Expert Systems With Applications 292 (2025) 128555

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Ultrasound report generation with fuzzy knowledge and multi-modal large language model

Ziming Li [a,1], Mingde Li [b,1], Wei Wang [b,*], Qinghua Huang [a,c,*]

[a] School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, 127 West Youyi Road, Beilin District, Xi'an, 710072, Shaanxi, PR China
[b] Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, Ultrasomics Artificial Intelligence X-Lab, The First Affiliated Hospital of Sun Yat-Sen University, No. 58 Zhongshan 2nd Road, Guangzhou, 510080, Guangdong, PR China
[c] School of Mechanical Engineering, Translational Research Center, Shanghai Yangzhi Rehabilitation Hospital(Shanghai Sunshine Rehabilitation Center), Tongji University, 4800 Cao an gong Road, Jiading Distrct, Shanghai, 201804, PR China

ARTICLE INFO

Keywords:
Breast
Liver
Multimodal large language model
Report generation
Thyroid
Ultrasound image

ABSTRACT

Ultrasound report generation is a critical component of computer-aided diagnosis, aimed at alleviating the workload of radiologists during scanning procedures and enhancing diagnostic efficiency. Despite advancements in automatic report generation technologies, the development of a unified framework for generating reports across diverse anatomical regions in ultrasound imaging remains a significant challenge. In this study, we propose a novel and efficient multimodal large language model framework specifically designed for ultrasound report generation. Our framework leverages fuzzy theory to extract essential anatomical knowledge from statistical features, thereby providing more accurate and context-aware guidance throughout the report generation process. Furthermore, we propose a novel evaluation metric designed to assess both the precision and the clinical significance of the generated reports, leveraging insights derived from deep domain expertise. In contrast to traditional evaluation methods, this metric offers a more comprehensive and clinically meaningful assessment. To validate the efficacy of our framework, we conduct extensive experiments on both a publicly available dataset and a proprietary dataset collected from the First Affiliated Hospital of Sun Yat-sen University. We also supplemented our proprietary ultrasound dataset with an external validation set collected from Foshan Sanshui Hospital and The First Affiliated Hospital of Guangzhou. Experimental results demonstrate that our approach consistently achieves state-of-the-art performance across multiple evaluation metrics, highlighting its robustness and adaptability. These findings underscore the potential of our framework in advancing the accuracy and clinical applicability of ultrasound report generation.

1. Introduction

Ultrasound imaging is a non-invasive diagnostic tool that provides clear visualization of soft tissues and blood flow, aiding in lesion detection and comprehensive disease diagnosis and treatment planning (Huang et al., 2024; Lin, Chen, Chen, & Garibaldi, 2023; Roy & Maji, 2020). However, its reliance on manual probe operation leads to variability in imaging angles and content across patients, requiring radiologists' expertise for accurate interpretation. As ultrasound imaging scales, real-time interpretation becomes increasingly challenging, necessitating the development of automated methods for image analysis.

To address this, researchers have explored multimodal data fusion between images and text (Liu, Chen, Wang, Lu, & Zhang, 2024). Deep learning based image captioning techniques, using convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) (Song et al., 2019) or long short-term memory (LSTM) networks (Gao, Guo, Zhang, Xu, & Shen, 2017) for text generation, have made significant progress. These techniques have paved the way for automated medical report generation. However, the recursive nature of RNNs presents challenges in long-text generation, leading to issues such as vanishing gradients and limitations in generating comprehensive lesion reports.

To improve this, hierarchical RNNs have been proposed for more effective long-text understanding. Jing, Xie, and Xing (2018) used sentence- and word-level LSTMs for enhanced accuracy in chest X-ray report generation. Further work (Du et al., 2023; Pang, Li, & Zhao, 2023)

* Corresponding authors.
E-mail addresses: 2019201687@mail.nwpu.edu.cn (Z. Li), limd25@mail2.sysu.edu.cn (M. Li), wangw73@mail.sysu.edu.cn (W. Wang), qhhuang@nwpu.edu.cn (Q. Huang).
[1] First author.

https://doi.org/10.1016/j.eswa.2025.128555
Received 1 March 2025; Received in revised form 8 May 2025; Accepted 7 June 2025
Available online 11 June 2025

has focused on generating diagnostic reports for specific medical modalities and anatomical regions.

