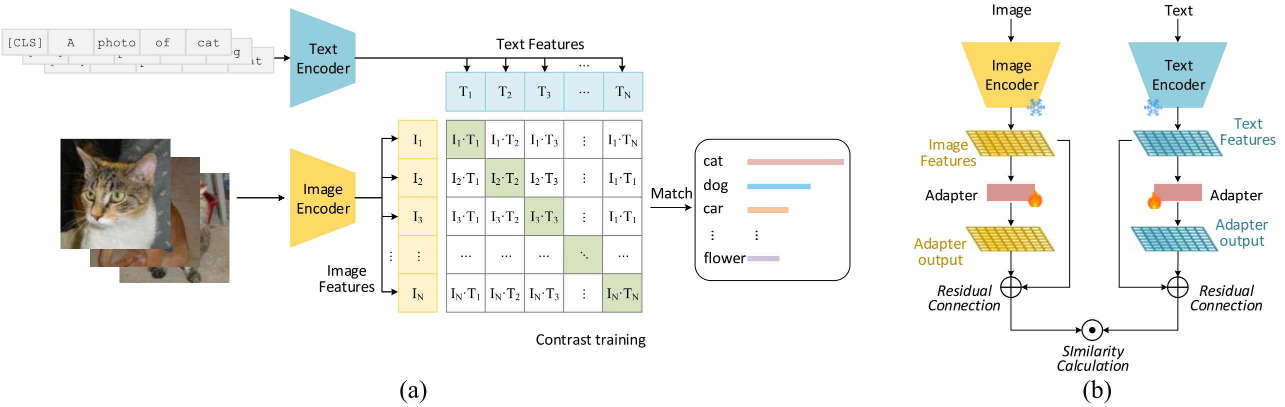

[CLS]
A
photo
of
cat
Text Encoder
Text Features
T_1 T_2 T_3 ⋯ T_N
Image Encoder
Image Features
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ $I_1 \cdot T_1$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ $I_1 \cdot T_1$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ $I_N \cdot T_N$
Match
cat
dog
car
flower
Contrast training
(a)
Image
Text
Image Encoder
Text Encoder
Image Features
Text Features
Adapter
Adapter
Adapter output
Adapter output
Residual Connection
Residual Connection
Similarity Calculation
(b)

Fig. 1. (a) Original CLIP, joint training of image and text encoders to align paired image–text inputs by maximizing similarity within each training batch. (b) CLIP-Adapter, adapter modules are inserted at the output of the frozen image and text encoders. Only the Adapter layers are trained, while a residual connection preserves part of the original encoder features.

enabling the generation of personalized representations tailored to individual clients. While this method offers valuable insights for advancing image–text retrieval tasks in FL, it relies on the training dataset of clients as the source for early retrieval. This reliance raises potential concerns about data privacy that warrant careful consideration. Although recent research has advanced FL-based cross-modal retrieval, challenges persist. Many existing methods are limited to specific client scenarios and overlook the impact of data partitioning strategies and client data heterogeneity, which can greatly influence model performance. To address this, our experimental section explores various nonindependent and identically distributed (non-IID) scenarios, evaluating their effects on retrieval performance and model adaptation.

III. Methodology

A. Preliminaries

1) Contrastive Language-Image Pretraining: CLIP is a powerful DL model designed for image–text contrastive learning. Trained on a large dataset of image–text pairs, CLIP learns to associate visual and textual information by mapping both modalities into a shared representation space. This enables the model to evaluate the degree of similarity between images and text, facilitating an understanding of multimodal relationships. OpenAI offers several pretrained versions of CLIP, including ViT-B/16, ViT-B/32, and ViT-L/14. Those models are trained on 400 million image–text pairs collected from the Web. CLIP typically consists of two components: 1) an image encoder and 2) a text encoder. The text encoder processes embedded text, which is often expanded during pretraining to provide richer semantic context. For example, image labels are formatted as sentences like "a photo of {}," where {} represents the label. This predefined structure is known as the hard prompt template H. The text encoder extracts feature vectors from these sentences, while the image encoder converts the corresponding images into visual feature vectors. During training, contrastive learning is used to align the textual and visual features by maximizing their cosine similarity, ensuring the model learns meaningful multimodal representations. Fig. 1(a) shows the architecture of CLIP.

