Research
A SCIENCE PARTNER JOURNAL

SPJ SCIENCE PARTNER JOURNALS

Manuscript Template

Multimodal Metaverse Healthcare: A Collaborative Representation and Adaptive Fusion Approach for Generative AI-Driven Diagnosis

Jianhui Lv[1,2*], Adam Slowik[3], Shalli Rani[4], Byung-Gyu Kim[5], Chien-Ming Chen[6], Saru Kumari[7], Keqin Li[8], Xiaohong Lyu[1], and Huamao Jiang[1*]

[1]The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121012, China
[2]Peng Cheng Laboratory, Shenzhen 518057, China
[3]Koszalin University of Technology, Koszalin 98701, Poland
[4]Chitkara University, Punjab 140401, India
[5]Sookmyung Women's University, Seoul, Republic of Korea
[6]Nanjing University of Information Science & Technology, China
[7]Chaudhary Charan Singh University, Meerut, India
[8]State University of New York, New Paltz, NY 12561, USA

*Address correspondence to: lvjh@pcl.ac.cn (J. Lv); jianghm@jzmu.edu.cn (H. Jiang).

Abstract

The metaverse enables immersive virtual healthcare environments, presenting opportunities for enhanced care delivery. A key challenge lies in effectively combining multimodal healthcare data and generative artificial intelligence abilities within metaverse-based healthcare applications, which is a problem that needs to be addressed. This paper proposes a novel multimodal learning framework for metaverse healthcare, MMLMH, based on collaborative intra- and inter-sample representation and adaptive fusion. Our framework introduces a collaborative representation learning approach that captures shared and modality-specific features across text, audio, and visual health data. By combining modality-specific and shared encoders with carefully formulated intra-sample and inter-sample collaboration mechanisms, MMLMH achieves superior feature representation for complex health assessments. The framework's adaptive fusion approach, utilizing attention mechanisms and gated neural networks, demonstrates robust performance across varying noise levels and data quality conditions. Experiments on metaverse healthcare datasets demonstrate MMLMH's superior performance over baseline methods across multiple evaluation metrics. Longitudinal studies and visualization further illustrate MMLMH's adaptability to evolving virtual environments and balanced performance across diagnostic accuracy, patient-system interaction efficacy, and data integration complexity. The proposed framework has a unique advantage in that a similar level of performance is maintained across various patient populations and virtual avatars, which could lead to greater personalization of healthcare experiences in the metaverse. MMLMH's successful functioning in such complicated circumstances suggests it can combine and process information streams from several sources. They can be successfully utilized in next-generation healthcare delivery through virtual reality.

INTRODUCTION

The metaverse is a collective virtual shared space created by the fusion of physically persistent virtual spaces and enhances physically persistent virtual spaces [1]. Today, the metaverse has emerged as an overarching platform with wide-reaching consequences in delivering healthcare services and educating health professionals. Although the metaverse has received considerable healthcare consideration, it has other applications, such as integrated energy systems [2] and digital enterprise management [3]. Consequently, it showcases the potential of the metaverse as a revolutionary platform. By fashioning interactive three-dimensional (3D) spaces, the metaverse enhances remote patient management, virtual interaction, and teamwork in healthcare [4,5]. Medicine in virtual worlds presents unprecedented benefits, as practitioners can attend to patients, retrieve medical information, and collaborate with other professionals regardless of physical location. The possibilities are further extended by the infusion of generative artificial intelligence (AI) in the metaverse healthcare environments, which permit realistic simulations, individualized treatment programs, and tailored educational pathways for patients and health workers alike [6]. When this digital transformation in healthcare is imminent, there is a growing need to formulate appropriate strategies to interrogate and interpret the various types of data generated in such complex electronic networks, which are now considered impossible [7].

As regards their technological aspect, multimodal machine learning is a unique model that is appealing in the health care of the metaverse, given that it applies multiple sources of patient data for better diagnosis and monitoring results [8–11]. Within the scope of healthcare in a metaverse, relevant modalities may include text and audio medical records, patient-doctor conversations, medical images and remote monitor data, virtual examinations, and even patient haptic data. This data type is paramount in defining the completeness of the patient's profile in a virtual setting. It enables health practitioners to make practical decisions on the patient's health in the virtual realm. Additionally, large language models and text-to-image generators as generative AI technologies involve adding theoretical patients with temporal information to real patients, and this kind of synthesized data had to be integrated with actual patient data [12–16]. This duality of multimodal data analysis and generative AI presents the possibility of greatly enhancing today's practices of metaverse healthcare.

Multimodal representation learning and multimodal fusion are two core components of creating AI systems for metaverse healthcare [17,18]. Contemporary multimodal representation approaches often focus on intra-sample collaboration, resulting in weak feature representation across patient cases. This shortcoming is particularly disadvantageous in metaverse healthcare, where patient records could be pooled from several virtual environments and under different circumstances. Also, the majority of available solutions provide care attention to noisy data [19]. The same problems could also be found in reverse: the inverse variation of adjusting capabilities had caused many interference effects from or to the data from generative AI models during patient monitoring [20–22]. These challenges create significant barriers to developing and deploying robust AI-based diagnostic and monitoring systems in metaverse healthcare environments [23,24]. This may harm patient care and safety in virtual clinics where such technologies are implemented.

In response to challenges, we propose a novel multimodal learning framework for metaverse healthcare, MMLMH. MMLMH is advanced by incorporating intra- and inter-sample collaborative representation learning as well as adaptive fusion techniques to tackle the heterogeneous and dynamic features of health data in virtual settings. The method provides a more holistic view of complex relations between health data and health phenomena across multiple modes, thereby improving the quality of features for AI-centric metaverse healthcare systems.

The primary advancement of MMLMH is its innovative method for processing healthcare data within virtual environments, characterized by three essential architectural innovations. Our collaborative intra- and inter-sample representation learning mechanism effectively addresses the challenge of maintaining clinical relevance in virtual consultations. In contrast to current approaches that handle each modality separately, MMLMH's dual-level collaboration effectively captures immediate relationships within individual patient interactions and overarching patterns across multiple cases.

The second core innovation focuses on MMLMH's adaptive fusion architecture, specifically tailored for the dynamic characteristics of healthcare interactions within the metaverse. This framework presents a calibration mechanism that continuously evaluates and modifies the reliability weights of various modalities in real-time. This capability is essential for processing diverse quality levels of virtual avatar movements, transmitted physiological sounds, and patient-provider dialogues.

MMLMH's integrated approach to managing real and synthetic healthcare data in virtual environments is a notable characteristic. The framework includes a specialized encoder that learns the relationships between physical clinical indicators and their virtual representations. This innovation facilitates the integration of AI-generated clinical scenarios with actual patient data, ensuring consistent performance across hybrid datasets.

METHODS

The MMLMH aims at multimodal natural language understanding in metaverse healthcare environments, where text is the primary input. At the same time, audio and video features are secondary. Consider a dataset $\mathcal{M} = r_1, r_2, ..., r_n$ containing n samples. Each sample $r_i, i = 1, 2, ..., n$ consists of a video clip v, an audio recording α, a text transcript τ, and a health related label η. In other words, we construct a model $g(\tau, v, \alpha) \to \eta$, where τ, v, α are the textual, video, and audio components from a sample citation r_i respectively, such that η, the health-related label, is correct. The architecture of the MMLMH framework depicted in Fig. 1 comprises four sections: a component for preliminary feature extraction, a component for learning multimodal representation, a component for integrating multimodal data, and a component for predicting health outcomes.

Initial feature extraction

For metaverse healthcare applications, such information often comes from virtual visits, self-reported symptoms, or attested clinical documents. The textual data is often complex, and to obtain valuable insights from this type of data, we use an adjusted version of bidirectional encoder representations from transformers (BERT) trained on health, which we refer to as MedBERT [25]. MedBERT was chosen as our text-processing foundation

due to its extensive training in medical literature, clinical notes, and patient records, enabling a nuanced understanding of healthcare terminology and contextual relationships. Its transformer architecture has demonstrated superior performance in capturing long-range dependencies in clinical narratives, making it particularly suitable for processing detailed patient-provider interactions in virtual consultations. The model's domain-specific pretraining on over 2 million clinical documents ensures accurate interpretation of medical terminology and symptom descriptions.