However, simply categorizing reports into normal and abnormal regions is insufficient for recognizing multiple lesions caused by different diseases. This often leads to semantic loss and ambiguity, affecting the accuracy of the generated reports (Pang et al., 2023). To address this, some studies incorporate label encodings into report generation networks to guide disease feature recognition and text generation (Li et al., 2024b; Qin, Zhang, & Guo, 2023). While this improves accuracy, it requires clinician-provided annotations, limiting automation and generalizability across diverse datasets.

Large language models, with their massive parameters and pre-trained data, have demonstrated superior text comprehension capabilities. Their transformer-based structure leverages full-text attention mechanisms for capturing and understanding global features, while the parallel architecture significantly accelerates text generation. This has positioned LLMs as state-of-the-art techniques for natural language generation tasks. Researchers have also begun exploring methods to align multimodal features between image extraction models and text, with pre-trained models like CLIP (Radford et al., 2021) offering a promising solution. Multimodal large language models (e.g., Qwen-VL (Bai et al., 2023), LLava (Liu, Li, Wu, & Lee, 2023)) have adopted such approaches to align different modalities theoretically. As large-scale probabilistic models, multimodal LLMs require vast amounts of image-text data to learn rich cross-modal representations, enhancing the model's knowledge representation of domain-specific content. Using this technique, researchers have trained models on large-scale biomedical datasets, improving their understanding of medical images and achieving promising results in medical imaging tasks such as CT, MRI, and endoscopy report generation.

Compared to other medical imaging modalities, ultrasound images present lower clarity and are prone to motion artifacts due to patient movement, such as breathing (Huang et al., 2023b). Unlike fixed-position imaging methods like CT or MRI, ultrasound often exhibits irregular shifts in lesion regions, making it difficult to construct accurate region descriptions, thereby hindering the model's ability to learn aligned ultrasound-text multimodal features. Additionally, the propagation characteristics of sound waves in soft tissues often produce artifacts due to the structural properties of different tissues, further degrading the image quality of lesion areas in ultrasound images (Yoon, Khan, Huh, & Ye, 2019). This increases the difficulty for models to accurately extract features from ultrasound images, making it challenging to generate detailed reports that capture both the pathological characteristics and tissue structures.

To address these challenges, we propose a novel ultrasound report generation algorithm based on a multimodal large language model that meets the clinical demand for multi-site ultrasound report automation. Our approach mitigates the effects of noise in ultrasound imaging by integrating omics-based statistical knowledge extraction methods and using a supervised learning strategy to construct embedding techniques that encode anatomical site recognition knowledge, providing guidance for report text generation. To enhance the model's capability in generating reports across multiple anatomical sites, we employ an ensemble learning strategy, constructing a site-specific model activation mechanism to drive text generation, thereby improving the model's recognition and comprehension of different diseases across multiple sites.

Additionally, we introduce a novel factual consistency mechanism, analyzing whether the generated text omits or misrepresents factual information from a semantic perspective, complementing the limitations of cosine similarity-based evaluations. To validate the effectiveness of our model and evaluation framework, we conduct experiments on both publicly available dataset A and a self-collected dataset comprising 5000 breast ultrasound cases, 5000 thyroid ultrasound cases, and 5000 liver ultrasound cases collected from different medical centers. One example's components are shown in Fig. 1. The results demonstrate that our approach achieves state-of-the-art (SOTA) performance in generating ultrasound reports through multi-site and multi-disease fusion analysis.

