Multimedia Systems (2024) 30:222
https://doi.org/10.1007/s00530-024-01422-9

SURVEY

A survey of multimodal federated learning: background, applications, and perspectives

Hao Pan[1] · Xiaoli Zhao[1] · Lipeng He[2] · Yicong Shi[1] · Xiaogang Lin[1]

Received: 16 March 2024 / Accepted: 15 July 2024 / Published online: 29 July 2024

Abstract
Multimodal Federated Learning (MMFL) is a novel machine learning technique that enhances the capabilities of traditional Federated Learning (FL) to support collaborative training of local models using data available in various modalities. With the generation and storage of a vast amount of multimodal data from the internet, sensors, and mobile devices, as well as the rapid iteration of artificial intelligence models, the demand for multimodal models is growing rapidly. While FL has been widely studied in the past few years, most of the existing research was based in unimodal settings. With the hope of inspiring more applications and research within the MMFL paradigm, we conduct a comprehensive review of the progress and challenges in various aspects of state-of-the-art MMFL. Specifically, we analyze the research motivation for MMFL, propose a new classification method of existing research, discuss the available datasets and application scenarios, and put forward perspectives on the opportunities and challenges faced by MMFL.

Keywords Federated learning · Multimodal learning · Multimodal federated learning · Machine learning

1 Introduction

Humans perceive the real world through various senses and process information with their brains. Similarly, computers receive data in multiple modalities through various sensors and input devices, and output predictions using Machine Learning (ML) models. Today, an unmatched amount of data is being generated and stored every single day on the internet, mobile devices, and sensors (cameras, microphones, gyroscopes, etc.) at an unprecedented rate. The modalities of these data vary significantly depending on the deployment platform of ML models.

1.1 Motivation of MMFL

On social media, image and text data complement each other. In most cases, the sentiments expressed in images and texts are matched, but sometimes it is also possible that the text may be positive while the image is negative [1]. Video and audio data are also a prevalent form of content on social media, containing versatile details and rich features [2]. Various internet companies and mobile phone manufacturers are hoping to use this data to enhance their user experiences, with some companies even exploring the possibilities of using artificial intelligence agents to provide tailored experience for individual users.

In the industrial sector, it is a common practice for factories to install numerous independent sensors in order to achieve assembly line automation. Each of these sensors generate a large volume of one-time industrial data every single day. Enterprises hope to use this data for intra- or

Communicated by Bing-kun Bao.

✉ Xiaoli Zhao
zhaoxiaoli@sues.edu.cn

Hao Pan
ph@sues.edu.cn

Lipeng He
lipeng.he@uwaterloo.ca

Yicong Shi
m320121104@sues.edu.cn

Xiaogang Lin
lxg@sues.edu.cn

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, China

[2] Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

inter-enterprise collaboration without investing more in the cost required for data collection.

In the medical field, the existence of modality gaps [3] (i.e., the inability to process them together due to differences in data modality) [4] means that patient examination images, electrocardiograms, and symptom information cannot be fully utilized by healthcare professionals to provide patient care. For that reason, the integration of healthcare and artificial intelligence has been one of the most popular research direction in recent years.

Multimodal learning (MML), as an effective solution for the aforementioned problems enabled the ability for ML models to process and correlate information from various modalities. Recently, MML has been gaining increasing popularity in the field of machine learning, opening up a variety of application opportunities such as cross-modal retrieval, image captioning, video summarization, emotion recognition, and medical diagnosis assistance [5–9].

However, the aggregation and utilization of massive amounts of data, including private records on personal mobile phones, confidential corporate data in industrial equipment, and sensitive medical records of patients in hospitals, inevitably draws the public's concern regarding data privacy and security. More and more countries and regions are implementing policies for data governance, such as China's "Data Security Law" and "Personal Information Protection Law", the European Union's "General Data Protection Regulation", and the United States' "Data Privacy and Protection Act". This presents a significant challenge for ML service providers.

Federated learning (FL), as a method that allows multiple clients to collaboratively train or optimize models without sharing their local raw data, has become the best alternative to traditional distributed machine learning [10]. A significant amount of existing research demonstrated that FL is highly effective in addressing data privacy and security issues in distributed environments; and FL has been widely applied to image or text datasets [11–15]. But that is far from sufficient for real-world use cases.

We can bridge the modality gaps and enable information from different modalities to complement and relate to each other via MML, we can legally and compliantly get massive amounts of data via FL. Therefore, we can consider integrating FL with MML, and propose the concept of Multimodal Federated Learning (MMFL). Figure 1a and b demonstrate two different frameworks of FL and MMFL. (**a**) demonstrates the most common architecture in FL (Horizontal Federated Learning). Local clients send encrypted gradients w_k to the server, which aggregates them and sends back the w^* to the clients for updating models. Throughout this process, clients do not share datasets with each other. (**b**) shows the multimodal federated learning architecture, with technical details hidden. The key difference is that (**b**) needs to consider more scenarios(different modality datasets and different model), which is more in line with the real world. Now, our question is: How to prevent the leakage of raw data throughout the process while ensuring that performance is not excessively compromised?

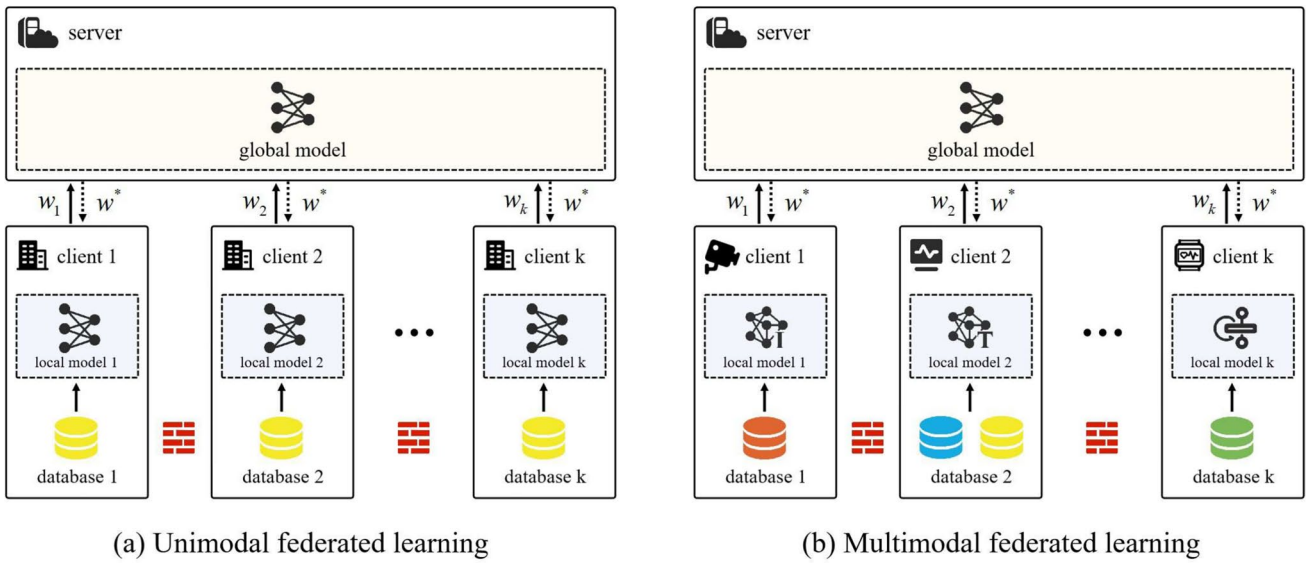

Fig. 1 Illustration of federated learning and multimodal federated learning

1.2 Our contributions

Based on our observations, MMFL is a topic that is in urgent need of further research. Most of the methods proposed so far focused on their own domains, lacking a universal standardization. We have collected a large number of ideas and research to conduct a comprehensive literature review on MMFL. Our contributions are as follows:

- We provide a new problem definition for Multimodal Federated Learning, in which we propose a new classification method for MMFL. We review the existing works related to MMFL using this classification method, highlighting their respective focuses and shortcomings.
- We provide a comprehensive summary of the datasets, evaluation metrics, and benchmarks available for research in MMFL.
- We summarize and evaluate the areas where MMFL is most likely to be applied soon and highlight the issues that demand attention in future work.

The rest of the paper is organized as follows. In Sect. 2, we provide a detailed introduction to the concept and related research of MMFL. In Sect. 3, we propose a new taxonomy that takes into account real-world applications and actual environments. In Sect. 4, we summarize multiple datasets, metrics and benchmarks. In Sect. 5, we summarize the applications of MMFL based on different modalities. Lastly, we propose some perspectives for future research of MMFL in Sect. 6, and conclude this paper in Sect. 7.

2 Overview

In this section, we establish MMFL as a novel paradigm, and provide an overview of MMFL problem formulation and relevant machine learning techniques. A brief chronology is shown in Fig. 2, article counts were collected from Google Scholar. The continuously increasing number of articles indicates that this is a research direction that is receiving more and more attention.

2.1 Problem definition

Modality refers to the way in which something happens or is experienced [16], most ML studies focus on a single modality (unimodal), where the number of modalities, denoted by M, is equal to 1. In contrast, if a research question involves $M \geq 2$ modalities, it is referred to as multimodal.

Multimodal Federated Learning is a collaborative training process involving multiple clients, each with diverse modality settings and data, conducting learning tasks without disclosing their local raw data. Specifically, MMFL aims to learn a classification model that can accurately predict the labels of local multimodal samples. From a research perspective, in addition to ensuring data security, under a given benchmark, an effective MMFL model should perform better than unimodal FL or models produced by local training, given the same hyperparameter settings.

Inspired by FL, in the MMFL system, we have K clients participating in the federation, indexed by k, who aggregate on one central server. The proportion of clients participating in training each round is C. The training process of MMFL is also similar to FL, consisting of E local epochs and T rounds of global training. For example, during the communication of the t^{th} round, $C \cdot K$ clients train in parallel for E epochs. In the entire system, there are M modalities, and each client possesses

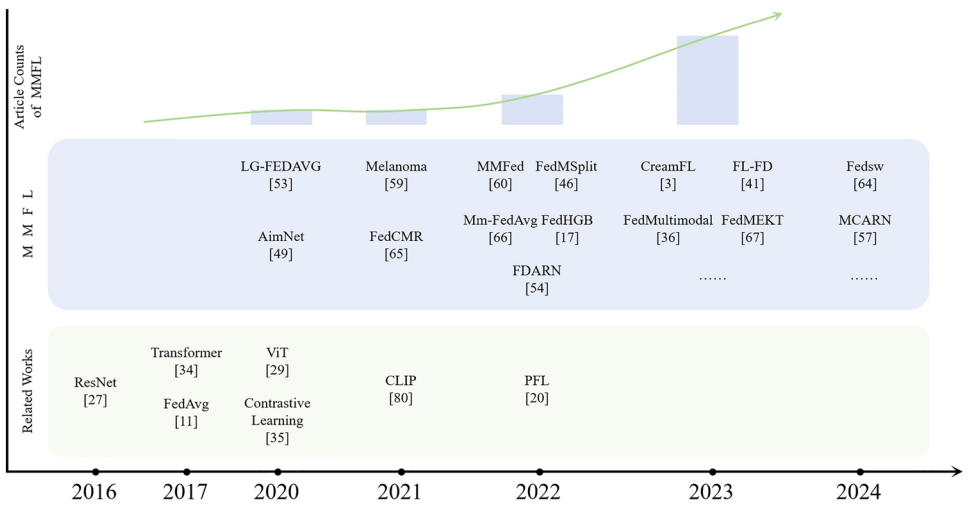

Fig. 2 A brief chronology of MMFL

M_k modalities of data. Unimodal refers to having one ($M_k = 1$) modality, while multimodal corresponds to $M_k \geq 2$. Based on the combination of modalities, the system can accommodate a maximum of $2^M - 1$ distinct client types.