2) Adapter: CLIP-Adapter was introduced by Gao et al. [32] in prior research, it is an innovative method aimed at achieving PEFT for CLIP models. Unlike traditional fine-tuning approaches that often require extensive changes to the model architecture, the CLIP-Adapter adds only a few learnable bottleneck linear layers to the language and image encoder branches of CLIP. This design allows the original CLIP backbone to remain unchanged, preserving its pretrained knowledge while enabling task-specific adaptation.

Formally, given the input image I and a set of text categories T, the image features f and text features g in the CLIP backbone are computed as follows:

$$\begin{aligned} f &= \text{Visual-Encoder}(I) \\ g &= \text{Text-Encoder}(\text{Tokenizer}([H; T])). \end{aligned}$$

Then, two learnable feature adapters, $A_v(\cdot)$ and $A_t(\cdot)$, are introduced. Each consists of two linear layers that refine f and g, respectively

$$\begin{aligned} A_v(f) &= \text{ReLU}\left(f^T w_{v1}\right) w_{v2} \\ A_t(g) &= \text{ReLU}\left(g^T w_{t1}\right) w_{t2} \end{aligned}$$

where w_{v1}, w_{v2} and w_{t1}, w_{t2} are the bottleneck linear layer weights for the visual and text branches, respectively. ReLU is the activation function.

To preserve the original knowledge encoded in CLIP, the feature adapters incorporate residual connections. The constants α and β, known as residual ratios, control the balance between retaining the original representations and integrating the newly learned adaptations

$$\begin{aligned} f &= \alpha A_v(f)^T + (1-\alpha) f \\ g &= \beta A_t(g)^T + (1-\beta) g. \end{aligned}$$

During training, the parameters of $A_v(\cdot)$ and $A_t(\cdot)$ are optimized using the contrastive loss from the original CLIP model

$$\mathcal{L}_\theta = -\frac{1}{N} \sum_i^N \log \frac{\exp\left(g_i^\top f_i / \tau\right)}{\sum_{j=1}^N \exp\left(g_j^\top f_i / \tau\right)}$$

where N is the total number of training examples.

The architecture of this approach is depicted in Fig. 1(b). By adopting the CLIP-Adapter in this work, we aim to leverage its efficient fine-tuning methodology, enhancing our model's performance without the computational overhead associated with retraining large portions of the CLIP architecture.

B. FLAIR

In this section, we present the proposed framework FLAIR, which encompasses three key components.

1) Enhanced Industrial Image–Text Dataset: As [33] highlights, the industrial domain presents unique challenges involving uncommon objects and scenes, rendering commonly used datasets inadequate for fine-tuning or transfer learning. Instead, datasets tailored to industrial scenarios are essential for effective training. However, the scarcity of publicly available ILIDs significantly limits the application of large multimodal foundation models in this field. To address this limitation, we propose a dataset expansion method based on data augmentation. This approach involves generating new images through techniques, like color change, rotation, and cropping, while keeping the corresponding text descriptions. Consequently, each image–text pair generates multiple variants, leading to a substantial expansion of the industrial dataset. The data augmentation techniques we apply include adjusting color (*ColorJitter*), flipping horizontally (*RandomHorizontalFlip*), rotating (*RandomRotation*), resizing and cropping (*RandomResizeCropping*), and converting to grayscale (*Grayscale*). It is important to note that this operation is applied exclusively to the training set during experiments, adhering strictly to the principle of maintaining an unchanged test set for DL training.

2) Federated PEFT for CLIP: FL enables collaborative model training across decentralized clients while preserving data privacy. Unlike centralized approaches, FL ensures that raw data remains local to each client, making it well-suited for privacy-sensitive applications.

In our proposed Federated PEFT framework, named FLAIR, we adopt a hybrid model structure where the base model parameters θ are frozen and shared across clients. Each client i maintains its own adapter parameters $\Delta\theta_i$, which are lightweight and task-specific. During local training, only $\Delta\theta_i$ is updated, while θ remains unchanged. This design minimizes communication costs and supports efficient fine-tuning.