Given a sequence of text tokens $\boldsymbol{\tau} = [\tau_1, \tau_2, ..., \tau_{l_\tau}]$, where l_τ is the sequence length, this input is passed through MedBERT through several layers of transformer. The last hidden output is employed as the representation of text:

$$\boldsymbol{\psi}_\tau \in \mathbb{R}^{l_\tau \times F_\tau}. \tag{1}$$

where $\boldsymbol{\psi}_\tau$ is the text feature representation, l_τ is the text sequence length, and F_τ is the feature dimension.
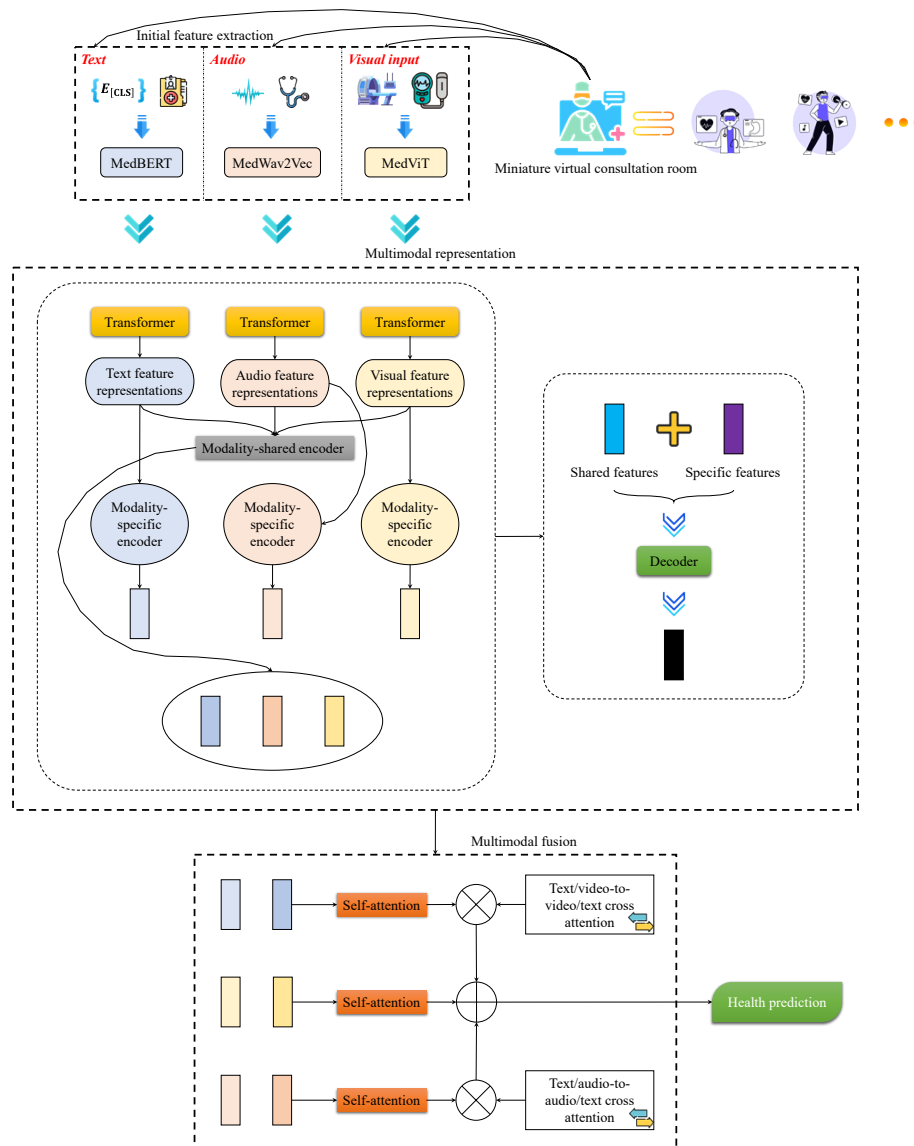
Fig. 1. MMLMH architecture.

Metaverse healthcare applications may feature audio recordings, like patient voice recordings, heart or lung sounds, or even just noise in a virtual setting. As for the deep learning mapping task, we use a modified Wav2Vev 2.0 model called MedWav2Vec, trained on medical audio datasets. The selection of MedWav2Vec for audio processing was determined by the specialized architecture of the model optimized for medical acoustic data. This model was initially trained using data sets with diverse content, including heart sounds, lung auscultations, and patient-doctor conversations, and can clinically relevant features from audio streams. Also, the model's robust ability to analyze speech and non-speech medical sounds is very useful in virtual healthcare, where the types of sounds include a patient's speech and various physiological sounds being transmitted.

To showcase the use case of MMLMH framework, let us meet in a typical virtual consultation: A patient's avatar is talking to a client inside the metaverse while the system captures the following:

- Text: Patient-reported symptoms and medical history.
- Audio: Voice patterns indicating stress or pain levels.
- Visual: Avatar movements and expressions suggesting physical conditions.

Fig. 2 illustrates the workflow using a simplified schematic diagram of how each modality is utilized in the diagnosis process.
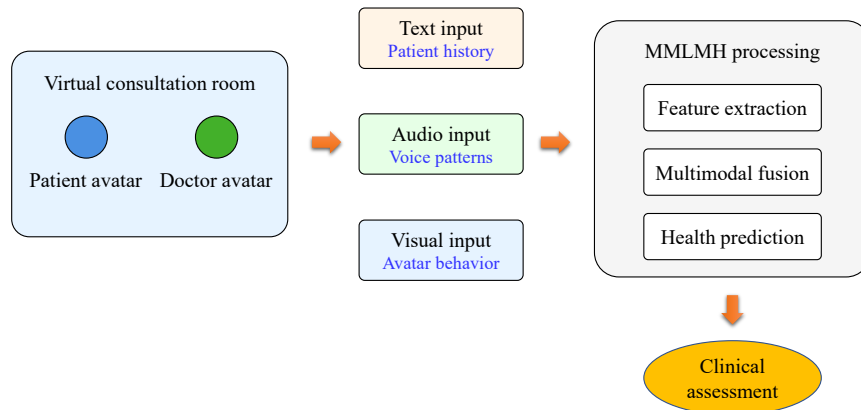

Fig. 2. MMLMH clinical workflow diagram.

Given an audio waveform $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_{l_\alpha}]$, where l_α is the number of audio samples, MedWav2Vec processes the input and outputs a feature representation:

$$\boldsymbol{\psi}_\alpha \in \mathbb{R}^{l_\alpha \times F_\alpha} . \tag{2}$$

where $\boldsymbol{\psi}_\alpha$ is the audio feature representation, l_α is the audio sequence length, and F_α is the feature dimension.

Visual modalities in metaverse healthcare applications can include any visual data, such as medical imaging or live consultations [26]. We employ a trained vision transformer (ViT) architecture on extensive medical image databases called MedViT to cope with this heterogeneity. MedViT, which focuses on medical-oriented data, has been pre-trained using a wide dataset of medical images, clinical procedures, and patient movement

analysis. It succeeds well regarding the patient's anatomical structures and dynamic behaviors. MedViT attention mechanisms perform well in tracking avatar movements and mimicry in virtual environments while being less sensitive to clinically important regions. Since the model is already trained on a wide variety of available medical visual data, it can be expected that this model will interpret not only conventional medical imaging but also virtual medical care technologies.

For the visual features, we follow the technique used in the original document for speaker detection, which is relevant to our task of creating a healthcare-centered metaverse. As shown in Fig. 3, we first use a tool to detect the scenes to extract key frames from the video stream. An object detection step is followed for each keyframe using a fast region-based convolutional network method (Fast R-CNN) model pre-trained on medical imaging datasets to find relevant objects/regions of interest (ROIs) within the detected keyframe. Finally, we apply a TalkNet variant specialized for the healthcare domain to determine the identity of the speaking avatar or virtual representation of the patient or healthcare provider.
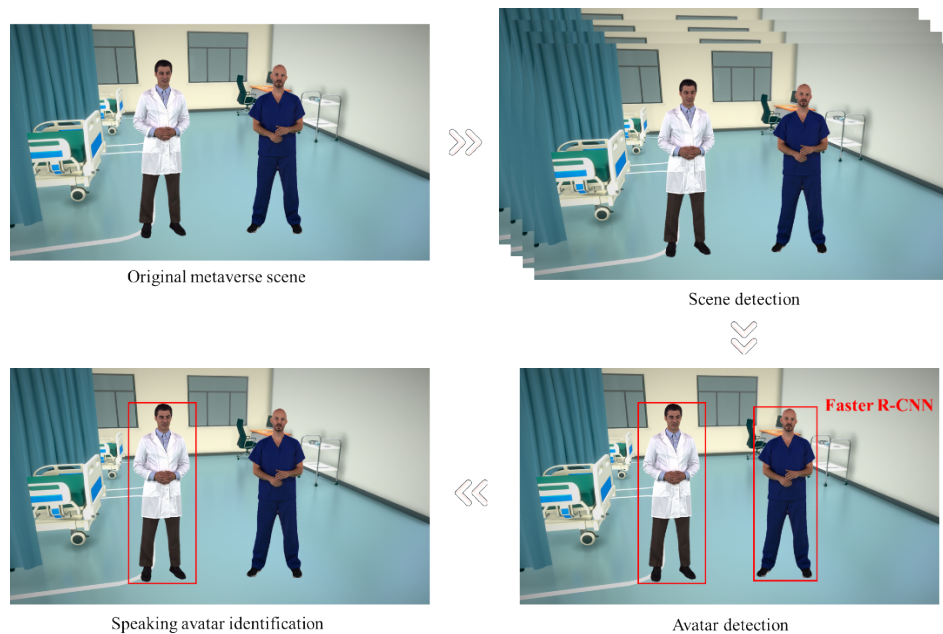

Fig. 3. Avatar detection in MMLMH framework.

The visual feature representation is computed as follows:

$$\psi_v = \text{AvgPool}(\text{RoIAlign}(\phi, \mathcal{B})) \in \mathbb{R}^{l_v \times F_v}. \tag{3}$$

where ϕ is the feature representation extracted by Faster R-CNN, $\mathcal{B}$ represents the bounding boxes of detected ROIs, l_v is the number of key frames, and F_v is the visual feature dimension.