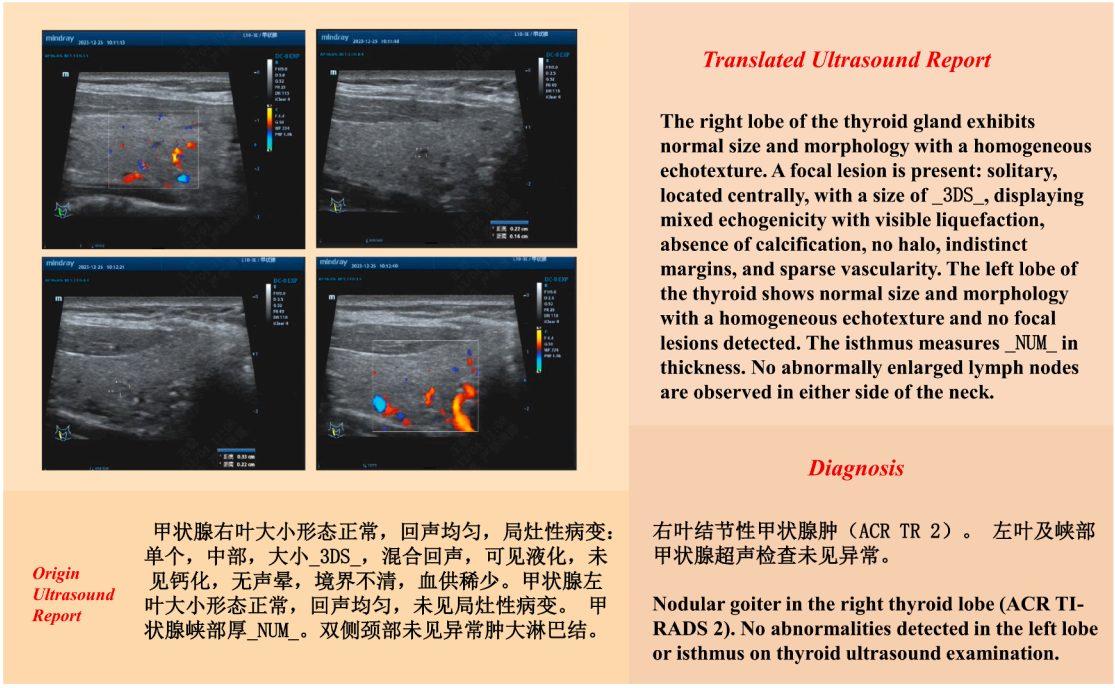


Fig. 1. Example of ultrasound report generation. We adopted the numerical replacement rules proposed by Li et al. (2024b) to address issues such as the lack of a scale, which prevent the model from performing numerical computations and subsequently affect the generation accuracy. Specifically, we use _2DS_, _3DS_, _SCM_, _SMM_, and _LOC_ to represent two-dimensional values, three-dimensional values, one-dimensional centimeter-level values, one-dimensional millimeter-level values, and values in positional data, respectively.

Our main contributions are as follows:

- We propose an innovative framework utilizing a mixture-of-experts strategy for enhanced ultrasound report generation with site-specific feature amplification is proposed, improving visual-textual alignment without additional annotations, thereby enhancing generation efficiency.
- We develop a novel approach combining uncertainty strategies with supervised learning to enhance site-specific recognition in ultrasound imaging is developed, improving model understanding and report accuracy by focusing on anatomical site characteristics.
- We introduce a fact-based similarity comparison mechanism that performs a comprehensive accuracy analysis from multiple scales. This method enables a more effective evaluation of the precision of generated ultrasound reports.

The rest of the article is organized as follows: Section II provides a brief of the related works. Section III presents the technical methods underlying our proposed framework. Section IV describes the design and implementation of the experiments and the comprehensive experimental results. Section V presents the conclusion and the future direction.

2. Related works

In the followings, We review the existing work in two key areas: the development of multi-modal large language models and advancements in image captioning algorithms.

2.1. Multi-modal large language models

In natural language processing (NLP) tasks, large language models have demonstrated remarkable reasoning capabilities by increasing both data scale and model parameters, achieving state-of-the-art (SOTA) results across various text processing tasks. When it comes to tasks that require advanced reasoning, such as audio, vision, and other non-text modalities-like image inference and visual question answering-researchers have begun to leverage the strong feature processing and long-context retrieval capabilities of LLMs (Yin et al., 2024). Some methods directly input non-text modalities and use multi-layer cross-modal attention to fuse these features within the LLM, as seen in models like Flamingo (Alayrac et al., 2022). These methods effectively utilize multimodal features while maintaining input integrity. However, as LLM parameters increase, this approach significantly escalates model complexity, resulting in higher computational demands. This makes it unsuitable for scaling to models with tens of billions of parameters, thereby limiting the overall learning capacity.