In this setting, the e-commerce parent company acts as the central server, coordinating training across multiple distributed warehouses, each representing a client. Warehouses hold private data such as product images, descriptions, and part metadata, which is never shared externally. Instead of transmitting the full model, only the adapter parameters $\Delta\theta_i$ are sent to the server after local training.

The central server aggregates the received adapter updates from all k clients via parameter-wise averaging

$$\Delta\theta = \frac{1}{k}\sum_{i=1}^{k}\Delta\theta_i.$$

This approach integrates knowledge from diverse sources while maintaining a compact model. The updated global adapter $\Delta\theta$ is then redistributed to all clients, enabling continual refinement. The training workflow is illustrated in Fig. 2.

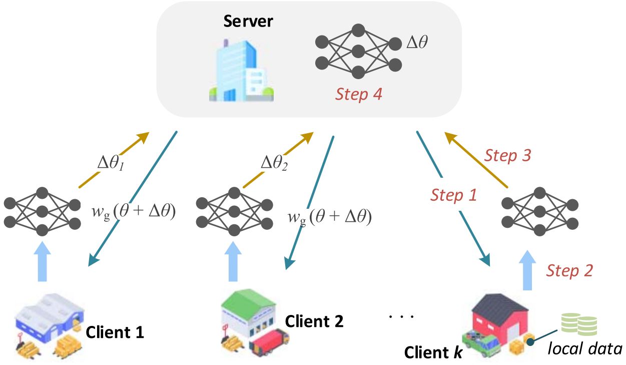


Fig. 2. Training process of the proposed FLAIR framework. Steps: (1) Each client downloads the latest global model $w_g = \theta + \Delta\theta$ from the server. (2) The client performs local training on its dataset, during which only the adapter modules $\Delta\theta_i$ are updated while the base parameters θ remain frozen. (3) The trained adapters $\Delta\theta_i$ are uploaded to the server. (4) The server aggregates all received adapters $(\Delta\theta_1, \Delta\theta_2, \dots, \Delta\theta_k)$ to update the global adapter parameters.

3) Advanced E-Commerce Recommendation System: The global model, trained using the FL framework, is integrated into the parent company's e-commerce platform to enable seamless and intuitive industrial parts search, accessible to users without technical expertise. After synchronizing the final global model, each warehouse client encodes local image data into feature vectors using a CLIP encoder, creating a data retrieval library. The server-side system processes users' natural language queries, encodes them via a text encoder, and matches them against the retrieval libraries of participating warehouses to identify the best-matching parts. For example, a user entering steel gear for machinery can retrieve relevant images leveraging the collective knowledge of all warehouses. The platform then displays the availability of parts across different locations and recommends the warehouse that optimizes delivery time and cost. Users can also perform image-to-text retrieval by uploading photos of parts, assisting those who may not know the name or specifications.

Additionally, the federated model, trained on diverse datasets from multiple warehouses, enhances user experience through intelligent product suggestions. Customers searching for specific items, such as bolts, may receive recommendations for complementary or substitute products like washers, nuts, or alternative bolt versions. These recommendations are determined by cosine similarity scores, where higher scores indicate greater relevance. By incorporating insights from all warehouses, the system not only improves search success rates but also offers valuable recommendations that align with user needs. This approach ensures a comprehensive and efficient search experience, bridging the gap between technical complexity and user accessibility.

IV. Experimental Results

A. Dataset

The ILID is the first multimodal unstructured dataset specifically designed for the industrial domain. It comprises

TABLE I
NON-IID SCENARIO DATA PARTITIONING

Scenario ID	Partitioning Type	Characteristics	Client 0	Client 1	Client 2	Client 3	Client 4
S1	Source	Quantity skew, different styles	1004	3980	3222	1422	187
S2	Dir(0.1)	Extreme feature skew	1003	654	3219	3087	1852
S3	Dir(0.5)	Moderate feature skew	1132	1834	2462	2248	2138
S4	Pareto	Long-tail, quantity skew	7852	491	491	490	490

images of industrial components paired with detailed textual annotations. The dataset was curated using Web crawling tools, sourcing data from five distinct online stores. ILID captures a wide spectrum of industrial parts, from small items like washers, hinges, and bearings to larger equipment such as lifts and trailers. In total, the ILID dataset includes 12 270 samples, with each sample featuring an image accompanied by five types of textual data: 1) a short label; 2) a long label; 3) a detailed description; 4) material information; and 5) surface material specifications. This comprehensive information makes ILID a valuable resource for developing and testing multimodal applications in industrial contexts.