The size of the datasets are related to the number of data samples. To simplify the issue, we assume that each client has only one large dataset. Dataset $\mathbf{D}_k$ has N samples, and its size is N. We define multimodal datasets as:

$$\mathbf{D}_k = \left\{ \left(\mathbf{X}_k^i, y_k^i \right) \right\}_{i=1}^{N} \quad , \tag{1}$$

$$\mathbf{X}_k^i = \left(x_k^{m_1}, x_k^{m_2}, \ldots, x_k^{m_{M_k}} \right)_i \quad , \tag{2}$$

$$\mathbf{D}_k = \left\{ \left(x_k^{m_1}, x_k^{m_2}, \ldots, x_k^{m_{M_k}}, y_k \right)_i \right\}_{i=1}^{N} \quad , \tag{3}$$

where x_k^m is the data sample of m-modality in client k, y_k represents the corresponding sample labels. $\mathbf{X}_k^i = (\cdot)_i$ represents the i^{th} data sample of the k^{th} dataset. For example, a vision-signal-text multimodal dataset, $\mathbf{X}^i = \left(x^{image}, x^{signal}, x^{text} \right)_i$.

During the local training process, each client has a target function f_k in every round, we define:

$$f_k(\omega_k) = \frac{1}{N} \sum_{i=1}^{N} \ell_k \left(\mathbf{X}_k^i, y_k^i, \omega_k \right), \tag{4}$$

where ω_k is the model parameters of each local client, it typically utilizes Stochastic Gradient Descent (SGD) for updates. ℓ_k is a user specific loss function.

If multimodal data is integrated in a late-fusion approach, the computation of ℓ_k will be very complex. The contributions of different modal data to ℓ_k are inconsistent. We consider that it will be affected by context, data heterogeneity, and downstream tasks, and we can define:

$$\ell_k \left(\mathbf{X}_k^i, y_k^i, \omega_k \right) = \sum_{p=1}^{m_{M_k}} \mathbb{A}_k^{m_p} \mathbb{B}_k^{m_p} \left(\mathbb{C}_k \left(\mathbf{X}_k^i, \omega_k \right), y_k^i \right), \tag{5}$$

where $\mathbb{A}_k^{m_p}$ represents the sum weight of modality m_p, $\mathbb{B}_k^{m_p}$ denotes the loss function of m_p. $\mathbb{C}_k$ is the local model of client-k, which is typically a classifier that outputs predictions. $\mathbb{C}_k$ can be designed as a fusion operation, it can also be a fully connected layer, or directly perform average predictions [17, 18].

We obtain the global target function F, which is defined as follows:

$$\min_{\omega^*} F(\omega_k) = \sum_{k=1}^{K} \varphi_k f_k(\omega_k), \tag{6}$$

here φ_k denotes the weight of each client. In distributed optimization algorithms, φ_k is the same for each client, whereas in unimodal FL, φ_k depends on the number of samples or communication, this parameter is one of the keys to addressing the heterogeneity issue in MMFL. ω^* represents the global aggregation weight, which is generally sent back to each client.

A specific example [19] is that if a multimodal dataset is image-text pair, we can define:

$$\mathbf{D} = \left\{ \left(\mathcal{X}^i, \mathcal{Y}^i \right) \right\}_{i=1}^{N}, \tag{7}$$

where $\mathcal{X}^i$ represents images and $\mathcal{Y}^i$ represents text, $\left(\mathcal{X}^i, \mathcal{Y}^i \right)$ is the i^{th} image-text pair with associated semantic information. At this point, the formula above can be simplified, where y can be interchanged with $\mathcal{Y}$.

2.2 Preliminaries of MMFL

The framework research for unimodal FL is relatively rich, giving rise to personalized FL (PFL) [20] and federated multitask learning (FMTL) [21]. Our real world is comprised of multimodal big data, not unimodal data, making MML a critical piece of the puzzle for practical applications of FL. The key to MML lies in the representation, translation, fusion [22], alignment [16, 23, 24] of data features, and co-learning. We believe that the focal point of MMFL research is data representation, fusion, and FMTL.

2.2.1 Representation

Representation extraction is usually the most significant differences part in multimodal architectures. Representations are typically features with lower dimensions than the original data. In the context of multimodal learning, we use the term feature and representation interchangeably [16].

Generally, during the FL process where each client only has access to a single modality of data, each local client generates representations independently, then are concatenated or fused. This overlaps with the traditional multimodal learning process. Mathematically, we define the multimodal representation as:

$$x_m^* = f(x_1^*, x_2^*, ..., x_k^*), \tag{8}$$

where $x_1^*, ..., x_k^*$ is unimodal representations. $f(\cdot)$ can be a deep neural network, restricted Boltzmann machine, or RNN.

And unimodal representations can be defined as:

$$x^* = H(\mathbf{X}, \omega_k), \tag{9}$$

where $\mathbf{X}$ represents the set of unimodal data x, and ω denotes the current local weight parameter. $H(\cdot)$ is generally a local model.

Then, we use global modal to get predictions:

$$y' = G(x^*, \omega^*), \tag{10}$$

here, y' denotes predictions, ω^* denotes the current global weight parameter, $G(\cdot)$ is a global modal.

Specifically, the most popular models used for image representation learning are CNN, such as LeNet [25], AlexNet [26], ResNet [27], VGGNet [28], and ViT [29]. For handling natural language, there are RNN (LSTM [30] and GRU [31]), Seq2seq [32], Word2vec [33], and Transformer [34]. Among these, RNN is more versatile, it can handle sequences of varying lengths and is also capable of processing modalities with temporal sequences, such as signal, video, and audio.

Some studies attempt to reduce the distance between representations by adopting contrastive learning [35], proposing Co-representations. Each modality has a corresponding projection function, which we define as:

$$f(x_1^*) \rightleftharpoons g(x_2^*). \tag{11}$$

After a series of mappings, we obtain the respective multimodal representations.

2.2.2 Fusion mechanism

Fusion plays a significant role in enhancing the performance of multimodal models [36], the representation and fusion of multimodal data in deep learning are akin to data mining, where researchers predominantly rely on four pioneering architectures that DBN [37], SAE [38], CNN, and RNN, for improvements [22]. In general, multimodal data fusion can consider three approaches: data-level fusion, feature-level fusion, and decision-level fusion [39]. In actual algorithms, they are further divided into early-fusion, late-fusion, and hybrid-fusion.

Data-level fusion involves directly integrating the input data before extracting a feature, and feature-level fusion refers to fusing after the client has extracted the feature, so both of them belong to early-fusion. In response to early-fusion, we typically employ methods such as direct feature concatenation, modal translation, element-wise multiplication, attention, and auto-encoder [40]. Decision-level fusion is a kind of voting mechanism that does not fully exploit the complementarity between modalities. Qi et al. [41] believe that data-level fusion can preserve more information from different modalities, but it potentially losing semantic information between different modalities. [16] categorized fusion strategies into two categories: Model-agnostic, which does not directly depend on specific ML methods; and Model-based, such as kernel-based approaches, graphical models, and neural networks. As shown in Fig. 3, we use different colors to represent technological differences and the same color to indicate similar technological principles.

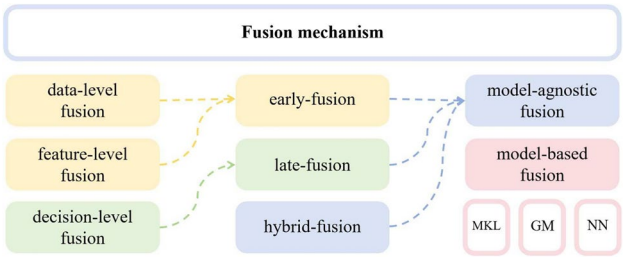

Fig. 3 Fusion mechanism taxonomy

2.2.3 Federated multitask learning

FMTL is another approach in FL settings for handling non-IID data. Smith et al. [21] reported the MOCHA framework, an extension of COCOA [42, 43], designed to tackle the implementation of FMTL in distributed network. They contributed solutions to deal with statistics heterogeneity [44] and system heterogeneity (eg. straggler [45]). Chen and Zhan [46] pointed out that MOCHA likes other methods addressing statistics heterogeneity in FL, overlooks the input modality settings. For instance, local clients in the real world may have heterogeneous sensor settings. To address the challenges of inconsistent local client sensor settings and the randomness of local client selection, they proposed FedMSplit, which employs a dynamic and multi-view graph structure. This structure uses a graph-based attention to extract and aggregate modalities. Moreover, the model is divided into shareable blocks for exploitation and exploration. Multitask learning focuses on training a model that adapts to each local client through shared parameters, or use personalized approaches to enhance the performance of multiple local clients.

3 Proposed taxonomy of MMFL

To the best of our knowledge, there does not currently exist a shared consensus among researchers on the classification methods for MMFL. In this section, we will propose a new taxonomy to introduce the specific technical details of MMFL, while also providing a systematic and multi-faceted literature review of existing works. Furthermore, we will also introduce other referenceable taxonomies to provide a comprehensive perspective.

Based on the data distribution, unimodal FL can be divided into three categories: horizontal, vertical [47], and federated transfer learning [48]. Compared to the others, horizontal FL is easier to implement, thus most MMFL algorithms fall into this category.

We believe that employing FL methods to handle multimodal data will make the situation more complex from the

client's perspective. This is because in addition to considering the statistics heterogeneity between unimodal, we also need to consider the modality heterogeneity between different modalities. Furthermore, there are also client sensor setting heterogeneity, local model heterogeneity, and more that we need to pay attention to. Unimodal FL is classified based on the sample feature space and sample IDs. In MMFL, each modality has a different feature space, and sample IDs may be the same or different. Identical sample IDs but different modalities are more like grounding problem [49], for example, certain images, certain videos, and certain texts all refer to a foundational concept in human understanding. However, in different modality combinations, the grounding will also vary, so we believe that classifying based on sample IDs is outdated, and currently lacking a standardized categorization framework [4, 18, 50].

Considering that the core of MML is the fusion of data representations, and the core issue in FL research is addressing heterogeneity, we recommend classifying MMFL based on the modality distribution of the participating parties in FL. We categorize it into three categories: modality exclusive, modality complete, and modality incomplete. Our classification method takes into account modality heterogeneity (inconsistent or missing of modalities), and model heterogeneity (inconsistent or missing of models). As shown in Fig. 4, to facilitate understanding, we use a limited number of clients and abstract model icons, and conceal various possible extension components. (**a**) The figure on the left illustrates three clients, each with a single modality of data and different models. (**b**) The middle figure shows three clients, each with multiple modalities of data and different models, with the client marked with an asterisk indicating it possesses all modalities of data from the participating clients in the training. (**c**) The figure on the right depicts three clients where the modalities involved in training are known, but there is no client with a complete set of modalities, or the modalities and models of the data are unknown(i.e., modal-model agnostic [51, 52]).