Other approaches adopt a supervised learning framework to connect multimodal feature extraction layers with the LLM layers by introducing embedding layers at the feature level. These methods construct cross-modal embeddings to integrate non-text features, such as visual data, into text-based inputs. For instance, models like LLaVa-onevision(Llava-OV) (Li et al., 2024a) and LLaVa-med (Li et al., 2023a) employ a two-layer multilayer perceptron (MLP) structure as an embedding model. Similarly, BLIP2 (Li, Li, Savarese, & Hoi, 2023b), based on a transformer architecture, proposes a learnable query construction mechanism known as the q-former, which uses three different self-attention masking techniques: bi-directional masking, multi-modal masking, and uni-modal masking. This approach aligns visual and textual features, creating a unified representation of both image and text modalities, which is then input into the LLM. Such methods can disassemble the LLM and effectively utilize pretrained multimodal model weights, such as those from CLIP and ImageBind (Girdhar et al., 2023).However, these approaches may lead to the loss of detailed features in non-text modalities, and when dealing with highly complex modalities, maintaining accurate fusion becomes challenging. As a result, multimodal LLMs based on these connection-layer methods tend to exhibit suboptimal performance in tasks that demand strong reasoning capabilities.

Therefore, it is essential to adjust the modality fusion methods according to the specific requirements of different tasks. Additionally, incorporating external knowledge can assist in aligning multimodal features or tokens with text modality representations, allowing for the seamless integration of textual instructions into the input for large language models. This approach ensures that the diverse modalities are effectively harmonized within the model, enhancing its performance across a variety of complex tasks.

2.2. Image captioning

Image captioning, an interdisciplinary task combining computer vision and NLP, aims to enable computers to generate meaningful textual descriptions of images. Traditional methods relied on template matching, using algorithms like SIFT and SURF, but struggled with modeling object relationships. For instance, Yang, Teo, Daumé, and Aloimonos (2011) used a template-based approach with predefined sets of objects, actions, and prepositions, yet their methods failed to address complex relational patterns effectively. Kulkarni et al. (2013) incorporated conditional random fields for a three-step image understanding process: identifying objects, attributes, and spatial relations. Despite improvements, these methods were limited to low-level pattern recognition and struggled with issues like occlusion and non-scale transformations.

With the advent of deep learning, CNNs became popular for image feature extraction, overcoming some of these limitations. CNNs can capture deep semantic features, making them better suited for analyzing complex images (Huang, Jia, Ren, Wang, & Liu, 2023a). RNNs, particularly in encoder-decoder architectures, facilitated contextual text generation (Sharma, Dhiman, & Kumar, 2023). Kiros, Salakhutdinov, and Zemel (2014) and Vinyals, Toshev, Bengio, and Erhan (2015) optimized multi-modal embeddings for improved text generation. Despite these advances, issues remain in heterogeneous feature fusion, and RNNs still face challenges with long-range dependencies, limiting their ability to maintain long-text coherence.

In contrast, transformer architectures, with their parallel computation and global context capture, offer advantages in generating long texts (Bai & An, 2018). Vision transformers, enhanced by cross-modal attention, improve the fusion of visual and textual features (Cornia, Stefanini, Baraldi, & Cucchiara, 2020). Yang et al. (2021) introduces AMAnet, a multi-label classification network with spatial-semantic attention. However, low-parameter transformer models still struggle with contextual depth and syntactic complexity, particularly with limited training data.

Recent advancements focus on large-scale pre-trained models like BERT (Devlin, Chang, Lee, & Toutanova, 2019) and T5 (Raffel et al., 2019), which leverage extensive datasets and multi-layer transformers for superior text understanding. Cross-modal models like CLIP(Xu, Zhao, Li, & Wang, 2023) and ALGCN have shown significant promise in image-text fusion, with applications in image captioning (Mall et al., 2023) and remote sensing tasks (Wang et al., 2023). Huh, Park, and Ye (2023) Proposes a LangChain-based pipeline integrating image-analysis tools with an LLM for breast ultrasound report generation. These methods, however, require additional fine-tuning and task-specific guidance, which introduces data and annotation dependencies.

3. Methods

3.1. Expert system-enhanced multimodal large language model

We propose a novel methodology within the multimodal large language model framework, using the Llava-OV architecture as the core structure, as shown in Fig. 2. Our model is divided into three components: the multimodal language model, fuzzy knowledge extraction embeddings, and a fact-based evaluation system. The multimodal model leverages Llava-OV and integrates image and text features through innovative instructions. The fuzzy knowledge extraction algorithm aids