B. Experimental Setup

The ILID dataset comprises samples from different sources, which we divided among 5 clients. Each client holds a varying number of samples: To create a balanced evaluation, 20% of the data from each client was randomly selected for the test set. To thoroughly assess the performance of the proposed method across different non-IID scenarios, we design four distinct experimental settings. Table I outlines the characteristics of each scenario, along with the corresponding client data distribution. Specifically, **S1**, Data partitioned according to 5 distinct online shop sources: *norelem*, *rexroth*, *maedler*, *jungheinrich*, and *ganter*. **S2**, Dirichlet distribution with a concentration parameter of 0.1, i.e., Dir(0.1). **S3**, Dirichlet distribution with a concentration parameter of 0.5, i.e., Dir(0.5). **S4**, Pareto distribution where a small fraction of clients (e.g., 20%) holds the majority of the data (e.g., 80%).

The study uses the pretrained CLIP ViT-B/16 model from OpenAI.[1] Built on the vision transformer (ViT) architecture, it divides images into 16 × 16 patches and uses a Transformer to extract global features. The ViT-B/16 model has 12 Transformer layers with a hidden dimension of 768, an image encoder for 224 × 224 images, and a text encoder that tokenizes input using BERT. Both encoders output 512-D feature vectors. The image encoder has 86.2 M parameters, while the text encoder has 37.8 M. For federated PEFT training, an adapter is added to both encoders, with a bottleneck layer reduced by a factor of 4. This results in 131K trainable parameters per adapter, totaling 262K. This reduces the number of trainable parameters, saves computational resources, and lowers communication overhead. The AdamW optimizer is used for local fine-tuning, with a cosine annealing learning rate schedule from 10^{-4} to 10^{-5} over 5 local epochs. Gradient clipping (max norm of 1) and mixed precision training are also applied. Training uses a batch size of 64, testing samples 32, and 50 communication rounds between clients and the server. Additional, the applied data augmentation pipeline includes the following transformations: color jittering (brightness = 0.4, contrast = 0.4, saturation = 0.4, hue = 0.2); random rotation up to 30 degrees; horizontal flipping; random resized cropping to 214 pixels with a scale range between 0.8 and 1.0; and grayscale conversion applied with a probability of 30%.

In the experiment, to explore whether varying text inputs impact the results, we experimented with four different types of text prompts: 1) long label; 2) short label; 3) long label with description; and 4) short label with description. For the first two types, we used the prompt template "an industrial photo of a {}," replacing {} with either the long or short label. For the last two types, we used "an industrial photo of a {}, {}," where the first {} is filled with the label, and the second {} contains the corresponding description. For empirical comparison, we selected several baseline models.