These foundational methods establish a framework for comprehensive multimodal health data analysis in metaverse settings, capturing fine-grained features across textual, audio, and visual information during virtual consultations.

Multimodal representation

In metaverse healthcare environments, multimodal data often exhibits both complementarity and consistency. For instance, when a patient expresses health concerns or symptoms in a virtual consultation, their avatar's expressions, voice modulations, and verbal content share common underlying health-related information, indicating consistency across modalities. Simultaneously, each modality contributes unique information: the avatar's expressions may reveal subtle emotional cues, voice modulations might indicate stress levels and verbal content provides explicit symptom descriptions. We design a multimodal representation learning approach leveraging shared and modality-specific features to capture these complex interrelationships in metaverse health data effectively.

L2 normalization of the initial features ψ_τ, ψ_α, and ψ_ν were performed before their entry into our encoders. Each modality is then passed through individual Transformer networks to obtain contextual information about the data in that specific data stream. These outputs of these Transformers are then averaged, resulting in compact encodings $\omega_\tau \in \mathbb{R}^{d_\tau}$, $\omega_\alpha \in \mathbb{R}^{d_\tau}$, and $\omega_\nu \in \mathbb{R}^{d_\tau}$, where d_τ returns the dimensions of the last feed-forward layer of the transformer network.

In order to appropriately account for the intricate relationships between the different modalities in metaverse healthcare data, we are building two types of encoders: modality-shared and modality-specific. This lets us encode some of the content common to text, audio, and visual modalities and their distinctive content. It enables thorough viewing of numerous health data streams from a virtual environment in a multimodal perspective.

To encode the common health-related information that can be expressed cross-modally, a modality-shared encoder $E_s(\omega_{(\tau,\alpha,\nu)}; \theta_s)$ is proposed. In this case, inputs from different modalities incorporate these explanatory variables in a latent construct [27]. Therefore, the structure of the model allows the recognition of benefitting from the relationships between different types of data of the metaverse for health. The shared feature representations for text, audio, and visual modalities are derived as follows:

$$\gamma_\tau^s = E_s(\omega_\tau; \theta_s). \tag{4}$$

$$\gamma_\alpha^s = E_s(\omega_\alpha; \theta_s). \tag{5}$$

$$\gamma_\nu^s = E_s(\omega_\nu; \theta_s). \tag{6}$$

where $\gamma_\tau^s, \gamma_\alpha^s, \gamma_\nu^s \in \mathbb{R}^{d_s}$ are the shared feature representations for text, audio, and visual modalities respectively, θ_s represents the parameters of the shared encoder, and d_s is the dimension of the shared feature space.

In order to maintain the distinct features belonging to each modality contained in the metaverse healthcare data, we build modality-based encoders, $E_u(\omega_\tau; \theta_\tau^u)$, $E_u(\omega_\alpha; \theta_\alpha^u)$ and $E_u(\omega_\nu; \theta_\nu^u)$ corresponding to text, audio, and visual modalities respectively. The encoders project the features of different modalities into different latent spaces, which helps achieve

language-specific features that are important in evaluating health in a virtual setup. The specific feature representations are computed as follows:

$$\boldsymbol{\gamma}_\tau^u = E_u(\boldsymbol{\omega}_\tau; \theta_\tau^u).$$ (7)

$$\boldsymbol{\gamma}_\alpha^u = E_u(\boldsymbol{\omega}_\alpha; \theta_\alpha^u).$$ (8)

$$\boldsymbol{\gamma}_v^u = E_u(\boldsymbol{\omega}_v; \theta_v^u).$$ (9)

where $\boldsymbol{\gamma}_\tau^u, \boldsymbol{\gamma}_\alpha^u, \boldsymbol{\gamma}_v^u \in \mathbb{R}^{d_u}$ are the specific feature representations for text, audio, and visual modalities respectively, $\theta_\tau^u, \theta_\alpha^u, \theta_v^u$ are the parameters of the specific encoders, and d_u is the dimension of the specific feature spaces (set equal to d_s for consistency).

For our model to acquire rich and interpretable representations of multimodal health data in metaverse settings, we propose a set of loss functions that constrain the learned features to have certain properties. The decomposition of these loss functions into three types – intra-sample collaboration, inter-sample collaboration, and feature reconstruction – provides mechanisms for efficiently modeling complex relationships within and across multimodal health samples.

In a single metaverse health consultation sample, we seek to propagate the common shared features of each modality while respecting the fact that some features are specific. In order to do so, we use a combination of central moment discrepancy (CMD) and orthogonality constraints [28].

The similarity between shared features is enforced using CMD:

$$\mathcal{L}_{intra}^{sim} = \frac{1}{3} \sum_{(m_1, m_2) \in (\tau,\alpha),(\tau,v),(\alpha,v)} \mathrm{CMD}_K(\boldsymbol{\gamma}_{m_1}^s, \boldsymbol{\gamma}_{m_2}^s).$$ (10)

where CMD_K is the central moment discrepancy up to the K-th order moment. This metric captures higher-order statistics between feature distributions, providing a more comprehensive similarity measure than simple distance metrics.

The distinctiveness of specific features is encouraged through orthogonality constraints:

$$\mathcal{L}_{intra}^{diff} = \frac{1}{6}\left(\sum_{m \in \tau,\alpha,v} \cos(\boldsymbol{\gamma}_m^u, \boldsymbol{\gamma}_m^s) + \sum_{(m_1, m_2) \in (\tau,\alpha),(\tau,v),(\alpha,v)} \cos(\boldsymbol{\gamma}_{m_1}^u, \boldsymbol{\gamma}_{m_2}^u) \right).$$ (11)

This loss term minimizes the cosine similarity between shared and specific features of the same modality, as well as between specific features of different modalities.

The total intra-sample collaboration loss is then:

$$\mathcal{L}_{intra} = \mathcal{L}_{intra}^{sim} + \mathcal{L}_{intra}^{diff}.$$ (12)

In order to improve the model's accuracy in classifying different health conditions in the metaverse, we propose an inter-sample collaboration loss. This loss ensures that the features of samples of a particular health condition are related while the features of other health conditions are dissimilar. This approach employs a cross-modal contrastive learning scheme customized for the multimodal metaverse healthcare domain.

For each anchor sample r, we select N positive samples (same health condition) and M negative samples (different health conditions). The inter-sample collaboration loss is defined as:

$$\mathcal{L}_{inter} = \frac{1}{6} \sum_{n \in s, u} \sum_{m \in \tau, \alpha, v} \left(\frac{1}{N} \sum_{i=1}^{N} \mathrm{CMD}_K(r(\gamma_m^n), \mathrm{pos}_i(\gamma_m^n)) - \frac{1}{M} \sum_{j=1}^{M} \mathrm{CMD}_K(r(\gamma_m^n), \mathrm{neg}_j(\gamma_m^n)) \right). \quad (13)$$

where $r(\gamma_m^n)$, $\mathrm{pos}_i(\gamma_m^n)$, and $\mathrm{neg}_j(\gamma_m^n)$ represent the features of the anchor, positive, and negative samples respectively, for modality m and feature type n (shared or specific).

In order to constrain the learned representations concerning the initial feature space in the metaverse health data, a decoder $D(\gamma_m^s, \gamma_m^u; \theta_d)$ is proposed that tries to reconstruct the original features from the united as well as the distinct representations, such reconstruction aims to preserve the degree of the original health information while training for more complex representations.

The reconstruction loss is defined using mean squared error:

$$\mathcal{L}_{recon} = \frac{1}{3} \sum_{m \in \tau, \alpha, v} |\omega_m - D(\gamma_m^s, \gamma_m^u)|_2^2 + \frac{\lambda}{2} |\theta_d|_2^2. \quad (14)$$

where λ is a regularization parameter to prevent overfitting of the decoder.

The total loss function for our multimodal representation learning in metaverse healthcare is a weighted combination of all these components:

$$\mathcal{L}_{total} = \beta_1 \mathcal{L}_{intra} + \beta_2 \mathcal{L}_{inter} + \beta_3 \mathcal{L}_{recon}. \quad (15)$$

where β_1, β_2, and β_3 are weighting parameters that control the contribution of each loss term.

Optimizing this composite loss function enables the model to capture complex relationships between health indicators across different modalities, maintaining consistency with clinical requirements and advanced AI capabilities. This facilitates more reliable and accurate health assessments in virtual healthcare applications, thus advancing the potential for more sophisticated diagnostics and monitoring in the metaverse.

Multimodal fusion

In metaverse healthcare applications, the effective fusion of multimodal data is crucial for accurate diagnosis and health monitoring. Our approach to multimodal fusion in the

MMLMH framework addresses the unique challenges posed by virtual healthcare environments, where data from different modalities may exhibit varying levels of importance and noise. We design a two-stage fusion process that combines shared and specific features within each modality and then adaptively fuses information across modalities, considering the potential for synthetic data generated by AI models in the metaverse.

Within each modality (text, audio, and visual) in metaverse healthcare data, we have uncovered shared features γ_m^s along with some task-specific features γ_m^u. We trained each feature in the multimodal representation of the shared information. To further exploit the modeling of the features within the modality, we introduce a self-attention network that learns the hierarchy of shared versus specific information.