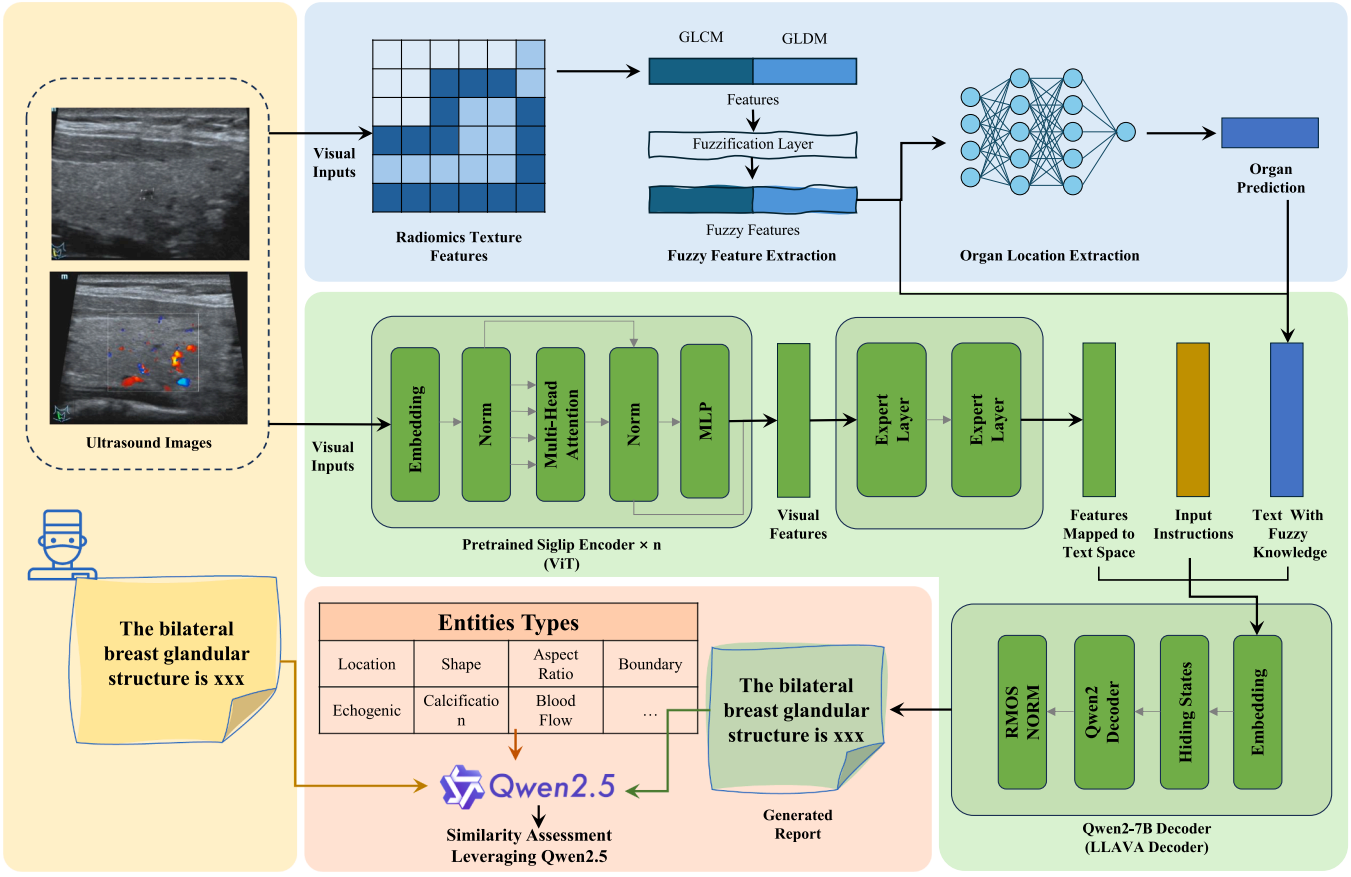


Fig. 2. Structure of the novel multi-modal large language model for ultrasound report generation. The blue region indicates the Fuzzy Knowledge Extraction and Prompt Construction Network developed, the green region denotes the Expert System-Enhanced Multimodal Large Language Model built, and the red region corresponds to the Fact-Based Evaluation Metrics computation method established.

the model in interpreting ultrasound images, while the fact-based evaluation system provides a scientific metric for assessing ultrasound image generation accuracy, validating the model's reliability.