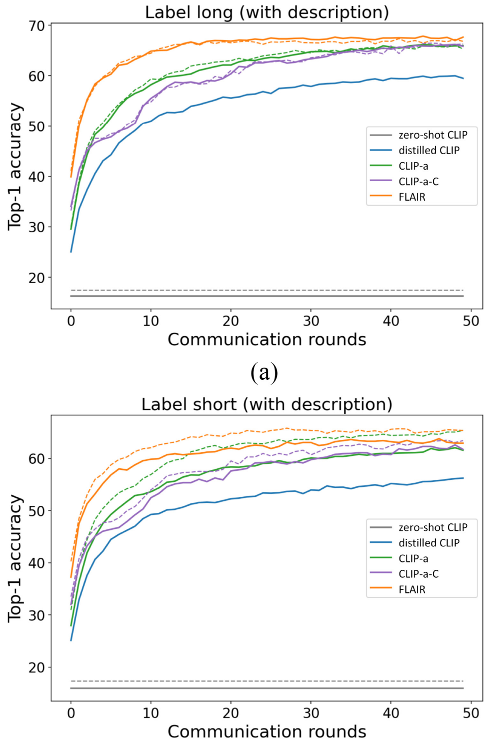
1) *Zero-shot CLIP*, A pretrained CLIP model directly used for evaluation without any fine-tuning.
2) *Distilled CLIP*, A lightweight version of CLIP distilled on the Conceptual Captions 3M (CC3M) and Conceptual 12M (CC12M) datasets, with ViT-B/16 serving as the teacher model, as proposed by [34]. This model includes 5.6M parameters for the image encoder and 21.3M for the text encoder.
3) *CLIP-a*, only the adapter modules of the CLIP model are fine-tuned, and other parameters are frozen.
4) *CLIP-a-C*, only the adapter modules of the CLIP model are fine-tuned, within a centralized training environment.

To evaluate the performance of the image–text retrieval system, we measured the key metrics: Top-1 Accuracy, which calculates the percentage of cases where the best match between the image and text features is correct.

C. Results

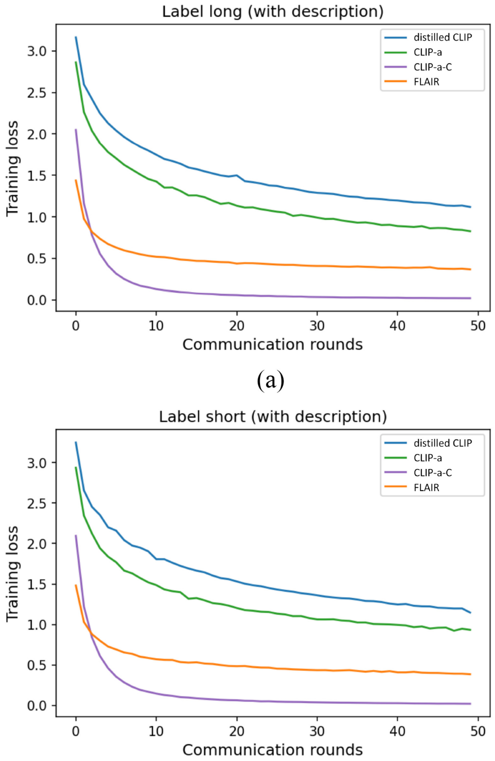
1) Image–Text Retrival: The experimental evaluation results of non-IID scenario S1 are presented in Fig. 3. In Fig. 3(a), the results are based on long labels as text input. The solid line represents the use of long labels alone, while the dotted line incorporates descriptions as supplementary information. From the top-1 accuracy curve, FLAIR consistently leads, converging the fastest within the same number of training epochs. It is followed by CLIP-a, which, notably, does not use any data augmentation strategy, unlike FLAIR. This highlights the role of the proposed data augmentation strategy in enhancing the diversity of the training dataset, which in turn accelerates convergence and improves the accuracy of the model. CLIP-a-C represents the model in a centralized training environment. Lacking local training adjustments, it does not achieve optimal accuracy within 50 epochs, underlining the importance of multiparty cooperative training based on FL. Then, the ranking is the distilled

[1] https://github.com/openai/CLIP



(b)

Fig. 3. Top-1 accuracy comparison of different methods. (a) Label long alone (solid) versus with description (dotted) as text input. (b) Label short alone (solid) versus with description (dotted) as text input.



(b)