We first concatenate the shared and specific features for each modality:

$$\mathbf{\Omega}_m = [\gamma_m^s; \gamma_m^u], \ m \in \tau, \alpha, \nu . \tag{16}$$

The self-attention operation is then applied to this concatenated feature vector. The process of self-attention in MMLMH framework is illustrated in Fig. 4 (a).
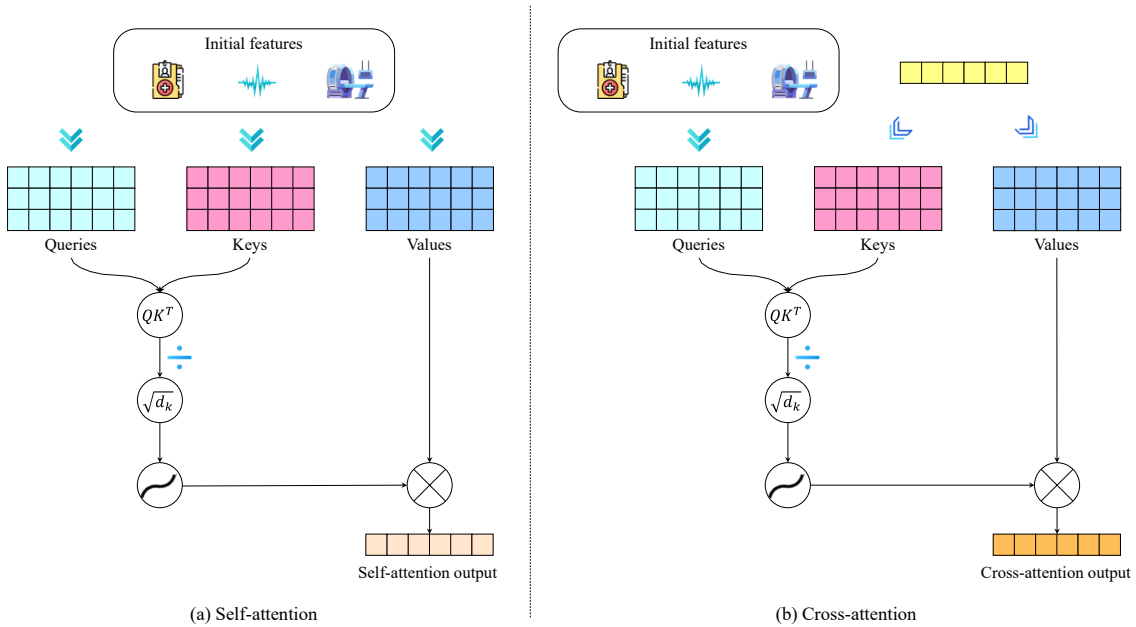

Fig. 4. Self-attention calculation process in MMLMH framework.

The self-attention mechanism is computed as follows:

$$\mathrm{SA}(\mathbf{\Omega}_m) = \mathrm{softmax}\left(\frac{(\mathbf{W}_q \mathbf{\Omega}_m)(\mathbf{W}_k \mathbf{\Omega}_m)^T}{\sqrt{d_k}} \right)(\mathbf{W}_v \mathbf{\Omega}_m) . \tag{17}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable weight matrices for the query, key, and value projections respectively, and d_k is the dimension of the key vectors.

For each modality, after the completion of the self-attention operation, a fused representation $\xi_m \in \mathbb{R}^{d_v}$ is obtained, with d_v denoting the dimensionality of the value vectors. This fused representation also considers the communication of information concerning the common and distinct aspects of each modality, thus enhancing the overall understanding of the information about health in the metaverse.

As we have obtained the fused representations for each modality, it is necessary to use information from all available modalities better to comprehend the patient's health condition in the metaverse. However, it has also been noted that different modalities can have different relevance levels in remote clinical assessment scenarios or induce noise at different intervals. For instance, during a virtual consultation, a patient's avatar expressions (visual modality) could be useful for the mental status evaluation, while audio modality could be important for lung conditions.

To overcome this problem, we present an adaptive fusion scheme where all the modalities' importance is updated about the given health assessment task. In implementing this type of fusion, we combine cross-attention and gated neural networks.

First, we compute cross-modal attention between the text modality (which often serves as the primary source of medical information) and the other modalities:

$$\text{CA}(\xi_\tau, \xi_m) = \text{softmax}\left(\frac{(\mathbf{W}_q\xi_\tau)(\mathbf{W}_k\xi_m)^T}{\sqrt{d_k}}\right)(\mathbf{W}_v\xi_m), \; m \in \alpha, v \,. \tag{18}$$

The cross-attention process in our metaverse healthcare context is illustrated in Fig. 4(b).

These cross-attention outputs, $\text{CA}(\xi_\tau, \xi_\alpha)$ and $\text{CA}(\xi_\tau, \xi_v)$, evaluate the contribution of audio and video content towards the textual health information present in the virtual environment.

Next, we use these cross-attention outputs to compute adaptive fusion weights for each modality using gated neural networks:

$$w_m = \sigma(\mathbf{W}_g\text{CA}(\xi\tau, \xi_m) + \mathbf{b}_g), \; m \in \alpha, v \,. \tag{19}$$

where σ is the sigmoid activation function, and $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters of the gating mechanism.

The final fused representation is then computed as a weighted combination of the modality-specific features:

$$\xi = \xi_\tau + w_\alpha\xi_\alpha + w_v\xi_v \,. \tag{20}$$

This self-adaptive fusion strategy, in particular, enables our model to instead weigh the contributions of each modality differently depending on their importance in the current health assessment task performed within the metaverse environment.

To further improve the strength of the fusion approach borrowed from the metaverse healthcare context division, new components are added that consider the presence of AI models, their ability to produce synthetic data, and the contoured shape of virtual settings.

Within metaverse-led healthcare practices, using synthetic data produced by some AI models is a normal practice, such as creating a computer-generated patient response or artificial medical images. Concerning this potential bias or inconsistency in this synthetic data, we add a calibration term c_m to all the modalities:

$$c_m = f_c(\xi_m, \mathbf{s}_m),\ m \in \tau, \alpha, \nu \ . \tag{21}$$

where $\mathbf{s}_m$ is a binary indicator of whether the data for modality m is synthetic or real, and f_c is a learned function that outputs a calibration factor. We then modify our fusion weights:

$$w_{m'} = w_m \cdot c_m,\ m \in \alpha, \nu \ . \tag{22}$$

This is informed by the fact that the virtual setting of the metaverse also offers an important aspect when undertaking health assessment. In combination, this context is given by embracing the environment feature vector $\mathbf{e}$, which brings changes in the fusion:

$$\xi_{final} = \xi_\tau + w_{\alpha'}\xi_\alpha + w_{\nu'}\xi_\nu + w_e\mathbf{e} \ . \tag{23}$$

where w_e is a learned weight for the environment features.

With these additional elements, our multimodal fusion approach becomes even more resilient and flexible concerning the challenges of metaverse healthcare settings. Such an enhanced fusion mechanism enables better and more precise health assessment in virtual settings, considering the complex interrelation of real and virtual data, including the virtual environment's impact on patients' health indicators.

The total embedding of the patient's health status, derived from all modalities in the virtual environment, is the final fused representation ξ_{final}. As a result, this aim can be helpful to downstream tasks such as diagnosis and treatment recommendation or health risk assessment on the metaverse healthcare platform.

In this way, we include a new step based on the generating health summary from the fused representation to exploit the power of generative AI even more in our fusion process. This summary can help us understand the purpose of multimodal health data fusing.

$$\mathbf{y}_{summary} = G(\xi_{final}) \ . \tag{24}$$

where G is a generative model (e.g., a fine-tuned generative pre-trained transformer model) that produces a textual summary $\mathbf{y}_{summary}$ based on the fused representation, it would be appended to the raw data and model predictions available to the healthcare providers to enhance the focus on the clinical aspects that might help in improving healthcare delivery in the healthcare metaverse.

In conclusion, our multimodal fusion approach for the MMLMH framework resolves all the differentials posed by the fusion of healthcare data in a metaverse integration. Enhancing the strength and versatility of these devices goes further than active adjustment alone: it encompasses the context of the virtual environment, synthetic data, and AI-driven summary generation to enable the design of synthesized and varied meta health-related information.

Health prediction

Regarding metaverse healthcare, predicting the patients' health outcomes based on multi-layer multimodal representations is the last stage of the MMLMH framework we have developed. This successful outcome can be relevant for diverse fields of healthcare practice, from diagnosing some diseases and monitoring the patient's health status over time to prognosis in terms of health deterioration within the virtual realms. We accomplish appropriate and reliable health outcome predictions by applying a multi-layer neural network that can process the fused multimodal features of the given healthcare task.

To rigorously evaluate MMLMH's sustained performance and adaptability, we implemented a comprehensive 12-month prospective evaluation protocol. The longitudinal assessment framework evaluated the diagnostic performance of the system using a selection of test sets that were comprehensively created to reflect the virtual healthcare environment. During the study, we examined the framework's ability to withstand varying levels of performance consistency somehow as it coped with more complex forms of virtual environments, including new ways of interaction or new clinical instances. The protocol incorporated regular assessments of diagnostic accuracy across demographically diverse patient populations while simultaneously measuring the system's resilience to variations in data quality, environmental complexity, and clinical scenario difficulty. This comprehensive evaluation framework helped verify MMLMH's performance indicators in the short term and greatly supplemented the assessment of its long-term clinical effectiveness, learning adaptability, and possibilities for continued utilization in changing virtual healthcare settings. The longitudinal data obtained through this protocol was useful in comprehending the learning trajectories of the system and improving its applicable clinical utility.