3.1.1. Siglip vision encoder

For precise image comprehension, accurate feature extraction algorithms are critical. We use the Siglip encoder (Zhai, Mustafa, Kolesnikov, & Beyer, 2023), as shown in Fig. 3, based on the Vision Transformer (ViT) (Dosovitskiy et al., 2021) framework, as the image encoder. Siglip aligns visual and textual modality features in a shared embedding space using contrastive learning, maximizing semantic consistency for paired image-text pairs and minimizing similarity for non-paired ones, as in Eq. (1). This enhances alignment and increases similarity between real images and corresponding descriptions. Siglip also employs pairwise Sigmoid loss to focus on each image-text pair, reducing reliance on other samples and mitigating issues of global normalization and weight averaging.

$$\mathcal{L}_{\text{Siglip}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\log(\sigma(s_{ii})+\tau) + \sum_{j=1,j\neq i}^{N}\log(\sigma(-s_{ij})+\tau)\Big] \tag{1}$$

Where:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{2}$$

In the equation, s_{ij} represents elements within the similarity weight matrix. Here, i indicates the index of the input image element, and j refers to the index of the text description. Correspondingly, the input image-text pairs can be represented as x_i and y_j. We denote the Vision Transformer-based visual encoder as f_{image} and the transformer-based text encoder as f_{text}. τ is the temperature parameter, which is utilized to regulate the scaling of the similarity scores between image and text embeddings.The calculation of s_{ij} is presented in Eq. (3).

$$s_{i,j} = \frac{f_{image}(x_i) f_{text}(y_j)}{\|f_{image}(x_i)\| \|f_{text}(y_j)\|} \tag{3}$$

In the equation, $||\cdot||$ represents the $L2$ norm of a vector. This allows us to obtain the distance between the image-text pair inputs, derived through the visual and textual encoders, within the constructed cross-modal embedding space, which serves as the similarity evaluation score.

The Siglip encoder (Tschannen et al., 2025) also exhibits exceptional zero-shot learning capabilities, primarily due to its transformer-based bidirectional sequence structure, which enhances model scalability. This scalability enables the construction of large-parameter models, capable of handling larger datasets and more complex tasks, thus acquiring generalized knowledge. Siglip's embedding space does not require specific labels for training; instead, it utilizes textual descriptions within the space as labels, allowing the model to match images with arbitrary descriptions and assess their similarity. Assuming a new category $\hat{y}$, the textual input for this category in the cross-modal embedding space is represented as $[y_0 \sim y_j, \hat{y}]$, leading to a similarity matrix S expressed as S_{ij+1}. This probabilistic capability allows the model to interpret images in new categories based solely on textual descriptions. Given the extensive disease categories in ultrasound imaging and the limited data for some conditions, Siglip's zero-shot learning is crucial for achieving effective results.