In Table 1, we provide a comprehensive summary of existing MMFL researches, and conduct an in-depth analysis of representative works in the following sections, including their application scenarios, network structures, datasets, metrics, shortcomings, and so on. Furthermore, we compare them from various perspectives. In Table 1, to maintain consistency, we have assigned abbreviations for architectures that do not have specific names in the literature.

3.1 Modality exclusive

Modality Exclusive MMFL (ME-MMFL) refers to an exclusive scenario where clients participating in federated training each have only one type of modal data or only one type of sensor, meaning that each participant is a unimodal client. It is important to note that the modalities of different clients can be the same. In some studies, this setting is also referred to as Cross-modal MMFL [57]. Formally, we define a ME-MMFL setup as:

$$D = \left\{ D_1^{m_1}, D_2^{m_2}, ..., D_k^{m_{M_k}} \right\}, \tag{12}$$

for instance, D_a^{text} and D_b^{text} may appear at the same time.

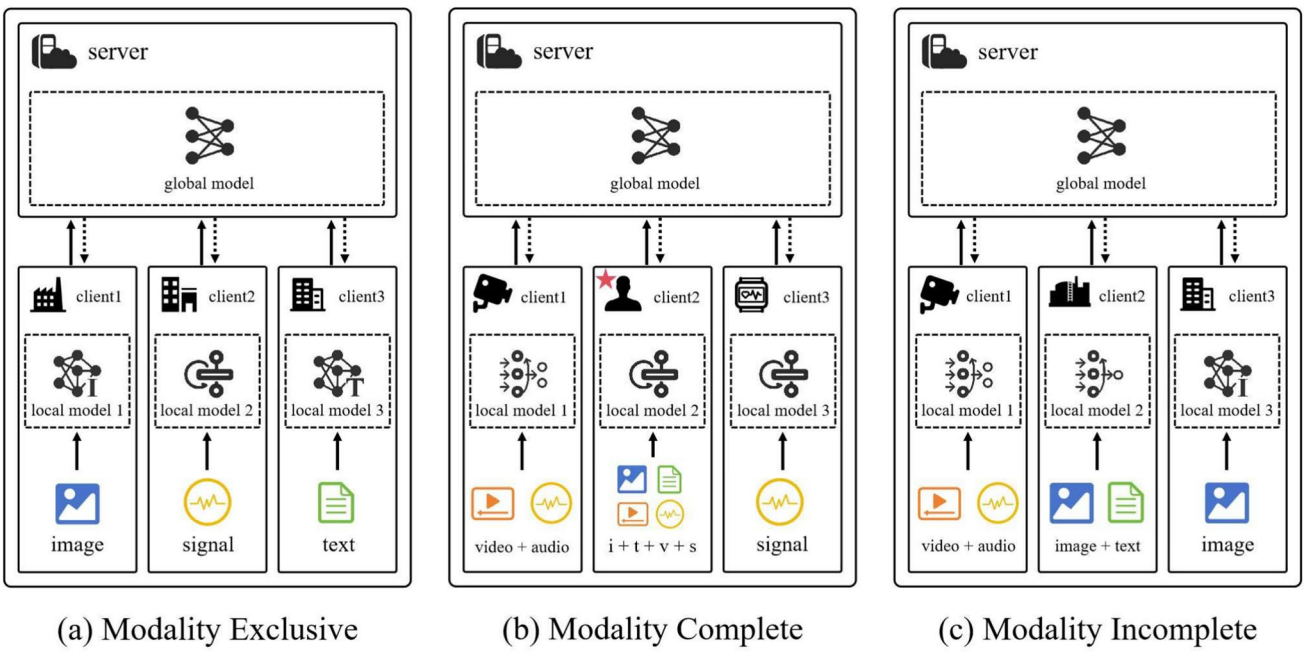

Fig. 4 Proposed taxonomy of MMFL

Table 1 An overview of existing works of MMFL

Category		Existing work	Year	Task	Core method
ME-MMFL		LG-FEDAVG [53]	2020	Image-Text	Representation learning and early-fusion
		FDARN [54]	2022	Human activity recognition	A feature-disentangled network
		CFL [55]	2022	COVID-19 diagnosis	Clustered learning and edge calculation
		FedFusion [56]	2023	Remote sensing telemetry	Manifold learning and in-orbit fusion
		MCARN [57]	2024	Human activity recognition	Modality-collaborative network
		FedMEMA [58]	2024	Brain tumor segmentation	Modality-specific encoders and multimodal anchors
MC-MMFL	TP	AimNet [49]	2020	Image-Text	Extract the fine-grained reps by bonding V-L tasks
		Melanoma [59]	2021	Melanoma detection	Late-fusion by concatenating the output features
		MMFed [60]	2022	Human activity recognition	Co-attention and personalization method
		FedHGB [17]	2022	Video classification	Hierarchical gradient blending
		FedMultimodal [36]	2023	Benchmark	Divide the training process into six parts
		FL-FD [41]	2023	Human fall detection	Transform time-series data into image data
		AutoFed [61]	2023	Vehicle driving automation	Auto-encoder-based data imputation method
		FedCLIP [62]	2023	Image-Text	An attention based adapter
		Harmony [63]	2023	Alzheimer's disease monitoring	Modality-wise FL and federated fusion learning
		Fedsw [64]	2024	Medical image report	CNN and Transformer
	TR	FedCMR [65]	2021	Image-Text	Aggregate the updates of all common subspaces
		Mm-FedAvg [66]	2022	Human activity recognition	Auto-encoder
		FedMEKT [67]	2023	Human activity recognition	Distillation-based embedding knowledge transfer
		CreamFL [3]	2023	Image-Text	Contrastive Learning and Knowledge Distillation
		PFedPrompt [68]	2023	Image-Text	Prompt training and use local personalized attention
		PmcmFL [69]	2023	Image-Text	Introduce a prototype library
		FedUSL [70]	2024	Driving fatigue detection	Project multimodal sensing data onto a unified space
MI-MMFL		FedMSplit [46]	2022	General	Graph-based attention
		MMVFL [71]	2023	General (VFL)	Two-step multimodal model for Vertical FL
		FedSea [72]	2023	Search and classification	Use domain adversarial learning for feature alignment
		mmFedMC [73]	2024	General	Selecting modalities and local clients
		DisentAFL [51]	2024	General	Disentanglement of asymmetric knowledge into symmetric

TP and TR respectively refer to transmission parameters and transmission representations

In this case, handling each unimodal data separately at the data-level is a convenient approach. To the best of our knowledge, Liang et al. [53] were the earliest explorers. In their research on how to introduce representation learning into FL, they used a multimodal dataset VQA to verify the applicability of LG-FEDAVG. They employed LSTM and ResNet-18 [27] as local and global models. The global model uses element-wise multiplication to achieve early-fusion of the image and question layers, and LG-FEDAVG performed better than FedAvg. This was a mainstream early-fusion method at that time, but a large number of experiments have shown that the accuracy is usually not high. The reason is that element-wise multiplication cannot truly allow the information of the two modalities to complement each other. During the COVID-19 pandemic, how to quickly diagnose pneumonia symptoms was a challenge at that time. Qayyum et al. [55] designed a MMFL method for diagnosing COVID-19, focusing their research on data from X-ray and Ultrasound modalities. They based their approach on clustered FL and edge computing, using the VGG16 [28] model to process the two types of data separately, and then sharing a multimodal model. The improved accuracy and recall rate prove that the method is effective. However, the article only discusses a limited dataset and model, the method may be difficult to generalize to more modalities, and edge computing brings additional heterogeneity issues.

These types of methods are almost all direct modifications of FedAvg [11], where clients exchange model parameters with the server, which may result in the loss of many data features. Moreover, due to the lack of consideration for modal alignment, they are also unable to capture the semantic relationships between different modalities. Then, some research works attempt to exchange different model parameters or vectors for different modalities, such as encoder parameters, in order to preserve as much useful features of individual modalities and complementary information between modalities as possible.

Yang et al. [54] proposed a method called FDARN for the human activity recognition problem in 2022. FDARN involves Modality-Agnostic and Modality-Specific representation learning. They mapped the data on local clients into a Modality-Agnostic feature space and then shared it among different clients. Specifically, this approach maintains separate private classifiers and a shared classifier, and it has demonstrated an increase in accuracy compared to traditional FL. Subsequently, they expanded the model and further proposed MCARN [57]. They designed a flexible angular margin adjustment scheme and a relation-aware global–local calibration mechanism to mitigate the differences between modalities, and supplemented with extensive experiments to demonstrate that the method performs well in both balanced and unbalanced modality scenarios. However, the various encoders and classifiers in the proposed framework pose challenges for practical application, as clients require sufficient computational power and resources. Moreover, the accuracy of MCARN does not show a significant improvement compared to other baseline models.

Li et al. [56] focused on the issue of satellite detection and proposed a manifold-driven multimodal fusion framework known as FedFusion. The research involved HSI satellites and LiDAR satellites, with each satellite acting as a client. These clients do not need to transmit raw data back to earth for centralized processing, they can also collaborate with each other, ultimately learning an optimal global model. FedFusion consists of the local data module, global data spreading, and classifier. The core of the fusion strategy is the SVD module. Specifically, the local features of a client can be decomposed into three smaller vectors through SVD. These vectors are then distributed to each client for reconstruction and subsequent concatenation. Dai et al. [58] focused on the problem of brain tumor image segmentation and designed a relatively simple MMFL framework called FedMEMA. The server has four different encoders and a multimodal fusion decoder, while also using a cross-attention module to calibrate missing modalities, ultimately forming a multimodal representation. On a particular medical dataset, this approach demonstrates superior performance in comparison to its predecessors. Nonetheless, the methodology presented in the article tackles solely the heterogeneity that exists among different modalities, failing to address the intricacies of intra-modality heterogeneity and client-side heterogeneity.

In response to the situation in MMFL where each client possesses only one modality, researchers have achieved breakthroughs from various perspectives. In general, these works exemplify three distinct fusion strategies: one is direct connection, another is a common subspace approach, and the third is based on attention mechanisms. Methods like LG-FEDAVG, which establish direct connections, find it harder to retain more of the original information compared to approaches like FDARN and MCARN that map onto a common subspace. However, this shared subspace could be adversely affected by negative data, leading to suboptimal outcomes. The attention mechanisms can mitigate these challenges. Nonetheless, these methodologies are indeed paving the way for addressing more intricate scenarios in the future.

3.2 Modality complete

Modality Complete MMFL (MC-MMFL) indicates that there is a client that possesses all modalities of data, with some overlap among the clients' data. Specifically, one scenario is where all clients have multimodal data with consistent modality types. Another scenario is that the implementation of the method depends on the client having a complete modal type, with the remaining clients providing supplementary information. We can define a MC-MMFL setup as:

$$D = \left\{D_1^{m_1}, D_1^{m_2}, D_2^{m_1}, D_2^{m_2}, D_3^{m_1}, D_3^{m_2}\right\} \tag{13}$$

or

$$D' = \left\{D_1^{m_1}, D_2^{m_1}, D_2^{m_2}, D_3^{m_2}\right\}. \tag{14}$$

Most existing studies can be categorized under this category. Regarding the extraction and fusion of representations, some architectures are similar to those in ME-MMFL. By reviewing a large number of research works, we categorize these innovative methods into two types: transmission parameters and transmission representations. As shown in Fig. 5, we summarize the frameworks involved in MC-MMFL into four types.