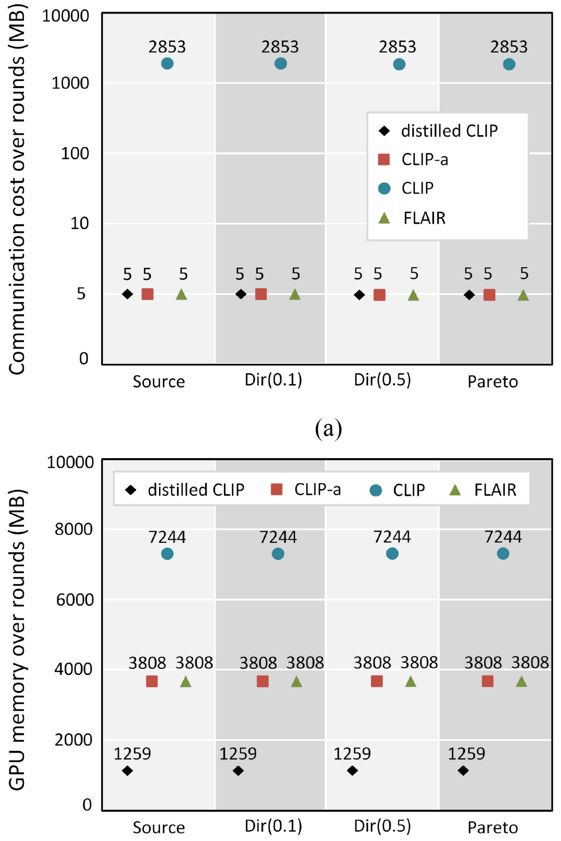
Fig. 4. Training loss comparison of different methods. (a) Label long with description as text input. (b) Label short with description as text input.

CLIP model. While it uses adapters for both image and text encoders, its performance is limited due to a smaller model capacity compared to standard CLIP (Vit-B/16). Although models trained via knowledge distillation are valuable for reducing inference load on devices, FLAIR provides a more balanced and effective solution in terms of comprehensive performance. It is important to note that zero-shot CLIP, while trained on the original pretrained knowledge, lacks fine-tuning on industrial datasets, making it less effective at correctly matching images and texts in industrial applications.

In Fig. 3(b), the experimental setup uses short labels as input. The solid line shows the results with only short labels, and the dotted line includes descriptions as supplementary information. The overall trends align with those in Fig. 3(a), but the accuracy is lower with short labels since they carry less information than long labels. Interestingly, when descriptions are used in conjunction with short labels, model performance improves, in contrast to the combination of long labels and descriptions. The latter does not provide additional benefits; instead, the longer text might introduce noise that hinders performance. On the other hand, the combination of short labels and descriptions enhances the accuracy. Additional, Fig. 4 illustrates the training loss of non-IID scenario S1. As all parameters of the zero-shot CLIP model are frozen, no training loss is observed. Among the other methods, CLIP-a-C converges more quickly due to its centralized training approach, while our FLAIR method follows closely behind. In contrast, the other two FL-based methods (distilled CLIP and CLIP-a) exhibit significantly higher losses.

Furthermore, comparative experiments were conducted to assess the proposed method against a scenario where all parameters of the CLIP model were fine-tuned, within the FL setting. When using long labels as text input, the top-1 accuracy of CLIP with full-parameter fine-tuning reached 75.37% at the 50th training epoch, outperforming our FLAIR, which achieved 67.63%. However, when short labels were used as text input, the training process for CLIP with full-parameter fine-tuning encountered a gradient explosion around the 35th epoch, causing the training loss to become *NaN*. This may be due to the deep architecture of image and text encoders of CLIP, where gradient accumulation during backpropagation leads to instability, especially when fine-tuning all parameters. while our FLAIR still runs normally and finally achieves an accuracy of 62.91%.

2) Communication Consumption: In FL, communication consumption plays a crucial role, referring to the cost of transmitting model updates between clients and the central server. This directly impacts the efficiency and scalability of the FL system. Fig. 5(a) shows the communication consumption per round for different methods under FL settings, measured by the amount of uploaded parameters. Notably, the three methods (distilled CLIP, CLIP-a, FLAIR) utilizing adapters for parameter fine-tuning exhibit significantly lower communication consumption than full parameter fine-tuning (CLIP), reducing it by 99.82%. Moreover, Fig. 5(b) presents the GPU memory usage of each method, where FLAIR ranks in the second tier, just above distilled CLIP, categorizing it as a low-consumption approach. Overall, Although in our FLAIR,