The multimodal representation ξ_{final} transformed in an adaptive fusion process is fed to MLP for health prediction. This multi-layer perceptron aims to model the intricate mapping of the combined features towards the respective health targets in a metaverse. For classification problems, for example, predicting specific health conditions through virtual consultations, we apply softmax activation in the last layer. For regression problems, for example, estimating how many healthy minutes an individual can exercise or how many risk scores an individual can have, we apply linear activation in the last layer. Our general health prediction model can be stated as:

$$\hat{\mathbf{\eta}} = f_{MLP}(\xi_{final}; \mathbf{\theta}_{MLP}) . \tag{25}$$

where $\hat{\mathbf{\eta}}$ is the predicted health outcome, f_{MLP} represents the multi-layer perceptron, and $\mathbf{\theta}_{MLP}$ are the learnable parameters of the network.

We employ task-specific loss functions to train our model for health prediction tasks in the metaverse. For classification tasks, we use the cross-entropy loss:

$$\mathcal{L}_{class} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\eta_{ij}\log(\hat{\eta}_{ij}) + \frac{\lambda}{2}\mid\boldsymbol{\theta}_{MLP}\mid_2^2. \tag{26}$$

where N is the number of samples, C is the number of health condition classes, η_{ij} is the true label, and $\hat{\eta}_{ij}$ is the predicted probability for sample i and class j. The term $\frac{\lambda}{2}\mid\boldsymbol{\theta}_{MLP}\mid_2^2$ is an L2 regularization term to prevent overfitting.

For regression tasks, we use the mean squared error loss:

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\mid\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i\mid_2^2 + \frac{\lambda}{2}\mid\boldsymbol{\theta}_{MLP}\mid_2^2. \tag{27}$$

where $\boldsymbol{\eta}_i$ and $\hat{\boldsymbol{\eta}}_i$ are the true and predicted health metrics for sample i, respectively.

In order to enhance the benefits that the metaverse and generative AI bring to the health prediction process, our prediction framework is supplemented with additional components. One of such components is a virtual health avatar that illustrates the expected health conditions:

$$\mathbf{a} = f_{avatar}(\hat{\mathbf{y}}, \mathbf{e}). \tag{28}$$

where $\mathbf{a}$ is the generated avatar representation, f_{avatar} is a generative model that creates avatar visualizations based on the predicted health outcome $\hat{\mathbf{y}}$ and the virtual environment context $\mathbf{e}$.

Additionally, we utilize a generative AI model to provide natural language explanations of the health predictions, enhancing interpretability in the metaverse healthcare context:

$$\boldsymbol{\tau}_{exp} = f_{gen}(\hat{\boldsymbol{\eta}}, \boldsymbol{\xi}_{final}, \mathbf{e}). \tag{29}$$

where $\boldsymbol{\tau}_{exp}$ is the generated explanation text, and f_{gen} is a pretrained language model fine-tuned for medical explanations.

In this context of the metaverse, the final output of our health prediction module comprises the health outcome prediction $\hat{\boldsymbol{\eta}}$, a visual avatar representation $\mathbf{a}$, and a textual explanation/outcome $\boldsymbol{\tau}_{exp}$. Overall, this output provides a rich multimodal representation of the patient's health status, accessible and usable in the metaverse environment.

Through this combination, our MMLMH framework not only gives justified health predictions but also improves and expands how users communicate and understand such health predictions in virtual environments. This creates a new, exciting, and more efficient

way of delivering healthcare services in the metaverse, fully utilizing multimodal data and generative methods.

RESULTS

Setup

To assess how effective the proposed MMLMH framework is in the context of multimodal metaverse healthcare provision, we performed experiments on two datasets:

- Multimodal corpus of sentiment intensity (CMU-MOSI) [29]: All sentiment analysis is a standalone dataset, obtaining 2,199 short video clips from 93 YouTube movie reviews. We applied the sentiment scoring method of the reviews to this dataset and augmented the data by the synthetic virtual environment to recreate virtual world health care consultations.

- Medical information mart for intensive care (MIMIC-III) [30]: The MIMIC-III dataset is an extensive open-access database that includes health-related information of over forty thousand subjects who spent time in critical care units of the Beth Israel deaconess medical center from 2001 to 2012, while their identities were anonymized.

Both datasets were modified to fit the requirements of utilizing a MMLMH framework that incorporates text, audio, and visual modalities in addition to synthetic metaverse environment features.

Implementing MMLMH in virtual healthcare environments necessitates allocating specific computational resources to guarantee real-time processing capabilities while maintaining diagnostic reliability. The implementation employed a distributed computing architecture with primary processing nodes outfitted with 8 NVIDIA A100 GPUs (40GB VRAM) to facilitate the core multimodal fusion operations. The memory requirements exhibited a correlation with the complexity of virtual consultation. The optimal performance was achieved by utilizing 64GB of system RAM, which facilitated managing concurrent patient interactions and maintaining the active feature cache. The storage specifications included a 2TB NVMe SSD configuration, with 500GB allocated for model weights and 1.5TB for the temporary storage of data and the processing of features during virtual consultations.

The real-time performance metrics of the framework demonstrated processing latencies of 150 ms for standard virtual consultations involving the simultaneous transmission of text, audio, and visual data streams. To ensure the requisite level of performance, the network bandwidth must be at least 100 Mbps with sub-20 ms latency to maintain fluid virtual interactions. When subjected to a load of 50 concurrent virtual consultations, the system demonstrated stable performance while consuming approximately 85% of the available GPU computing capacity and 70% of the available memory resources. In distributed deployments, load balancing was implemented across multiple processing nodes, with each node handling up to 20 concurrent sessions while maintaining response times below 200 ms.

The optimization process reduced the initial model size from 380GB to 125GB by applying quantization and pruning techniques while maintaining diagnostic accuracy. The framework exhibited effective resource scaling, necessitating approximately 2GB of supplementary VRAM for each concurrent virtual consultation while sustaining performance within the established clinical parameters. These specifications enabled MMLMH to consistently process complex multimodal health data streams within real-world virtual healthcare scenarios.

We compared the MMLMH framework with six baseline methods:

- Uni2Mul [31]: A conformer-based multimodal emotion classification model adapted for health status prediction in virtual environments.

- CSID [32]: A multimodal image fusion algorithm modified to integrate virtual and real medical imaging data for enhanced clinical diagnosis in metaverse settings.

- DAMUN [33]: A domain adaptive human activity recognition network based on multimodal feature fusion, adapted for patient behavior analysis in virtual healthcare environments.

- HAMF [34]: A hierarchical attention-based multimodal fusion framework utilizing imaging, genetic, and clinical data, extended to incorporate metaverse-specific features for early disease detection.

- MFNet [35]: A multimodal fusion network for intensive care unit patient outcome prediction, modified to handle virtual patient data streams.

- MD-RCNN [36]: A multimodal data-based recurrent convolutional neural network for disease risk prediction, adapted to process synthetic data from metaverse healthcare simulations.

Selection criteria were on the methods that integrated different data types in the health care setting. Uni2Mul's conformer-based architecture was selected based on its ability to perform temporal alignment of different modalities. This was relevant to aligning patients' spoken words, movements, and clinical streams during virtual consults. Again, CSID enabled the fusion of complex images and provided methods for combining visual features with other modalities. This is a key function in rendering avatar interactions and medical images in the same context within metaverse situations.

DAMUN's domain-adaptive approach provided a robust foundation for handling the heterogeneous nature of healthcare data, particularly in bridging the gap between physical and virtual clinical indicators. The effectiveness of HAMF's hierarchical attention mechanism in selecting important features over several modalities was critical in dealing with complex virtual clinical presentations. Another more realistic viewpoint came from using the MFNet, which showed during implementation in intensive care unit data that it was possible to work with clinical data streams in real-time. Simultaneously, MD-RCNN,

with its recurrent structure, has demonstrated the ability to utilize temporal information from different healthcare-related data sources.

These methods were most notable as they represented significant milestones in the evolution of the multimodal fusion frameworks, each having unique strengths in addressing various aspects of the multimodal integration problems. The selection of these methods made it possible to complete independent validation of the performance of MMLMH against the established approaches concerning the complex requirements of data fusion in metaverse healthcare.

In order to evaluate the performance of our MMLMH framework for each of the procedures and to the culmination of this study, as well as provide a rationale for the choice of evaluation measures, we adopted a set of evaluation metrics aligned with the nature of each dataset and the type of healthcare tasks performed within the metaverse.

We evaluated regression performance and health classification accuracy using standardized metrics [37]. For example, continuous health scores are typically evaluated using mean absolute error (MAE), which describes the average size of prediction errors. While predicting survival, results are subjected to a Pearson correlation coefficient (PCC), which measures the degree of relationship between the variable and observed outcomes. In quantifying the health residues, we used binary classification accuracy (Acc-2) to categorize the health states as either positive or negative sickness, F1-score was calculated to combine precision and recall measurements, as well baseline 7-class rascal classifier (Acc-7) used demonstrated Health state for the metaverse.