3.2.1 Transmission parameters

Transmission parameters refers to the process where each client participating in FL first fuses multimodal data, then uses a local model to train and obtain model parameters, and subsequently exchanges these model parameters with the server. Another common method involves each client setting different feature extractors and models for different modalities of data. In addition to exchanging model parameters, they also share the parameters of the feature extractors. As shown in Fig. 5a. The next subsections will specifically introduce the implementation, advantages, and disadvantages of these methods.

Specifically, local clients extract features from different multimodal data, then fuse these features, and the fused features are used to obtain model parameters through the model. This is similar to most methods in ME-MMFL. In 2021, Agbley et al. [59] applied FL to melanoma detection. Their proposed framework utilized a CNN (EfficientNet) and a custom network to process medical images and

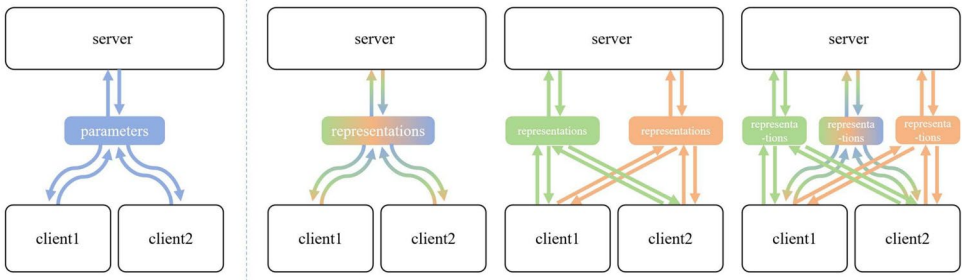

Fig. 5 Common methods of MC-MMFL

(a) Transmission parameters (b) Transmission representations

clinical data separately, and performed late-fusion to connect the two types of features. This method continues the traditional FL paradigm, with client and server exchanging model parameters. In terms of metrics such as accuracy, F1, and AUC, this method demonstrates that after FL is introduced to multimodality, it is also competitive compared to centralized learning. Moreover, this study only focused on the heterogeneity between different modalities, and the downstream tasks were relatively simple.

Xiong et al. [60] proposed MMFed in 2022, they design a Co-attention and personalization method based on Model-Agnostic Meta-Learning (MAML) [74], which achieved improved accuracy on video-signal datasets, focused on human activity recognition [75]. Their method is an early and more comprehensive attempt at MC-MMFL. Co-attention allows for the complementarity of information across different modalities, while MAML enables each client to have a personalized model rather than a global model. MMFed achieves federated training and protects data privacy by transmitting the parameters of the classifier and the attention module. Although the accuracy rate is higher than that of traditional FL methods, the value is still not high, which may be due to data imbalance or a relatively simple backbone. In addition, the communication cost also needs further investigation. Chen et al. [64] proposed a method for generating medical image reports based on MMFL, which utilizes CNN (Resnet-101/Densenet-121) to extract image features and obtain predictive labels. The report generator encodes visual features and predictive labels into a series of hidden features, and then use Transformer [34] in conjunction with textual data. In addition, they introduced Fedsw, which calculates a score for each client, and only updates the weights if the client's score is lower than the aggregated score. The method they proposed shows some improvement in BLEU scores compared to independent local training. However, these methods all overlook the fact that when there are differences in the quality of datasets, performance enhancement cannot be guaranteed through federated training. Additionally, using attention strategies brings about a high cost.

The work of Qi et al. [41] also only considers the scenario where modalities are consistent, but their clever fusion strategy is enlightening. Their FL-FD uses the Gramian Angular field method to encode one-dimensional time series data into two-dimensional images, and then stacks two types (2D time-series and processed camera data) of visual features to form a three-channel image for fusion. This is a clever approach that integrates all modal data as much as possible, improving accuracy while reducing model complexity, which is beneficial for IoT devices with limited performance. However, the limitations are also apparent, the method needs to be further expanded to other modal data and more complex downstream tasks.

The existing MMFL works are all addressing specific tasks within their respective fields. In order to provide a benchmark for MMFL, Feng et al. [36] proposed an end-to-end MMFL framework named FedMultimodal in 2023. They divided the training process into six parts: data partition, feature processing, multimodal models, fusion schemes, FL optimizers, and noise factor emulator. Considering the limited computational resources, memory, and battery life of the majority of clients involved in FL, they adopted a lightweight attention with a Conv+RNN architecture to fuse multimodal representation. The additional benefit is, FedMultimodal can be easily applied to most FL algorithms, such as FedAvg [11], FedProx [12], and FedOpt [76]. In the task of crisis recognition, the accuracy of this method for image-text modal data is relatively low. The main reason may be that the backbone network failed to extract more effective features.

In MC-MMFL, modality missing is a common issue, and Zheng et al. have provided a data imputation method based on inter-modal autoencoders, AutoFed [61]. It uses available modalities to fill in missing ones, thereby mitigating modality heterogeneity. This is a feature-level fusion that employs a cross-attention mechanism for alignment. Their model has achieved robust results (object detection) in autonomous driving scenarios. This is a preliminary attempt at intelligent transportation using the MMFL method, and the improvements in accuracy and recall rate demonstrate the effectiveness of the method. However, autonomous driving

is a sensitive issue, and many people are currently skeptical about this technology. The article does not discuss issues such as adversarial attacks, broader privacy protection, and the sustainability of the model. Including that they also point out that the incentive issue for participants cannot be resolved for the time being, which is a promising direction for research.

Due to the varying performance and data sizes of clients, the straggler problem can easily arise in FL, Ouyang et al. [63] have proposed Harmony. Harmony addresses MMFL issues through a two-stage framework, which includes modality-wise FL and federated fusion learning. Modality-wise FL helps to resolve the problem of imbalanced training latencies among different nodes and modalities in MMFL, while the latter is used to address the fusion challenge. After the two-stage training, the server sends both multimodal and unimodal models to all nodes, taking into account the possibility of sensor failures. The method has significantly improved accuracy compared to the baseline and has greatly reduced training latency.

These types of research works focus on the most common MMFL scenario, therefore, we further summarize the feature extractors, datasets, evaluation metrics, shortcomings, etc., used in various studies, as shown in Table 2.

Table 2 Summary of the transmission parameters representative methods

Method	Modality	Extractor	Dataset	Metrics	Weakness
AimNet [49]	Text	[77]	Flickr30k+MSCOCO+VQA	BLEU-4/METEOR/CIDEr/SPICE	Neglects the computational cost,lacks validation on more image-text tasks
	Image	Faster R-CNN			
Melanoma [59]	Text	Custom network	Skin lesions	Accuracy/F1/AUC	Focus only on the classification problem of modality heterogeneity
	Image	EfficientNet-B6			
MMFed [60]	Video	Inception-V3+Resnet50	Multimodal Data +MMC-PCL-Activity	Accuracy	The backbone is simple, model performance can be further improved
	Signal	LSTM			
FedHGB [17]	RGB	Recognizer3D	Kinetics+Finegym	Accuracy	The computational cost is high, should be verified in more scenarios
	Optical flow	Recognizer3D			
	Audio	AudioRecognizer			
FedMultimodal [36]	Text	MobileBERT/Distill-BERT	MELD+CREMA-D+KU-HAR +PTB-XL+UCF101+MiT10 +Hateful-Memes+MiT51 +CrisisMMD+UCI-HAR	Accuracy/F1/AUC/UAR	Unable to process unlabeled datasets, the fusion strategy is relatively simple
	Image	MobileNetV2/MobileViT			
	Video	MobileNetV2/MobileViT			
	Audio	MFCC			
	Signal	Conv+RNN			
FL-FD [41]	Video	Gray image	UP Fall	Accuracy/Recall/F1/Precision	Requires performance verification across more scenarios
	Signal	Gramian Angular field			
AutoFed [61]	Lidar	CNN	Radar RobotCar +nuScenes	IoU/Precision/Recall	Unable to address more complex security and incentive issues
	Radar	CNN			
FedCLIP [62]	Text	CLIP(ViT-B/32)/AlexNet	PACS+VLCS +Office-Home	Accuracy	Global model may be adversely affected by the client data
	Image	CLIP(ViT-B/32)/AlexNet			
Harmony [63]	Image	CNN/3D-CNN	Real-World testbed +USC+MHAD+FLASH	Accuracy	Does not discuss the issues of data distribution heterogeneous and transmissi on security
	Radar	3D-CNN			
	Audio	RNN			
	Signal	2D-CNN			
	Lidar	3D-CNN			
Fedsw [64]	Text	Transformer	Iu-xray	BLEU	Overlooks the communication and computational overhead
	Image	Densenet-121/Resnet-101			

3.2.2 Transmission representations

Transferring more abstract features during FL process has been a common idea in recent years. Recently, researchers have attempted to transmit representations that result from the fusion of multimodal data. It should be noted that "Transmission Representations" likely encompasses both representations and parameters. As shown in Fig. 5b, the left section illustrates each client either fusing or mapping all modalities into a single representation, which then participates in federated training. The middle section shows clients generating distinct representations for each modality, with each representation separately participating in federated training. The right section shows clients simultaneously maintaining different modality representations and a multimodal representation.

Zhao et al. [66] utilized auto-encoder to fuse data from different modalities, and introduced Mm-FedAvg framework, they assume that on multimodal clients, there are unlabeled but aligned data pairs, and the performance of the auto-encoder also depends on data with complete modalities. Specifically, the method transfers multimodal representations to align information from different modalities, which is more effective than ordinary fusion strategies, and the F1 of this method on three datasets is higher than the results of unimodal data. In addition, the framework takes into account the situation where client modalities are missing. Although this work addresses a limited set of heterogeneous issues, the problem scenarios it presents are practical, and the methods are pioneering.

Cross-modality retrieval is a research direction that has attracted much attention in multimodal learning. Inspired by DSCMR [78], Zong et al. [65] developed FedCMR, to address the challenge of handling multimodal data from each client in a FL setting. In their approach, they effectively utilized VGGNet for image data and BERT [79] for text data, creating a shared public sub-space modality for image-text and a linear classifier. The primary goal of FedCMR is to discover a globally consistent latent common subspace for modal fusion. The experimental results indicate that FedCMR shows superior performance over DSCMR, FedAvg, and FedProx in cross-modal retrieval tasks, highlighting its effectiveness in MMFL environments. However, to verify the generalizability of the method, this paper does not discuss situations where client data is severely unbalanced or extremely unevenly distributed.

Recently, Yu et al. [70] proposed the FedUSL framework for the problem of fatigue driving detection. In FedUSL, which is transmitted between the clients and the server are projection matrices. These projection matrices serve as the key components for exchanging information between the server and clients. Specifically, during each global iteration, the server collects the projection matrices from all clients, aggregates them, and then distributes the aggregated projection matrices to each client for the next round of learning. The experimental data includes video and signals, and the results demonstrate that the proposed method performs well even when only a limited amount of labeled data is provided for specific modalities. The method has reduced the number of communication rounds compared to the baseline, which contributes to rapid response in practical applications. However, broader privacy protection and computational costs are non-negligible considerations.