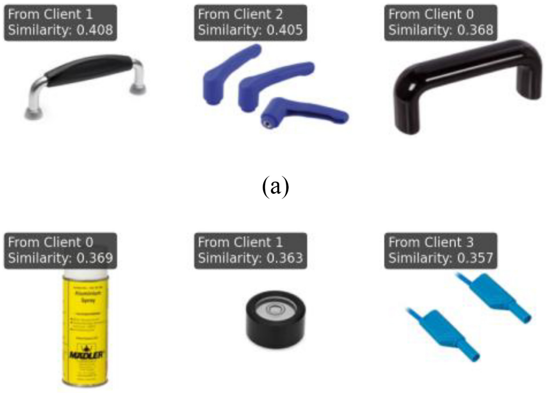
(a)

(b)

Fig. 5. (a) Communication costs. (b) GPU memory usage of different methods in four different non-IID scenarios.

PEFT introduces some performance degradation compared to full parameter fine-tuning, it significantly reduces communication costs. In FL, clients send the entire model to the server, which will lead to considerable bandwidth consumption and communication overhead. This issue is particularly severe in IoT environments, where general devices cannot handle such heavy data transmission. Overall, the compared results demonstrate that PEFT, applied within FL, offers a tradeoff between performance and efficiency, making it an ideal choice for scenarios with limited resources.

3) Recommendation System: Our FLAIR system is seamlessly integrated into the e-commerce platform, utilizing a global model trained by FL to facilitate industrial parts queries for users. Each warehouse client employs a CLIP encoder to generate image feature vectors. The server then processes the user's natural language query and matches it against the retrieval library of each warehouse using cosine similarity as the metric. This enables the system to provide intelligent product recommendations. Fig. 6 illustrates the results of this recommendation system when users input different queries. In Fig. 6(a), the user submits the query `u-handle`, and the system retrieves and displays product images from three different warehouse clients. Similarly, Fig. 6(b) shows the results for the query `floor cleaning`. The products retrieved by the recommendation system closely align with the user's query, demonstrating the system's effectiveness.



(a)

(b)

Fig. 6. Recommended products retrieved from different clients based on user queries. (a) For the query u-handle, similar handle-shaped items are returned according to scores. (b) For the query floor cleaning, semantically related items are retrieved from different clients.

In conclusion, the proposed advanced e-commerce recommendation system ensures both the privacy of warehouse data and the efficient allocation of products, striking a balance between speed and security. While our current evaluation focuses on quantitative retrieval performance, we recognize the importance of real-world user validation. In future work, we plan to conduct user studies with industry professionals to assess the practical usability, relevance, and effectiveness of the recommendation system. This validation will include user surveys, success rate analysis, and interaction studies, providing deeper insights into real-world performance and guiding further improvements.

V. Conclusion

This article introduces FLAIR, an innovative retrieval method for industrial products in e-commerce systems. FLAIR leverages the framework of FL to ensure privacy protection and enable seamless multiparty collaboration. The study presents a robust data augmentation strategy to enhance industrial parts data as image–text pairs, training the large foundation model CLIP through a federated, PEFT approach. Experimental results on the ILID industrial multimodal dataset validate the method's effectiveness and reliability for image–text retrieval. Furthermore, an advanced e-commerce recommendation system built using the trained model demonstrates potential for practical applications. We aspire for this study to serve as a reliable reference for industrial retrieval applications.

Despite its effectiveness, the proposed method has certain limitations. The data augmentation strategy, while beneficial, is not exhaustive, and variations in industrial scenarios may introduce cases that are not fully covered by the adopted augmentation techniques. Additionally, the CLIP-Adapter method used in this study has further room for optimization. One potential improvement is to more closely integrate the parameters of the visual and text adapters, co-training them rather than training them separately, which could lead to enhanced performance. Future work will focus on refining these aspects while also exploring the development of foundational models with extended capabilities, such as generative functionalities, to further advance this field and expand its practical impact.

Acknowledgment

The authors extend special thanks to Dr. Luisa Carracciuolo and Eng. Davide Bottalico for their constant and continuous support in utilizing the IBiSco HPC Cluster.

References

[1] Z. Li et al., "Federated learning in large model era: Vision-language model for smart city safety operation management," in *Proc. Compan. ACM Web Conf.*, 2024, pp. 1578–1585.