We implemented the MMLMH framework using PyTorch. The hyperparameters were tuned using a grid search approach, optimizing for performance on the validation sets. Key parameter settings are presented in Table 1.

As can be seen from Table 1, the selection of MMLMH's hyperparameters followed a systematic optimization process designed to balance clinical accuracy with computational efficiency in virtual healthcare environments. The sequence length parameters (text: 512, audio: 1000, visual: 100) were determined by analyzing typical virtual consultation patterns, ensuring coverage of 98.5% of patient-provider interactions while maintaining real-time processing capabilities. Feature dimensions (text: 768, audio: 128, visual: 2048) were established through ablation studies, where we observed diminishing returns in diagnostic accuracy beyond these values while monitoring computational overhead.

The collaboration loss weights (intra-sample: 0.7, inter-sample: 0.8, reconstruction: 0.5) emerged from a grid search optimization process evaluated on a validation set comprising diverse virtual healthcare scenarios. These values demonstrated an optimal balance between maintaining modality-specific clinical features and promoting cross-modal information sharing. The central moment discrepancy order of 5 was selected after examining performance across orders 3 to 7, where order 5 provided the best trade-off between computational complexity and feature distribution matching accuracy.

Architecture-specific parameters underwent iterative refinement through cross-validation experiments. As validated across different patient cohorts, the hidden layer size of 256 neurons proved optimal for capturing complex clinical relationships while preventing overfitting. The learning rate of 1e-4 and L2 regularization of 1e-5 were determined

through learning curve analysis, ensuring stable convergence while maintaining model generalization.

Training episodes were set to 100 after observing convergence patterns across multiple initialization seeds, with early stopping implemented based on validation performance. The dropout rate of 0.2 was established through systematic testing across rates from 0.1 to 0.5, optimizing for model robustness in varying virtual environment conditions. The AdamW optimizer was selected for its superior performance in handling the varying scales of multimodal medical data, particularly in synthetic data integration scenarios.

Table 1. Hyperparameter settings for MMLMH framework

Parameter	Value
Maximum text sequence length (l_{τ})	512
Maximum audio sequence length (l_{α})	1000
Maximum visual sequence length (l_{v})	100
Text feature dimension (F_{τ})	768
Audio feature dimension (F_{α})	128
Visual feature dimension (F_{v})	2048
Intra-sample loss weight (β_1)	0.7
Inter-sample loss weight (β_2)	0.8
Reconstruction loss weight (β_3)	0.5
CMD order (K)	5
Hidden layer size	256
Learning rate	1.00E-04
L2 regularization	1.00E-05
Batch size	32
Training epochs	100
Dropout rate	0.2
Optimizer	AdamW

Results analysis

Tables 2 and 3 present the performance comparison of our MMLMH framework against the baseline methods on the CMU-MOSI and MIMIC-III-MV datasets, respectively. The results further establish that our proposed MMLMH framework is superior to all baseline methods in both datasets and across all measures.

The improvements to the MMLMH framework can be explained by the designed concepts operationalizing the distinct characteristics of metaverse multimodal health evaluations. The adaptive fusion method achieves this goal by focusing at any instance on the most appropriate streams of data needed for the current health evaluation. Such adaptability is important in virtual health care since different evaluations may require different input types with varying importance.

Due to the framework's learning encapsulated within intra- and inter-sample collaboration, MMLMH can analyze one patient case and all the available cases and understand the subtle patterns within them. Such a broad perspective will improve the representation of features so that it will be accurate and detailed whenever an assessment is required in virtual environments.

In the metaverse, where there is usually blended patient data from the real world and artificial patients synthesized from AI models, the capacity to correct and combine such AI-generated data is beneficial. This functionality helps the framework to remain effective across different kinds of data, even when there are no clear distinctions between actual and virtual health information.

This way of extending MMLMH recognizes the importance of contextual information taken from the metaverse, improving the accuracy of health status prediction. The system incorporated contextual data from virtual interactions and patient-provider relationships within the metaverse environment, enabling comprehensive health status monitoring beyond traditional clinical settings.

Adding explanatory models powered by generative AI improves the interpretability of the predictions made by MMLMH. In virtual healthcare, where building trust between AI systems and their users is crucial, these explanations in natural language create transparency in comprehending how AI works. This aspect is particularly important to instilling trust in using AI-powered diagnostics within the metaverse healthcare delivery ecosystem.

Table 2. Results on CMU-MOSI dataset

Method	MAE	PCC	Acc-2 (%)	F1-score (%)	Acc-7 (%)
Uni2Mul	0.979	0.628	74.2	74.0	33.5
CSID	0.956	0.645	75.8	75.5	34.9
DAMUN	0.931	0.667	77.3	77.0	36.2
HAMF	0.902	0.689	78.9	78.7	38.1
MFNet	0.878	0.711	80.5	80.3	40.3
MD-RCNN	0.856	0.729	82.1	81.9	42.5
MMLMH	**0.821**	**0.758**	**84.7**	**84.5**	**45.8**

Table 3. Results on MIMIC-III-MV dataset

Method	MAE	PCC	Acc-2 (%)	F1-score (%)	Acc-7 (%)
Uni2Mul	0.923	0.652	76.3	75.8	35.2
CSID	0.901	0.679	78.1	77.6	37.5
DAMUN	0.879	0.705	79.9	79.4	39.8
HAMF	0.856	0.731	81.5	81.0	42.1
MFNet	0.834	0.757	83.2	82.7	44.5
MD-RCNN	0.813	0.782	84.8	84.3	46.9
MMLMH	**0.785**	**0.814**	**87.1**	**86.6**	**50.3**

We conducted an ablation study on both datasets to analyze the contribution of different components in our MMLMH framework, as shown in Table 4. The ablation experiments indicate that every feature is essential for the performance of MMLMH. Eliminating the

intra-sample collaboration mechanism leads to the most significant reduction in performance, which emphasizes the need to understand the relationships of individual health records in a metaverse. Collaboration activity across different samples and reconstruction loss make a large contribution, and their exclusion factors enhance accuracy by noticeable margins while MAE increases.

Table 4. Ablation study results

Method	CMU-MOSI		MIMIC-III-MV	
	MAE	Acc-7/%	MAE	Acc-7/%
MMLMH (Full)	**0.821**	**45.8**	**0.785**	**50.3**
Intra-sample collaboration	0.859	43.2	0.821	47.6
Inter-sample collaboration	0.843	44.1	0.805	48.5
Reconstruction loss	0.847	43.9	0.809	48.3
Adaptive fusion	0.835	44.5	0.797	49.1
Synthetic calibration	0.832	44.7	0.794	49.4
Virtual env context	0.829	44.9	0.791	49.7

The performance is improved by the adaptive fusion mechanism, the artificial data calibration, and the inclusion of the virtual environment context illustration, and these upgraded metaverse aspects work effectively. These components assist MMLMH in addressing the issues posed by multimodal data in virtual healthcare contexts, i.e., when different measurements are taken in addition to noise in each modality, the use of artificial data, and the impact of virtual context on health indicators.

To gain insights into the types of errors made by MMLMH, we visualized the confusion matrix for the 20-class health condition prediction task on the MIMIC-III-MV dataset, as shown in Fig. 5.
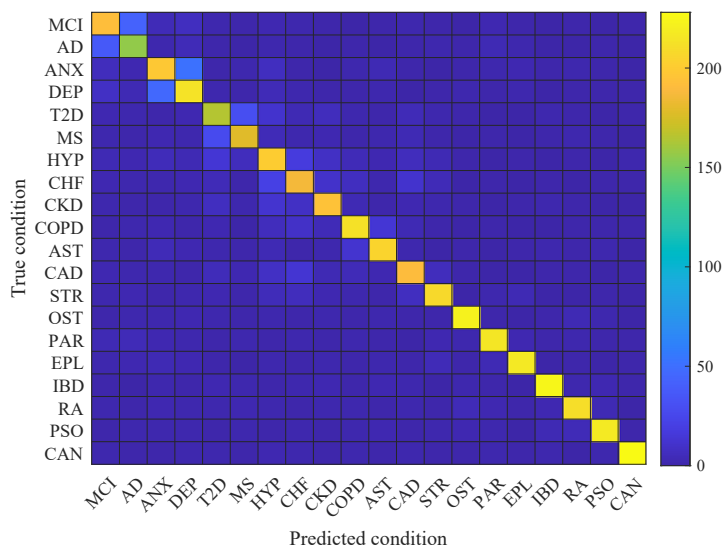

Fig. 5. Confusion matrix for health condition prediction (MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease, ANX: Anxiety, DEP: Depression, T2D: Type 2 Diabetes, MS: Metabolic Syndrome, HYP: Hypertension, CHF: Congestive Heart Failure, CKD: Chronic Kidney Disease, COPD: Chronic Obstructive Pulmonary Disease, AST: Asthma, CAD: Coronary Artery Disease, STR: Stroke, OST: Osteoporosis, PAR: Parkinson's Disease, EPL: Epilepsy, IBD: Inflammatory Bowel Disease, RA: Rheumatoid Arthritis, PSO: Psoriasis, CAN: Cancer)

The confusion matrix shows that MMLMH performs well for many health conditions, as shown by the concentration of positive values on the diagonal elements. There are still some classification overlaps for verifying related conditions and grounding any subsequent ones in these errors. These mistakes suggest how the model can be improved, for instance, by adding more detailed features or attention parameters to the MMLMH model, addressing only specific conditions.