If we interpret multimodal representations as the knowledge that models acquire from multimodal data, then the process of continuously enriching these multimodal representations can be seen as knowledge transfer. FedMEKT [67], proposes a semi-supervised knowledge transfer mechanism. It utilizes joint embedding knowledge transfer instead of parameter exchanging, which enhances communication efficiency and strengthens the model generalizability. Specifically, FedMEKT iteratively updates the generalized global encoder for local learning with joint multimodal embedding knowledge from clients through upstream and downstream multimodal embedding knowledge transfer. This approach reduces communication overhead and prevents reverse engineering, thereby enhancing privacy protection.

There are also some special frameworks, utilizing multimodal public datasets, and we classify this situation as MC-MMFL as well. Yu et al. [3] is the first to introduce knowledge distillation into MMFL, CreamFL restricts clients to exchange representation on a public dataset, using an improved contrastive learning method to reduce model drift, and designing a global-local contrastive aggregation method during the aggregation phase. Because it has a global representation for all modalities based on public datasets, in this work, local clients of any modality can interact with the global representation. Although this method has achieved high performance, the quality of the model is directly dependent on the quality of the public dataset, which also brings high computational and communication costs. In addition, in the presence of a malicious attacker, how to ensure the secure transmission and aggregation of representations remains an important issue.

In summary, both methods of transferring only parameters and transferring representations have their drawbacks. If only model parameters are transferred, the potential of multimodal data cannot be further explored. On the other hand, if multimodal representations are transferred between the clients and server, it significantly increases communication costs and the difficulty of training. Additionally, the security of transferring representations has not been rigorously evaluated in existing works.

3.3 Modality incomplete

Modality Incomplete MMFL (MI-MMFL) refers to a situation where none of the clients have data of all modalities, or the client modality is unknown (modal-model agnostic). This scenario is more intricate than the approaches discussed previously. Some research employs graph structures or shared modules to delve deeper into semantic information across different modalities.

In scenarios where the specific data modalities are not known in advance but the task is clear, it is possible to perform federated training across all available modalities and then carry out a local fusion process. However, this approach can lead to a higher computational burden. By strategically selecting clients that contribute significantly to the task at hand, the overall computational requirements can be substantially reduced. Yuan et al. [73] introduced the mmFedMC framework, it employs a decision-level fusion strategy, where they modularize the models and train each unimodal model separately. By carefully selecting modalities and local clients to minimized communication overhead, their experimental results proved the effectiveness of the approach. The shortcomings lie in the fact that mmFedMC sets too many hyperparameters, which is a clear disadvantage as not all datasets achieved high accuracy.

In more complex situations where both modalities and tasks are unknown, combining specific modalities with specific tasks and training these combinations or knowledge simultaneously can effectively enhance performance. FedMSplit [46] innovatively handled the issue of the missing of modality from multimodal settings by employing a graph-based attention. FedMSplit segments each client's data into two categories of blocks: one for blocks specific to individual modalities and the other for blocks that are globally shareable. By employing weighted aggregation, clients with similar statistical properties become more relevant to each other. However, they did not take into account the high computational cost and negative impact induced by server-side aggregations performed by different encoders. Recently, Chen and Zhang [51] proposed the DisentAFL. DisentAFL is a novel MMFL framework based on a two-stage knowledge disentanglement and gating mechanism, designed to address the ultimate scenario of MMFL, namely Modality-task Agnostic Federated Learning (AFL). They also point out that AFL is an essential path to achieving Artificial General Intelligence (AGI). The key to DisentAFL is to decompose the original asymmetric inter-client information sharing scheme into several independent symmetric inter-client information sharing schemes, each corresponding to a type of semantic knowledge learned from local tasks. The method was evaluated in five AFL simulation experiments, and the results show that its performance on AFL is superior to that of the baseline methods. Although this method provides a pathway for AFL to combine different modalities and tasks into knowledge, the complex network is bound to incur high computational and time complexity costs.

3.4 Other perspectives

In the previous parts of the section we reviewed most of the existing MMFL frameworks according to the proposed taxonomy. We will now revisit those methods or frameworks from other perspectives in order to provide a more complete and comprehensive view.

Supervised learning is an indispensable part of machine learning, but in the real-world, not all data can be effectively labeled [50, 78, 80]. Therefore, semi-supervised and unsupervised learning have recently gained popularity in FL. Similarly, existing MMFL frameworks can also be divided into Supervised MMFL, Semi-supervised MMFL, and Unsupervised MMFL (Self-supervised MMFL). Most ME-MMFL and MI-MMFL methods can be classified as Supervised MMFL, with representative works such as DisentAFL [51], FedMSplit [46], and MMFed [60]. Semi-supervised MMFL refers to situations where not all data have labels, which is a very common scenario because often times data collected by sensors are involved in learning before they can be labeled. Some of the representative works are MmFedAvg [66], CreamFL [3], FedUSL [70], and FedMEKT [67]. Unsupervised MMFL is a more complex setting, to be precise, there is currently no fully Unsupervised MMFL. Sun [81] has explored Unsupervised MMFL in a study of multimodal federated transfer learning. In their experimental setup, unimodal clients still require labeled data, while multimodal clients can use contrastive learning for unsupervised learning. Annotating massive amounts of multimodal data is expensive, and extensive professional knowledge is often a requirement for annotators. Therefore, Semi-supervised MMFL remains a promising research direction, and fully Unsupervised MMFL is going to be one of the challenges that needs to be addressed in future work [82].

Personalized FL is an extension of traditional FL, aiming to learn a personalized model for each participating client while protecting data privacy and achieving global optimization for the model [20, 83]. In MMFL, the modality gaps generate the demand of each client to have a personalized model. Hence, it is an issue worth further investigations. Based on this, we can categorize existing frameworks into Personalized-MMFL and Non-personalized-MMFL. Most studies only focus on one classification task within a single domain, and the resulting models are generally applicable. Therefore, these methods can be classified into Non-personalized-MMFL. FedMSplit and DisentAFL are two similar such frameworks. FedMSplit splits different modalities into shared blocks corresponding to the same task, while

DisentAFL correlates different modalities and different tasks to address more complex issues. pFedPrompt [68] was proposed to address the problem of personalized prompt, enabling the prompts learned by clients through the MMFL process to be fully personalized and aligned with the users' local characteristics. These Personalized-MMFL methods are an active area of research.

4 Datasets and benchmarks

4.1 Datasets of MMFL

To the best of our knowledge, there is currently no dataset specifically designed for MMFL. Therefore, following the convention in FL research, we partition unimodal datasets to simulate heterogeneous clients. Instead of listing out those unimodal datasets that are most readily accessible, we focus on categorizing some high-quality and highly adaptable multimodal datasets according to their application scenarios, as shown in Table 3. While only multimodal datasets are presented in the paper, we note that studies have shown that the results of hybrid training with both multimodal and unimodal data tend to be better than those with unimodal data alone or multimodal data exclusively [66].

Vision-Language Model (VLM), is recently a focal point of multimodal research, as image and text network has become relatively mature. Under the guidance of multimodal pre-training models such as CLIP [80], ViLT [84], and VLMo [85], the integration of image and text is becoming increasingly tight-knit.

- Flickr30k [86] is an image-text dataset containing over 31,000 images, which feature various regions and cultures. Each image is accompanied by 5 annotated sentences, totaling more than 150,000 sentences.
- MS COCO [19] is a dataset specifically designed for object detection, segmentation, localization, and captioning. The dataset has been divided into three parts: test, train, and val, with over 160,000 images, and corresponding annotated texts of various tasks. It is a high-quality multimodal dataset.
- VQA Dataset [87] is a Visual Question Answering dataset, Goyal et al. have annotated each image with its corresponding questions and answers. The dataset is also of the image-text type. It's an interesting dataset, and their complete balanced dataset contains approximately 1.1 million (image, question) pairs.

Multimodal Emotion Recognition (MER). Emotions are omnipresent in human daily life, influencing our judgments and decisions [88]. Multimodal models recognize human emotional states through modalities such as audio, vision, and text. Common data features include facial expressions, speech, actions, images, text, electroencephalography, and more. MER is extensively utilized across various domains, including social media emotion analysis, medical treatment, mental health, fatigue monitoring [89].

- IEMOCAP [90] is a video-audio-text dataset recorded by ten actors, consisting of 151 dialogue videos with 2 speakers per segment, totaling 302 videos in the dataset, and the corpus data spans approximately twelve

Table 3 Multimodal datasets

Application	Dataset	Modality	Ref size	Primary task
Vision-Language Model	Flickr30k	Image-text	31,000 * 5	Classification
	MS COCO	Image-text	160,000 * n	Classification
	VQA Dataset	Image-text	1,110,000	Question-answer
Multimodal emotion recognition	IEMOCAP	Video-audio-text	151 * 2	Classification
	MELD	Video-audio-text	1443 * 9	Classification
	CMU-MOSEI	Video-audio-text	3,228/23,000	Classification
Multimodal human recognition	Kinetics-400	Video-text	300,000	Classification
	UCF101	Video-audio-text	13,320	Classification
	UR fall detection dataset	Image-signal-text	70	Classification
Social media analysis	Hateful Memes	Image-text	10,000	Classification
	UR-FUNNY	Video-audio-text	1,866	Classification
	CrisisMMD	Image-text	18,100	Classification
Autonomous vehicles	Vehicle Sensor	Audio-signal-text	–	Classification
Healthcare	mHealth dataset	Signal-text	120 * 3	Classification
	PTB-XL	Signal-text	21,837	Classification
Object recognition	ModelNet40	Mesh-point cloud-text	12,311	Classification

hours. Each clip is annotated with 9 emotions (happiness, anger, neutral, disgust, fear, etc.).
- MELD [91] is an video-audio-text dataset containing over 1400 dialogues and 13,000 utterances, sourced from the TV series Friends. Each utterance in the dialogues is labeled with one of the following seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. In addition, each utterance is also marked as positive, negative, or neutral.
- CMU-MOSEI [92] is a video-audio-text dataset sourced from monologue videos on YouTube. The dataset includes over 65 h of annotated videos, featuring more than 1,000 different speakers and covering 250 distinct topics. These video clips are further segmented into over 23,000 sentences, encompassing six types of emotions.

Multimodal Human Recognition (MHR) is another popular research direction in multimodal learning [60, 93], where people typically train multimodal models using video-audio-signal-text to classify and understand human activities. Cameras can capture video, mobile devices can record audio, and smartwatches can gather signals, making these data easily accessible in real life. Therefore, MHR is a practical direction. Human recognition includes activities such as human action recognition [94, 95], pose recognition, and object localization, etc.