[2] L. Makatura et al., "How can large language models help humans in design and manufacturing?" 2023, *arXiv:2307.14377*.

[3] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. 36th Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744.

[4] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive APIs," 2023, *arXiv:2305.15334*.

[5] A. Josh et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[6] A. Yang et al., "Qwen2 technical report," 2024, *arXiv:2407.10671*.

[7] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[8] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23318–23340.

[9] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[10] W. Wang et al., "Image as a foreign language: BEIT pretraining for vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19175–19186.

[11] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021, *arXiv:2104.08691*.

[12] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[13] W. Wang et al., "VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–13.

[14] X. Zhou et al., "Hierarchical federated learning with social context clustering-based participant selection for Internet of Medical Things applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1742–1751, Aug. 2023.

[15] X. Zhou, W. Liang, A. Kawai, K. Fueda, J. She, and K. I.-K. Wang, "Adaptive segmentation enhanced asynchronous federated learning for sustainable intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6658–6666, Jul. 2024.

[16] F. Piccialli, D. Chiaro, P. Qi, V. Bellandi, and E. Damiani, "Federated and edge learning for large language models," *Inf. Fusion*, vol. 117, May 2025, Art. no. 102840.

[17] X. Zhou et al., "Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3191–3211, Oct. 2023.

[18] X. Zhou, W. Huang, W. Liang, Z. Yan, J. Ma et al., "Federated distillation and blockchain empowered secure knowledge sharing for Internet of Medical Things," *Inf. Sci.*, vol. 662, Mar. 2024, Art. no. 120217.

[19] X. Zhou et al., "Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 4, pp. 817–831, Apr. 2024.

[20] J. Chen et al., "Towards general industrial intelligence: A survey of continual large models in industrial IoT," 2024, *arXiv:2409.01207*.

[21] H. Zhang et al., "Large scale foundation models for intelligent manufacturing applications: A survey," 2023, *arXiv:2312.06718*.

[22] X. Zhou et al., "Reconstructed graph neural network with knowledge distillation for lightweight anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 11817–11828, Sep. 2024.

[23] K. Moenck et al., "Industrial segment anything–a case study in aircraft manufacturing, intralogistics, maintenance, repair, and overhaul," 2023, *arXiv:2307.12674*.

[24] J. Wang et al., "A framework and operational procedures for metaverses-based industrial foundation models," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2037–2046, Apr. 2023.

[25] C. Picard et al., "From concept to manufacturing: Evaluating vision-language models for engineering design," 2023, *arXiv:2311.12668*.

[26] J. Gong, M. Cheng, H. Shen, P.-Y. Vandenbussche, J. Jenq, and H. Eldardiry, "Visual zero-shot E-commerce product attribute value extraction," 2025, *arXiv:2502.15979*.

[27] L. Hua, X. Su, Y. Luo, S. You, and J. Long, "HieClip: Hierarchical CLIP with explicit alignment for zero-shot anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.

[28] G. Li, H. Lei, F. Liu, L. Li, and H. Jin, "AEPPFL: Accurate and efficient privacy protection federal learning in industrial IoT," *IEEE Internet Things J.*, early access, Apr. 17, 2025, doi: 10.1109/JIOT.2025.3562064.

[29] T. Zhang et al., "Generating synthetic data for unsupervised federated learning of cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 22569–22577.

[30] A. Li, Y. Li, and Y. Shao, "Federated learning for supervised cross-modal retrieval," *World Wide Web*, vol. 27, no. 4, p. 41, 2024.

[31] Y. Feng, F. Ma, W. Lin, C. Yao, J. Chen, and Y. Yang, "FedPAM: Federated Personalized augmentation model for text-to-image retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2024, pp. 1185–1189.

[32] P. Gao et al., "Clip-adapter: Better vision-language models with feature adapters," *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024.

[33] K. Moenck, D. T. Thieu, J. Koch, and T. Schüppstuhl, "Industrial language-image dataset (ILID): Adapting vision foundation models for industrial settings," 2024, *arXiv:2406.09637*.

[34] C. Yang et al., "CLIP-KD: An empirical study of CLIP model distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15952–15962.