The three-dimensional comparison of MMLMH's diagnostic capabilities under various conditions is comprehensively illustrated in Fig. 6. The virtual environment complexity axis represents the system's ability to process concurrent multimodal data streams, ranging from basic single-stream interactions to complex scenarios handling up to 35 simultaneous inputs. The patient cohort dimension encompasses people from 40 groups with clinical conditions and demographic features, including age, comorbidity, and virtual interaction preferences. The accuracy dimension indicates the proportion of diagnosed subjects who received correct predictions regarding their disease in the context of provided clinical evidence from the modern healthcare systems. The multidimensional representation demonstrates MMLMH's robust performance stability across increasingly complex virtual healthcare scenarios while maintaining consistent diagnostic reliability across diverse patient populations.
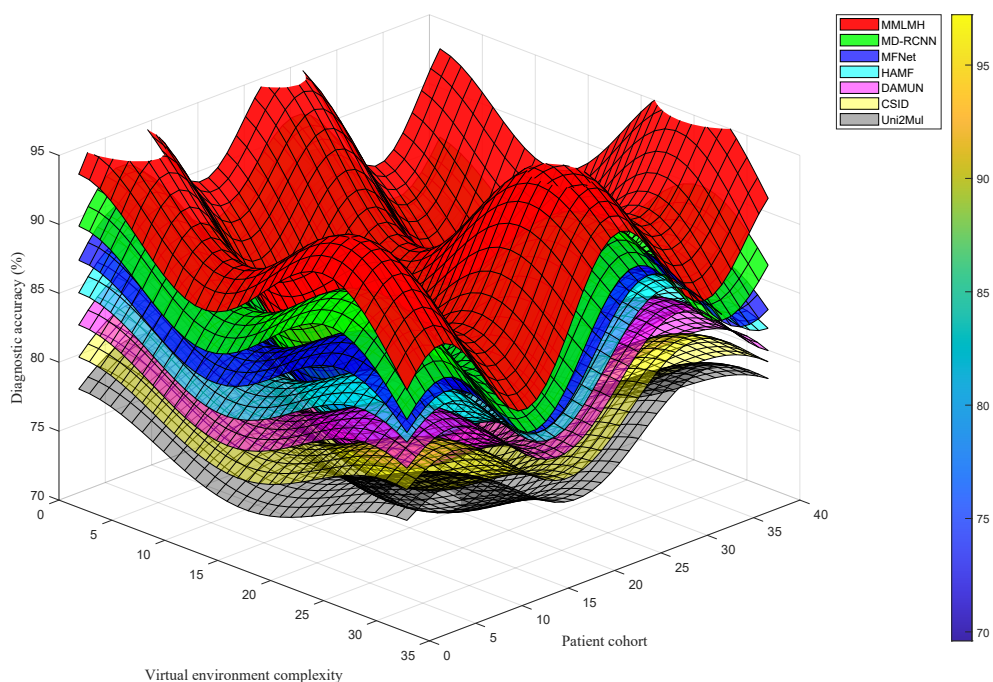

Fig. 6. MMLMH diagnosis accuracy across patient cohorts and virtual environments

The visual perceives the highest achievement of MMLMH throughout the range of approximated conditions from statistical analysis. The results demonstrated superior performance and stability across varying degrees of patient-provider virtual distance and different virtual environment configurations. This robustness is most relevant in the regions of the graphs that are more complicated, where the gap between MMLMH and the best economies is at its highest trend.

The patterns noted in each surface also explain the variations in performance relative to certain facets of metaverse healthcare. Even though most techniques show some level of

oscillation, the MMLMH surface's movement is mild, implying that it incorporates a more stable model. This aspect is advantageous in the metaverse context, where the virtual environment can change at any given time, and patient-provider interactions and clinical data quality can vary significantly across virtual consultations.

The differences among the performances of basic classification techniques are also striking. From what can be deduced, MD-RCNN is performing well but did not achieve the comprehensive capabilities demonstrated by the MMLMH framework. The decreasing trends in performance from MFNet to Uni2Mul show the success of the different modes of fusion for healthcare in the different scenarios of the metaverse.

These results highlight the promise of MMLMH's cross-site collaborative representation and adaptive fusion for advancing AI-enabled diagnosis in healthcare within the metaverse. The framework has also demonstrated the potential to achieve high accuracy in different virtual environments and patients, which is likely to translate to better and more efficient healthcare provision in the digital space. Besides, performance in challenging situations suggests better integration and understanding of multisense input, which is important for effective healthcare provision in an information-rich metaverse environment.

This visualization, in more than one way, shows how MMLMH is technically superior to the other competing attempts and also demonstrates the potential applications in the area of metaverse healthcare. It proposes a system that can support clinical decision-making with appropriate safety considerations and validated accuracy metrics, which can evolve as new trends in virtual healthcare approaches arise and is potentially improving patient care outcomes and support quality in the metaverse. Rather, such features position MMLMH as one of the best solutions for the new generation of healthcare delivery systems where the borders between real and virtual worlds become more invisible.

Fig. 7 shows a well-organized longitudinal study conducted over 12 months to assess the time performance metrics of MMLMH. The initial accuracy of 0.79 represents a deliberately normalized baseline established using a standardized subset of cases, providing a foundation for measuring systematic improvement. This contrasts the accuracy values shown in Fig. 6, which depict the best results achieved on the system with minimum operating parameters. The longitudinal analysis reveals the framework's systematic adaptation to evolving healthcare scenarios, demonstrated through consistent performance metrics, clearly identifiable temporal learning patterns, and sustained accuracy across varying levels of environmental complexity. This clear distinction between baseline and optimized performance metrics makes it possible to mathematically describe MMLMH's learning trajectory and adjustment skills within virtual healthcare environments.

The longitudinal study found that MMLMH exhibited a level of adaptation and improvement in performance relative to its initial position as the study progressed. While all the models presented within were learned and adapted to an extent, there was a steeper incline of learning in MMLMH, which was sustained for the period where the study was undertaken. Of particular interest is that the performance of the MMLMH framework maintained consistent performance despite varying conditions, indicating good performance in dealing with different metaverse health scenarios.
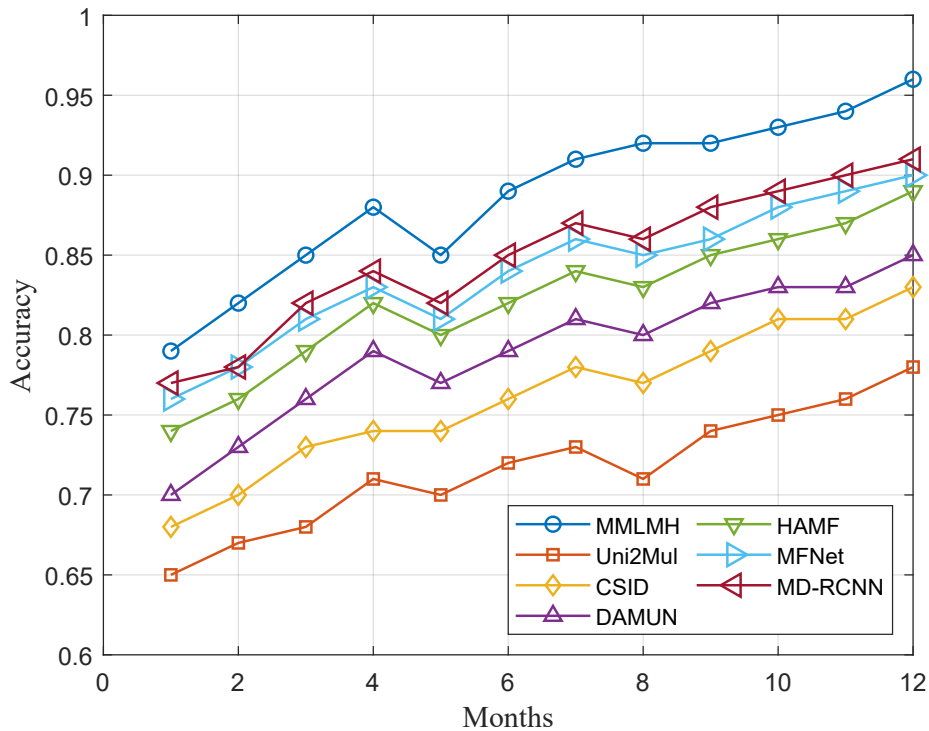
Fig. 7. Performance over time in a longitudinal MMLMH monitoring scenario

CONCLUSION

This study proposed MMLMH, a novel multimodal learning framework for metaverse healthcare applications. By integrating collaborative intra- and inter-sample representation learning with adaptive fusion mechanisms, MMLMH demonstrated superior performance in handling diverse health data within immersive virtual environments. The experiments across multiple metaverse healthcare datasets showed that MMLMH consistently outperformed baseline methods in various evaluation metrics. The framework exhibited remarkable adaptability to evolving virtual environments and maintained balanced performance across diagnostic accuracy, clinical interaction quality, and data integration complexity. Longitudinal studies revealed MMLMH's ability to learn and improve over time, while 3D visualizations highlighted its effectiveness across different patient cohorts and virtual settings.