- Kineics-400 [96], this is a video dataset of human actions with YouTube URLs, containing over 300,000 video sequences across 400 categories. Each clip is approximately 10 s long.
- UCF101 [97] consists of 13,320 videos with 101 sport-based action labels. All the videos are collected from YouTube and have a fixed frame rate of 25 FPS. However, only half of the data includes both video and audio information, which allows the dataset to be used for simulating situations that modality missing.
- UR fall detection dataset [98] is a dataset for human fall monitoring, containing 70 action sequences (30 fall events and 40 daily activities), with modalities including image-signal, where the images consist of original RGB and depth images. This dataset can be applied in medical settings or in projects aimed at supporting aging populations.

Social Media Analysis (SMA) has been widely applied in apps distribution, content Recommendation, government monitoring, and advertising marketing. It involves analyzing information on social media to determine the emotions, motivations, identities of the posters, and the trends and impact of information dissemination. Generally, the modalities involved image-video-audio-text.

- Hateful Memes [99] dataset is used for monitoring hateful speech in information, containing 10,000 image-text data, with corresponding words cover the images, and each image-text pair is labeled as either benevolent or malicious.
- UR-FUNNY [100] dataset is a multimodal collection designed to understand humor, encompassing video-audio-text. The data is sourced from public TED Talks, which have been transcribed and annotated with timestamps indicating when the audience laughs. We believe that this dataset, when integrated with AI assistants, can make models more humorous. Besides, it can help models to generate humorous text.
- CrisisMMD [101] dataset contains 18,100 image-text tweets from the X platform (Twitter), covering discussions about seven natural disasters in 2017. It provides annotations of three types. This dataset can help relevant personnel extract key information from tweets for rapid response during disasters.

There are also some multimodal datasets that play a significant role in the fields of IoT, healthcare, and object recognition. We have also listed them.

- Vehicle sensor [102], it contains 23 instances. Each instance includes acoustic features, seismic features, and infrared features. This dataset is typically used for vehicle identification and was initially processed using k-NN and SVM algorithms.
- MHealth dataset [103] includes records of body movements and signs of ten humen with varying conditions during 13 activities, including exercise, standing, sitting, and relaxation. These data are measured by various sensors and can be used not only for human activity recognition but also for heart rate prediction.
- PTB-XL [104] ECG dataset contains 21,837 10-second 12-lead ECG recordings from 18,885 patients. It is the largest freely accessible clinical 12-lead ECG waveform dataset to date. It plays a significant role in ECG analysis and medical diagnosis.
- ModelNet40 [105] dataset is a collection for 3D object recognition and classification. The original ModelNet40 consists of 12,311 CAD-generated meshes across 40 categories, with 9843 used for training and 2468 for testing. Each model is labeled, such as furniture, animals, everyday items, and so on. The improved version of ModelNet40 is in the form of point cloud-text.

4.2 Metrics of MMFL

The preceding content has provided a wealth of available multimodal datasets for VLM, multimodal emotion recognition, multimodal human recognition, social media analysis,

autonomous vehicles, healthcare, and object recognition. To facilitate further exploration and expansion in this field by future researchers, in the following section we will introduce some commonly used evaluation metrics.

Accuracy is a commonly used evaluation metric in machine learning. Since most datasets are designed for classification tasks, using Test Accuracy to assess the performance of individual models is considered reasonable. In binary classification tasks, the calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{15}$$

In methods such as MMFed [60], FedMSplit [46], FDARN [54]/MCARN [57], FL-FD [41], Harmony [63], FedFusion [56], DisentAFL [51], etc., Accuracy is used as the primary evaluation metric.

F_1 *Score* is the harmonic mean of Precision and Recall, a metric that takes into account both precision and recall, particularly suitable for scenarios where both precision and recall are highly important. The F_1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor model performance.

$$Precision = \frac{TP}{TP + FP} \quad , \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \quad , \tag{17}$$

$$F_1 = \frac{Precision \cdot Recall \cdot 2}{Precision + Recall} \quad , \tag{18}$$

where precision focuses on counting the number of samples predicted as positive by the model that are actually positive. A high precision means the model rarely predicts negative samples as positive erroneously. Recall measures how many of the actual positive samples are correctly predicted as positive by the model.

For example, FedCMR [65], CreamFL [3], and AutoFed [61] use precision and recall to compare with specific baselines, while Mm-FedAvg [66] uses the F_1 score to demonstrate the framework's performance. In addition, metrics such as AUC and BLEU are also used to evaluate MMFL frameworks.

4.3 Benchmarks of MMFL

Most existing research on MMFL focuses on a specific domain and rarely uses frameworks and overlapping datasets that are designed for general domains. Additionally, due to the involvement of different modalities and client settings, the baselines used in existing work are either unimodal FL or modifications of existing MMFL frameworks.

Under the complex settings of MMFL, we believe that existing single testing dataset evaluation methods are no longer suitable for future research, and this view is supported by the work in [36]. There are already many benchmarks for image-text multimodality, such as using multimodal datasets for testing, visual question answering, or simple zero-shot and few-shot approaches.

Liang et al. [106] have stated that the development of visual-language multimodal learning has experienced significant advancements, but other modalities are still quite underdeveloped (time-series, finance, set, force sensors, proprioception). To support their view, their MultiBench was designed to be a systematic and unified large-scale multimodal learning benchmark that covers 15 datasets, 10 modalities, 6 research domains, and 20 prediction tasks.

Unlike MML, MMFL requires a comprehensive consideration of various "gaps", such as modality gaps, model gaps, task gaps, and so on. Recent MMFL algorithms have been tested using modality complete datasets, and compared with classic FL algorithms that have been slightly modified. Feng et al.'s end-to-end framework transforms all modalities into a single representation, allowing FL to be directly experimented with and compared without modification. In their experiments, FedOpt [76] demonstrated state-of-the-art performance.

We agree with Radford et al.'s viewpoint [80] that, in addition to the existing dataset-based competition, we should encourage more multimodal learning algorithms to use zero-shot transfer to evaluate the model's learning ability, which closely aligns with the original goal of AI research. Additionally, the benchmarks for MMFL algorithms also require continuous exploration by researchers.

5 Applications

In the real world, unimodal FL faces limitations such as communication costs, data heterogeneity, and hardware differences. When the scope is expanded to multimodal, MMFL tasks become even more challenging. We have seen in Sect. 4 a variety of multimodal datasets each with a rich set of corresponding MMFL application scenarios, regardless of the quality and scale of the dataset. We summarize the main application scenarios of MMFL and classify them as shown in Fig. 6. VLM and IoT are the most widely used scenarios for MML and MMFL applications. Based on these two settings, it is possible to extend to more niche application areas. In this section, we will discuss some representative scenarios.

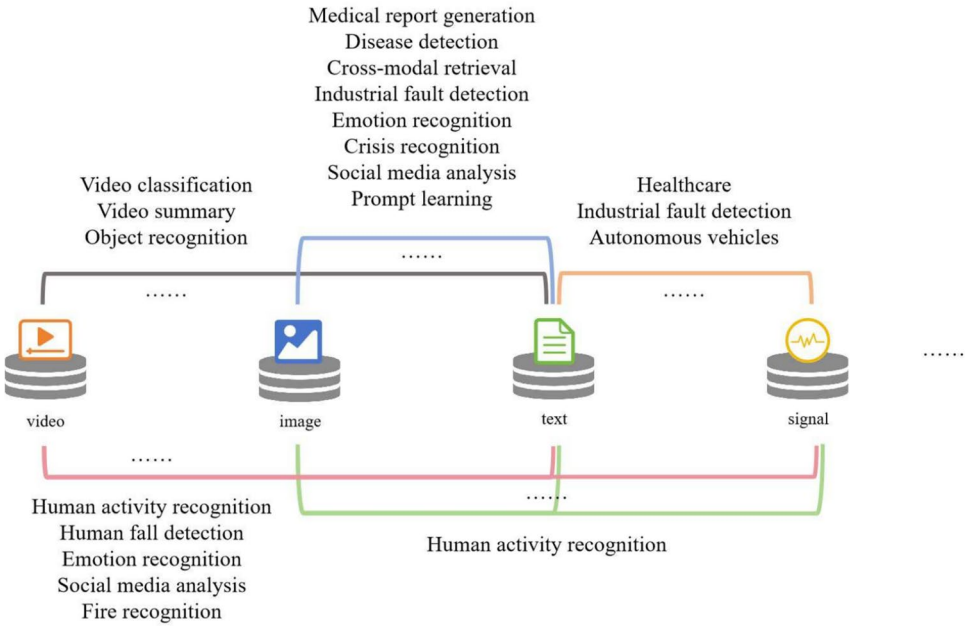

Fig. 6 Taxonomy of MMFL applications

5.1 Vision-language model

Thanks to the large-scale deployment of video and image collectors, vision-language tasks have become a core focus in both MML and MMFL. Based on rich image-text datasets, Vision-Language Model (VLM) is widely applied in image captioning, visual question answering, visual reasoning, image-text retrieval, text-to-image generation, content recommendation, medical assistance, social media analysis, and other areas [107]. The challenges for VLMs mainly lie in data distribution heterogeneity and client heterogeneity in FL. Large-scale VLMs like GPT-4 indeed perform well, but not all local clients have the capability to train or run such models. Therefore, implementing lightweight and general-purpose VLMs, and adaptively selecting local clients, are research directions for MMFL in the field of Vision-Language.

5.1.1 Cross-modal retrieval

Cross-modal Retrieval is a hot multimodal application in recent years, aiming to retrieve another corresponding modality through one modality. In MMFL, some works are also studying text-image cross-modal retrieval, mainly using contrastive learning [35] or mapping to a common subspace to process two types of modal data. FedCMR [65] had conducted earlier explorations of Federated Cross-modal Retrieval tasks and narrows the representation gap, by using a weighted aggregation based on the number of local data and categories. CreamFL [3] introduced the concepts of contrastive learning and CLIP [80] into FL, where each client only uses a common dataset to exchange knowledge. It outperforms the baseline model in both image recall of text and text recall of images.

5.1.2 Emotion recognition

In recent years, Emotion Recognition has sparked immense interest across a range of application fields, including social media, healthcare, and autonomous driving. Current methods for recognizing emotions are mainly based on visual cues, speech, or physiological signals. FedCMD [108] utilizes the cross-modal distillation technique, aiming to recognize the emotional state of drivers from unlabeled videos. In the education field, teenager psychological issues have become a significant societal concern, with various surveys indicating an increase in the suicide rate among teenagers. Some efforts have already been made to identify emotional states from social media content [109]. We suggest that keyboard input data could also be incorporated for a more comprehensive MMFL, facilitating early intervention by parents and schools. However, it is important to note that when training models for such issues, researchers should collect data with the consent of the teenagers and their parents, and ensure that sensitive data is anonymized [110].

5.1.3 Medical diagnosis

In the medical industry, different hospitals or clinics participate in FL using their own medical images, patient descriptions, diagnostic results, treatment plans, and patient activity videos, etc. This global model can learn a wealth of knowledge and provide personalized medical reports for each patient. However, due to the privacy and confidentiality

of medical data, there are significant issues with the sharing and analysis of this data. Therefore, MMFL frameworks such as Fedsw [64], FedMEMA [58], and Melanoma [59] have been proposed, paving the way for solving more complex medical problems in the future.

5.2 Internet of things

Plenty of Internet of Things (IoT) devices equipped with different sensors can lead to overlapping information, meaning that the features exhibited by an object may vary from different views. The modalities involved include video, image, text, signal, and time-series, and more. Connecting IoT devices in homes can enable smart home appliances, appealing to both furniture manufacturers and customers. Implementing FL on sensors within factory machinery and robots can lead to automatic fault detection and solution generation. Hospitals can monitor patient conditions in unattended wards through linked monitors and sensors. This often takes the form of inter-device FL. In general, current applications of distributed learning can naturally extend to MMFL.

5.2.1 Human activity recognition

Human Activity Recognition has garnered widespread attention in various fields such as health monitoring, smart homes, and smart cities. It generally involves modalities of data including video, depth images, audio, acceleration signals, and radar signals. In the past, Human Activity Recognition involved training on aggregated data, which could not effectively protect user privacy. MMFL is the optimal solution for this issue. Recently, the experiments of MMFed [60], Mm-FedAvg [66], FDARN [54], MCARN [57], FedMEKT [67] have demonstrated that using MMFL can effectively protect user privacy without significantly compromising accuracy. Additionally, Human Fall Detection [41] is one of the emerging directions, which can be applied in nursing homes, hospitals, and other fields, allowing vulnerable populations to receive timely protection and treatment.

5.2.2 Autonomous driving

In the intelligent transportation field, each street or city utilizes public data such as cameras, traffic signals, and crowd heat maps, along with personal vehicle cameras, vehicle sensors, and mobile phone signals, while ensuring smooth and secure data transmission, government departments or companies can help autonomous driving technology respond faster or provide the public with scientific travel suggestions in real time. AutoFed [61] is an early attempt at this application. With the popularization of new energy vehicles and AI-assisted driving, autonomous driving technology will be a future trend.

5.2.3 Embodied intelligence

In the industrial field, recent research has focused on Embodied Intelligence [111], where AI has already acquired various flexible bodies (robots). If it integrates with external cameras and sensors in FL, AI will also possess a brain. Currently, there is no research that combines robots with MMFL, this technology is scarce.

6 Perspectives

Many researchers have already answered the questions we posed at the beginning of this paper with their commendable MMFL frameworks, but MMFL technology is still far from reaching a mature state. In future work, we should also consider the following aspects and potential areas of improvement.

6.1 Modality heterogeneity

The core of modality heterogeneity is rooted in the fusion mechanism. In the current MMFL research landscape, modal data are either handled separately before integration or combined from the initial stage, with various innovative techniques applied to the fusion modules. The commonality among them is that they focus solely on the heterogeneity of the input side. True modality heterogeneity is not just about different data types but also about different semantics, which remains an open challenge. By reviewing a vast amount of literature, we believe that transmitting more abstract representations or sharing encoder parameters may facilitate a deeper integration of different modalities.

6.2 Hardware heterogeneity

In federated training, the diversity of client devices leads to hardware heterogeneity issues. This includes heterogeneity in computational power and communication capabilities. Under the conditions of FL, local models are relatively simple and the data is unimodal, making communication costs the primary concern. When client data is multimodal, computation costs can no longer be ignored. Some works simplistically deploy Transformer to run on local clients, which is unreasonable as not all clients have strong computation capabilities. In light of this, some researchers continuously activate clients with good computation power and communication conditions, which raises fairness issues. The issue of client selection and incentive in MMFL is one of the future research directions. Therefore, we recommend that researchers simplify complex models, or place most of the

computations on the server side or edge device, while simultaneously improving the model's convergence capabilities to reduce the communication burden.

6.3 Trustworthy MMFL

Trustworthy MMFL refers to the process of federated training of multimodal data that possesses robustness, privacy, and generalization. Robustness is in response to scenarios where participating parties may conduct Poisoning attacks, Backdoor attacks, and Adversarial attacks, which can affect the performance of the global model. Privacy pertains to the protection of data during the transmission of parameters, as well as the generation and fusion of data representations, where attackers or malicious participants might launch inference attacks aimed at reconstructing the original raw data. Generalization aims to ensure that the learned model can perform well in real-world environments. We suggest that researchers integrate technologies such as multi-party secure computation, differential privacy, homomorphic encryption, and blockchain more deeply with MMFL in their future work to further realize trustworthy MMFL. The presence of third-party servers introduces inherent security vulnerabilities. In subsequent research, it is advisable to explore the adoption of decentralized structures, like graph-based structures, to facilitate direct peer-to-peer communication among clients.

6.4 Unsupervised learning

In existing MMFL research, publicly available annotated multimodal datasets are an essential component, despite some works exploring semi-supervised MMFL. In real-world scenarios, the vast amount of data collected by sensors comes with limited annotations, and the complexity of modality alignment and client diversity increases the difficulty of knowledge sharing. Therefore, future work should consider using weak-supervised and unsupervised learning setups more.

6.5 Personalized-MMFL

The fundamental goal of MMFL is to achieve a model that performs better than individual unimodal models or few-sample multimodal models. The heterogeneity of modalities can lead to more challenging gradient convergence, making it potentially unfeasible to train a perfect global model. Personalized FL is an important direction for future research in FL, and the demand for personalization in MMFL is higher than ever. We have previously demonstrated some MMFL methods from the personalized perspective. Generally, Personalized-MMFL can be achieved through two approaches: one is to first learn a global model and then fine-tune it locally on the client, and the other is to modify the federated process to directly learn a personalized solution suitable for local clients. Both of these issues are at the initial stage of research and are a challenging problem.

6.6 Knowledge transfer

Knowledge transfer has become one of the hottest research directions in unimodal FL, but in addition to knowledge transfer within the same modality, cross-modal knowledge transfer from one modality to another poses a new challenge in MMFL scenario. For example, learning from image captioning tasks and transferring the acquired knowledge to video captioning tasks is a promising direction. There has already been work that introduces methods such as unimodal knowledge distillation into MMFL, but this is far from sufficient, multimodal knowledge distillation is still in its infancy. Additionally, for datasets with missing annotations, knowledge transfer can help supplement the dataset labels.

6.7 Prompt learning

Represented by GPT-4o, multimodal large language models (MLLMs) have further advanced the development of multimodal learning. Specifically, these MLLMs are based on text, with other modalities creating associations, allowing humans to input various modalities of information to interact with computers. Inspired by visual-prompt learning, a few studies have begun to explore prompt learning in MMFL, transferring prompts or adapters in federated training, such as supplementing text prompts with learned visual prompts. Under the FL setting, the collaborative training of multiple MLLMs is a promising research direction.

6.8 Interpretability

Interpretability in MMFL encompasses both the interpretability of modality alignment and the interpretability of the model itself. Researchers have already made breakthroughs in the interpretability of image-text, but the mapping relationships between other modalities remains a direction that needs exploration. The most direct method is to implement more parameter and representation visualizations in the research process in the future, thereby understanding the model's decision-making process. Furthermore, MMFL involves a large number of free parameters and, like FL, faces challenges in security auditing. If the model lacks interpretability, it could lead to potential security issues.

6.9 Novel benchmarks

In ML, when we evaluate new frameworks, we typically opt to compare them using the same datasets against multiple baselines. This approach is highly effective in the context of

unimodal FL, where the downstream tasks for each client are consistent. However, in MMFL, the downstream tasks for clients may not be uniform, and researchers may need to employ multiple multimodal datasets to test various downstream tasks. Additionally, due to the limited number of multimodal datasets, some annotated datasets exhibit biases to varying extents, such as language biases. We require a set of benchmarks that have been practically validated, taking into account security, fairness, accuracy, and generalization capabilities.

7 Conclusion

In this survey, we introduce the integration of multimodal learning with federated learning, MMFL. We provide a relatively comprehensive background, present the definition of MMFL, and review recent research works. We also propose a new classification approach for MMFL from the new perspective, list a wide array of multimodal datasets, metrics and benchmarks. Moreover, we also introduce a multitude of promising application directions and outline potential issues and future research directions. We are confident that our contributions will serve as a valuable resource for professionals and researchers in related fields that offer insights and guidance for the advancement of MMFL technologies.

Acknowledgements This work was supported by National Natural Science Foundation of China, 62376151. Science and Technology Commission of Shanghai Municipality, 22DZ2205600.

Author contributions H.P. wrote the main manuscript text. All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

1. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2506–2515 (2019)
2. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., Poria, S.: Towards multimodal sarcasm detection (an _obviously_ perfect paper). Preprint at arXiv:1906.01815 (2019)
3. Yu, Q., Liu, Y., Wang, Y., Xu, K., Liu, J.: Multimodal federated learning via contrastive representation ensemble. Preprint at arXiv:2302.08888v3 (2023)
4. Thrasher, J., Devkota, A., Siwakotai, P., Chivukula, R., Poudel, P., Hu, C., Bhattarai, B., Gyawali, P.: Multimodal federated learning in healthcare: a review. Preprint at arXiv:2310.09650 (2023)
5. Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L.: A comprehensive survey on cross-modal retrieval. Preprint at arXiv:1607.06215 (2016)
6. Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: a review. ACM Comput. Surv. **56**(3), 1–39 (2023)
7. Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S.W., Albuquerque, V.H.C.: A comprehensive survey of multi-view video summarization. Pattern Recogn. **109**, 107567 (2021)
8. Liang, P.P., Liu, T., Cai, A., Muszynski, M., Ishii, R., Allen, N., Auerbach, R., Brent, D., Salakhutdinov, R., Morency, L.-P.: Learning language and multimodal privacy-preserving markers of mood from mobile data. Preprint at arXiv:2106.13213 (2021)
9. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Trans. Med. Imaging **41**(10), 2598–2614 (2022)
10. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Found. Trends® Mach. Learn. **14**(1–2), 1–210 (2021)
11. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282 PMLR, (2017)
12. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. Learn. syst. **2**, 429–450 (2020)
13. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning, pp. 5132–5143 PMLR, (2020)
14. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. Preprint at arXiv:2102.07623 (2021)
15. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Adv. Neural Inform. Proc. Syst. **33**, 7611–7623 (2020)
16. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2018)
17. Chen, S., Li, B.: Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In: IEEE INFOCOM 2022-IEEE Conference on Computer Communications, pp. 1469–1478 IEEE, (2022)
18. Che, L., Wang, J., Zhou, Y., Ma, F.: Multimodal federated learning: a survey. Sensors **23**(15), 6986 (2023)
19. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755 Springer, (2014)
20. Tan, A.Z., Yu, H., Cui, L., Yang, Q.: Towards personalized federated learning. IEEE Transactions on Neural Networks and Learning Systems (2022)
21. Smith, V., Chiang, C.-K., Sanjabi, M., Talwalkar, A.S.: Federated multi-task learning. Advances in neural information processing systems **30** (2017)
22. Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. Neural Comput. **32**(5), 829–864 (2020)
23. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)

24. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/script: Alignment and parsing of video and text transcription. In: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10, pp. 158–171 Springer, (2008)
25. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint at arXiv:1409.1556 (2014)
29. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv:2010.11929 (2020)
30. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
31. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. Preprint at arXiv:1412.3555 (2014)
32. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems **27** (2014)
33. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Preprint at arXiv:1301.3781 (2013)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30,**(2017)
35. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 PMLR (2020)
36. Feng, T., Bose, D., Zhang, T., Hebbar, R., Ramakrishna, A., Gupta, R., Zhang, M., Avestimehr, S., Narayanan, S.: Fedmultimodal: A benchmark for multimodal federated learning. Preprint at arXiv:2306.09486 (2023)
37. Zhang, N., Ding, S., Zhang, J., Xue, Y.: An overview on restricted boltzmann machines. Neurocomputing **275**, 1186–1199 (2018)
38. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. Preprint at arXiv:1812.05069 (2018)
39. Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M., Falk, T.H.: A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Inform. Fusion **76**, 355–375 (2021)
40. Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y.: Multi-modal sensor fusion for auto driving perception: A survey. Preprint at arXiv:2202.02703 (2022)
41. Qi, P., Chiaro, D., Piccialli, F.: Fl-fd: Federated learning-based fall detection with multimodal data fusion. Inform. Fusion **99**, 101890 (2023)
42. Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., Jordan, M.I.: Communication-efficient distributed dual coordinate ascent. Advances in Neural Information Processing Systems **27** (2014)
43. Ma, C., Smith, V., Jaggi, M., Jordan, M., Richtárik, P., Takác, M.: Adding vs. averaging in distributed primal-dual optimization. In: International Conference on Machine Learning, pp. 1973–1982 PMLR, (2015)
44. Ye, M., Fang, X., Du, B., Yuen, P.C., Tao, D.: Heterogeneous federated learning: state-of-the-art and research challenges. ACM Comput. Surv. **56**(3), 1–44 (2023)
45. Reisizadeh, A., Tziotis, I., Hassani, H., Mokhtari, A., Pedarsani, R.: Straggler-resilient federated learning: Leveraging the interplay between statistical accuracy and system heterogeneity. IEEE J. Selected Areas Inform. Theory **3**(2), 197–205 (2022)
46. Chen, J., Zhang, A.: Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 87–96 (2022)
47. Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., Yang, Q.: Vertical federated learning: Concepts, advances, and challenges. IEEE Transactions on Knowledge and Data Engineering (2024)
48. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
49. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Federated learning for vision-and-language grounding problems. In: Proceedings of the AAAI Conference on Artificial Intelligence, **34**, pp. 11572–11579 (2020)
50. Lin, Y.-M., Gao, Y., Gong, M.-G., Zhang, S.-J., Zhang, Y.-Q., Li, Z.-Y.: Federated learning on multimodal data: a comprehensive survey. Mach. Intell. Res. **4**, 1–15 (2023)
51. Chen, J., Zhang, A.: On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, **38**, pp. 11311–11319 (2024)
52. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. Adv. Neural Inform. Proc. Syst. **33**, 3557–3568 (2020)
53. Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.-P.: Think locally, act globally: Federated learning with local and global representations. Preprint at arXiv:2001.01523 (2020)
54. Yang, X., Xiong, B., Huang, Y., Xu, C.: Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, **36**, pp. 3063–3071 (2022)
55. Qayyum, A., Ahmad, K., Ahsan, M.A., Al-Fuqaha, A., Qadir, J.: Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. IEEE Open J. Comput. Soc. **3**, 172–184 (2022)
56. Li, D., Xie, W., Li, Y., Fang, L.: Fedfusion: Manifold driven federated learning for multi-satellite and multi-modality fusion. IEEE Transactions on Geoscience and Remote Sensing (2023)
57. Yang, X., Xiong, B., Huang, Y., Xu, C.: Cross-modal federated human activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
58. Dai, Q., Wei, D., Liu, H., Sun, J., Wang, L., Zheng, Y.: Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, **38**, pp. 1445–1453 (2024)
59. Agbley, B.L.Y., Li, J., Haq, A.U., Bankas, E.K., Ahmad, S., Agyemang, I.O., Kulevome, D., Ndiaye, W.D., Cobbinah, B., Latipova, S.: Multimodal melanoma detection with federated learning. In: 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 238–244 IEEE, (2021)
60. Xiong, B., Yang, X., Qi, F., Xu, C.: A unified framework for multi-modal federated learning. Neurocomputing **480**, 110–118 (2022)

61. Zheng, T., Li, A., Chen, Z., Wang, H., Luo, J.: Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. Preprint at arXiv:2302.08646 (2023)
62. Lu, W., Hu, X., Wang, J., Xie, X.: Fedclip: Fast generalization and personalization for clip in federated learning. Preprint at arXiv:2302.13485 (2023)
63. Ouyang, X., Xie, Z., Fu, H., Cheng, S., Pan, L., Ling, N., Xing, G., Zhou, J., Huang, J.: Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In: Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services, pp. 530–543 (2023)
64. Chen, J., Pan, R.: Medical report generation based on multimodal federated learning. Comput. Med. Imaging Graph. **113**, 102342 (2024)
65. Zong, L., Xie, Q., Zhou, J., Wu, P., Zhang, X., Xu, B.: Fedcmr: Federated cross-modal retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1672–1676 (2021)
66. Zhao, Y., Barnaghi, P., Haddadi, H.: Multimodal federated learning on iot data. In: 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI), pp. 43–54 IEEE, (2022)
67. Le, H.Q., Nguyen, M.N., Thwal, C.M., Qiao, Y., Zhang, C., Hong, C.S.: Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. Preprint at arXiv:2307.13214 (2023)
68. Guo, T., Guo, S., Wang, J.: pfedprompt: Learning personalized prompt for vision-language models in federated learning. In: Proceedings of the ACM Web Conference 2023, pp. 1364–1374 (2023)
69. Bao, G., Zhang, Q., Miao, D., Gong, Z., Hu, L.: Multimodal federated learning with missing modality via prototype mask and contrast. Preprint at arXiv:2312.13508 (2023)
70. Yu, S., Yang, Q., Wang, J., Wu, C.: Fedusl,: A federated annotation method for driving fatigue detection based on multimodal sensing data. ACM Trans. Sensor Netw. (2024). https://doi.org/10.1145/3657291
71. Gong, M., Zhang, Y., Gao, Y., Qin, A., Wu, Y., Wang, S., Zhang, Y.: A multi-modal vertical federated learning framework based on homomorphic encryption. IEEE Transactions on Information Forensics and Security (2023)
72. Tan, M., Feng, Y., Chu, L., Shi, J., Xiao, R., Tang, H., Yu, J.: Fedsea: Federated learning via selective feature alignment for non-iid multimodal data. IEEE Transactions on Multimedia (2023)
73. Yuan, L., Han, D.-J., Wang, S., Upadhyay, D., Brinton, C.G.: Communication-efficient multimodal federated learning: Joint modality and client selection. Preprint at arXiv:2401.16685 (2024)
74. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135 PMLR, (2017).
75. Hu, M., Luo, M., Huang, M., Meng, W., Xiong, B., Yang, X., Sang, J.: Towards a multimodal human activity dataset for healthcare. Multimed. Syst. **29**(1), 1–13 (2023)
76. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., McMahan, H.B.: Adaptive federated optimization. Preprint at arXiv:2003.00295 (2020)
77. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J.C., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
78. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10394–10403 (2019)
79. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv:1810.04805 (2018)
80. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 PMLR, (2021)
81. Sun, Y.: Federated transfer learning with multimodal data. Preprint at arXiv:2209.03137 (2022)
82. Saeed, A., Salim, F.D., Ozcelebi, T., Lukkien, J.: Federated self-supervised learning of multisensor representations for embedded intelligence. IEEE Int. Things J. **8**(2), 1030–1040 (2020)
83. Wang, J., Yang, X., Cui, S., Che, L., Lyu, L., Xu, D.D., Ma, F.: Towards personalized federated learning via heterogeneous model reassembly. Adv. Neural Inform. Proc. Syst. **36** (2024)
84. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. International Conference on Machine Learning, 5583–5594 PMLR (2021)
85. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Adv. Neural Inform. Proc. Syst. **35**, 32897–32912 (2022)
86. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)
87. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6904–6913 (2017)
88. Zhao, S., Jia, G., Yang, J., Ding, G., Keutzer, K.: Emotion recognition from multiple modalities: fundamentals and methodologies. IEEE Signal Proc. Mag. **38**(6), 59–73 (2021)
89. Chaturvedi, V., Kaur, A.B., Varshney, V., Garg, A., Chhabra, G.S., Kumar, M.: Music mood and human emotion recognition based on physiological signals: a systematic review. Multimed. Syst. **28**(1), 21–44 (2022)
90. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
91. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. Preprint at arXiv:1810.02508v6 (2018)
92. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 1, pp. 2236–2246 Long Papers, (2018)
93. Huang, Y., Yang, X., Gao, J., Sang, J., Xu, C.: Knowledge-driven egocentric multimodal activity recognition. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **16**(4), 1–133 (2020)
94. Singh, R., Sonawane, A., Srivastava, R.: Recent evolution of modern datasets for human activity recognition: a deep survey. Multimed. Syst. **26**(2), 83–106 (2020)
95. Chao, X., Hou, Z., Mo, Y., Shi, H., Yao, W.: Structural feature representation and fusion of human spatial cooperative motion for action recognition. Multimed. Syst. **29**(3), 1301–1314 (2023)
96. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P.,

et al.: The kinetics human action video dataset. Preprint at arXiv:1705.06950 (2017)
97. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. Preprint at arXiv:1212.0402 (2012)
98. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Prog. Biomed. **117**(3), 489–501 (2014)
99. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: detecting hate speech in multimodal memes. Adv. Neural Inform. Proc. Syst. **33**, 2611–2624 (2020)
100. Hasan, M.K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.-P., et al.: Ur-funny: A multimodal language dataset for understanding humor. Preprint at arXiv:1904.06618 (2019)
101. Alam, F., Ofli, F., Imran, M.: Crisismmd: Multimodal twitter datasets from natural disasters. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
102. Duarte, M.F., Hu, Y.H.: Vehicle classification in distributed sensor networks. J. Parallel Distrib. Comput. **64**(7), 826–838 (2004)
103. Banos, O., Garcia, R., Holgado-Terriza, J.A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C.: mhealthdroid: a novel framework for agile development of mobile health applications. In: Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6, pp. 91–98 Springer, (2014)
104. Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Sci. Data **7**(1), 154 (2020)
105. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920 (2015)
106. Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., et al.: Multibench: Multiscale benchmarks for multimodal representation learning. Preprint at arXiv:2107.07502 (2021)
107. Li, X.: Tag relevance fusion for social image retrieval. Multimed. Syst. **23**(1), 29–40 (2017)
108. Bano, S., Tonellotto, N., Cassarà, P., Gotta, A.: Fedcmd: A federated cross-modal knowledge distillation for drivers emotion recognition. ACM Transactions on Intelligent Systems and Technology (2024)
109. Liang, P.P., Liu, T., Cai, A., Muszynski, M., Ishii, R., Allen, N., Auerbach, R., Brent, D., Salakhutdinov, R., Morency, L.-P.: Learning language and multimodal privacy-preserving markers of mood from mobile data. Preprint at arXiv:2106.13213 (2021)
110. Li, Z., Cheng, W., Zhou, J., An, Z., Hu, B.: Deep learning model with multi-feature fusion and label association for suicide detection. Multimed. Syst. **29**(4), 2193–2203 (2023)
111. Gupta, A., Savarese, S., Ganguli, S., Fei-Fei, L.: Embodied intelligence via learning and evolution. Nature Commun. **12**(1), 5721 (2021)