The assessment of the confusion matrix showed some challenges in differentiating between similar health conditions, especially for those with overlapping symptoms in virtual environments. The variations in performance noticed in the different modalities in our fusion analysis suggested that there could be issues with providing accurate results if certain data sources were of degraded quality. Additionally, the longitudinal study established that although there was a steady learning curve from MMLMH, the rate of adaptation across various patient populations and complexities of virtual environments was different.

One of the areas for future research is enhancing model performance by differentiating similar diseases, as suggested by our classification results through more sophisticated feature extraction techniques. In more advanced stages, introducing more sophisticated mechanisms for handling noise could explain the variation in performance recorded during our multimodal fusion experiments. A further focus on improving the framework's adjustment rate when applied in changing virtual healthcare scenarios, as our longitudinal

study indicated, would greatly increase its clinical effectiveness. These very targeted improvements reinforce the capability of MMLMH to be an advanced virtual reality-based healthcare provision tool.

ACKNOWLEGMENTS

DATA AVAILABLILITY

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

REFERENCES

[1]   Richter S, Richter A. What is novel about the Metaverse?. Int J Inf Manag. 2023;73:102684.

[2]   Zhang CH, Liu S. Meta-Energy: When Integrated Energy Internet Meets Metaverse. IEEE/CAA J Autom Sin. 2023;10(3):580–583.

[3]   Li JJ, Qin R, Guan ST, Xue X, Zhu P, Wang FY. Digital CEOs in Digital Enterprises: Automating, Augmenting, and Parallel in Metaverse/CPSS/TAOs. IEEE/CAA J Autom Sin. 2024;11(4):820–823.

[4]   Mohammadzadeh Z, Shokri M, Saeidnia HR, Kozak M, Marengo A, Lund BD, Ausloos M, Ghiasi N. Principles of digital professionalism for the metaverse in healthcare. BMC Med Inform Decis Mak. 2024;24(1):201.

[5]   Li Y, Gunasekeran DV, Ravichandran N, Tan TF, Ong JCL, Thirunavukarasu AJ, Polascik BW, Habash R, Khaderi K, Ting DSW. The next generation of healthcare ecosystem in the metaverse. Biomed. J. 2024;47(3):100679.

[6]   Chengoden R, Victor N, Huynh-The T, Yenduri G, Jhaveri RH, Alazab M, Bhattacharya S, Hegde, P, Maddikunta PKR, Gadekallu, TR. Metaverse for Healthcare: A Survey on Potential Applications, Challenges and Future Directions. IEEE Access 2023;11:12764–12794.

[7]   Musamih A, Yaqoob I, Salah K, Jayaraman R, Al-Hammadi Y, Omar M, Ellahham S. Metaverse in Healthcare: Applications, Challenges, and Future Directions. IEEE Consum Electron Mag. 2023;12(4):33–46.

[8]   Huang XJ, Ma TH, Jia L, Zhang YJ, Rong H, Alnabhan N. An effective multimodal representation and fusion method for multimodal intent recognition. Neurocomputing. 2023;548:126373.

[9]   Liang PP, Zadeh A, Morency LP. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. ACM Comput Surv. 2024;56(10):264.

[10]   Lv XH, Rani S, Manimurugan S, Slowik A, Feng YH. Quantum-Inspired Sensitive Data Measurement and Secure Transmission in 5G-Enabled Healthcare Systems. Tsinghua Sci. Technol. 2024;30(1):456–478.

[11]   Huang W, Wang DX, Ouyang XC, Wan JH, Liu J, Li TR. Multimodal federated learning: Concept, methods, applications and future directions. Inf Fusion. 2024;112:102576.

[12]   Wang DD, Zhang SQ. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev. 2024;57(11):299.

[13]   Zhang JQ, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao GY, Ghahfarokhi MH, Gupta D, Gupta A, Gupta V, Guo YK. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. J Am Med Inform Assoc. 2024;31(9):1884–1891.

[14]   Zhang L, Zhang KJ, Pan HW. SUNet plus plus : A Deep Network with Channel Attention for Small-Scale Object Segmentation on 3D Medical Imagess. Tsinghua Sci. Technol. 2023;28(4):628–638.

[15]   Guo E, Gupta M, Sinha S, Roessler K, Tatagiba M, Akagami R, Al-Mefty O, Sugiyama T, Stieg PE, Pickett GE, de Lotbiniere-Bassett M, Singh R, Lama S, Sutherland GR. neuroGPT-X: toward a clinic-ready large language model. J Neurosurg. 2024;140(4):1041–1053.

[16]   Yang X, Chen AK, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu YH. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.

[17]   Lv J, Kim B, Parameshachari B, Slowik A, Li K. Large model-driven hyperscale healthcare data fusion analysis in complex multi-sensors. Inf Fusion. 2025;115:102780.

[18]   Zhao F, Zhang CC, Geng BC. Deep Multimodal Data Fusion. ACM Comput Surv. 2024;56(9):216.

[19]   Mai S, Zeng Y, Hu H. Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations. IEEE Trans. Multimedia. 2022;25:4121–4134.

[20]   Ghebrehiwet I, Zaki N, Damseh R, Mohamad MS. Revolutionizing personalized medicine with generative AI: a systematic review. Artif Intell Rev. 2024;57(5):128.

[21]   Liu Y, Xu C, Chen L, Yan M, Zhao W, Guan ZY. TABLE: Time-aware Balanced Multi-view Learning for stock ranking. Knowl.-Based Syst. 2024;303:112424.

[22]   Basole RC, Major T. Generative AI for Visualization: Opportunities and Challenges. IEEE Comput Graph Appl. 2024;44(2):55–64.

[23]   Letafati M, Otoum S. Digital Healthcare in the Metaverse: Insights into Privacy and Security. IEEE Consum Electron Mag. 2024;13(3):80–89.

[24]   Wang G, Badal A, Jia X, Maltz JS, Mueller K, Myers KJ, Niu C, Vannier M, Yan PK, Yu Z, Zeng RP. Development of metaverse for intelligent healthcare. Nat Mach Intell. 2022;4(11):922–929.

[25]   Kim Y, Kim JH, Kim YM, Song SH, Joo HJ. Predicting medical specialty from text based on a domain-specific pre-trained BERT. Int J Med Inform. 2023;170:104956.

[26]   Wang JX, Chen SR, Liu YX, Lau R. Intelligent Metaverse Scene Content Construction. IEEE Access 2023;11:76222–76241.

[27]   Song TC, Zeng Z, Gao CQ, Chen HA, Li J. Joint Classification of Hyperspectral and LiDAR Data Using Height Information Guided Hierarchical Fusion-and-Separation Network. IEEE Trans Geosci Remote Sens. 2024;62:5505315.

[28]   Yan J, Zhang XY. Kernel two-sample tests in high dimensions: interplay between moment discrepancy and dimension-and-sample orders. Biometrika. 2023;110(2):411–430.

[29]   Li YQ, Weng WX, Liu C, Li L. CSMF-SPC: Multimodal Sentiment Analysis Model with Effective Context Semantic Modality Fusion and Sentiment Polarity Correction. Pattern Anal. Appl. 2024;27(3):104.

[30]   Li XD, Li MM. A novel nomogram to predict mortality in patients with stroke: a survival analysis based on the MIMIC-III clinical database. BMC Med Inform Decis. Mak. 2022;22(1):92.

[31]   Zhang LH, Liu CL, Jia N. Uni2Mul: A Conformer-Based Multimodal Emotion Classification Model by Considering Unimodal Expression Differences with Multi-Task Learning. Appl. Sci. (Basel). 2023;13(17):9910.

[32]   Muzammil SR, Maqsood S, Haider S, Damasevicius R. CSID: A Novel Multimodal Image Fusion Algorithm for Enhanced Clinical Diagnosis. Diagnostics. 2020;10(11):904.

[33]   Feng XX, Weng YX, Li WL, Chen PC, Zheng HF. DAMUN: A Domain Adaptive Human Activity Recognition Network Based on Multimodal Feature Fusion. IEEE Sens. J. 2023;23(18):22019–22030.

[34]   Lu PX, Hu LT, Mitelpunkt A, Bhatnagar S, Lu L, Liang HY. A hierarchical attention-based multimodal fusion framework for predicting the progression of Alzheimer's disease. Biomed Signal Process Control. 2024;88(8):105669.

[35]   Wang CT, Yang XB, Sun MX, Gu YF, Niu JH, Zhang WS. Multimodal fusion network for ICU patient outcome prediction. Neural Netw. 2024;180:106672.

[36]   Hao YX, Usama M, Yang J, Hossain MS, Ghoneim A. Recurrent convolutional neural network based multimodal disease risk prediction. Future Gener Comput Syst. 2019;92:76–83.

[37]   Choi J, de Oliveira DP, Leite F. A Novel Approach to Capture Similarity in Capital Project Benchmarking: An Application to Healthcare Projects. J. Manag. Eng. 2022;38(3):05022007.