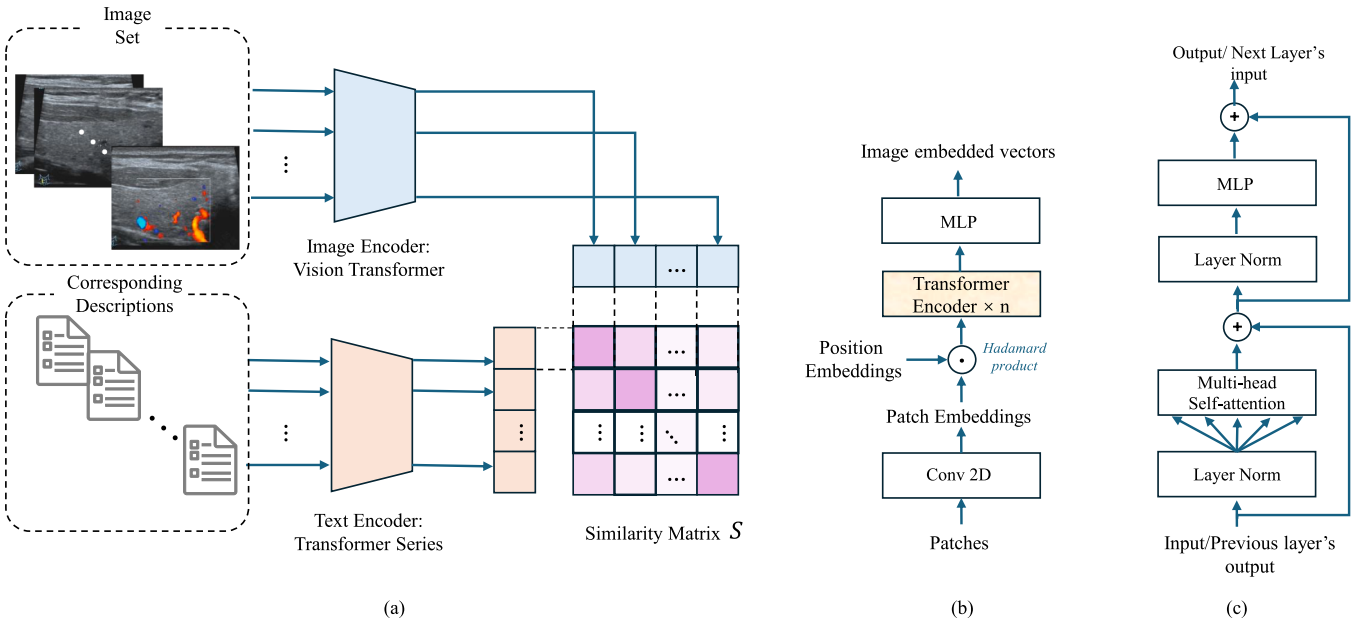


Fig. 3. Structure of the Siglip encoder with the training stragedy. (a) The training method of the Siglip network, which represents the implementation process of the sigmoid contrastive learning strategy; (b) The structure of the Siglip visual encoder network; (c) The structure of the transformer feature extraction layer.

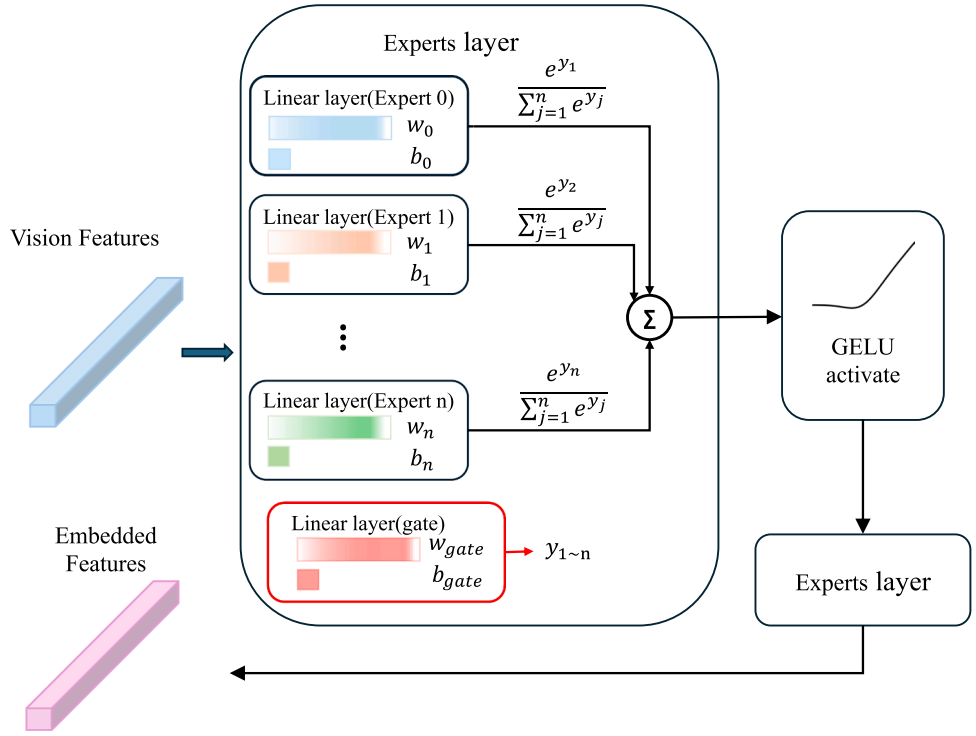


Fig. 4. The structure of the moe-enhanced projector, consisting of a mixture of n expert layers, along with a gating layer for controlling the expert outputs.

3.1.2. Llava based decoder

We employed Qwen2 as the language model for text comprehension and generation. Pretrained on multilingual data, including Chinese, English, and knowledge triples, Qwen2 benefits from efficient parameter tuning and knowledge distillation, enabling strong contextual understanding across various NLP tasks. However, as Qwen2 supports only text input, we integrated image information through a cross-modal feature fusion method proposed by LLaVA-Next. This method utilizes a multilayer perceptron (MLP) with two linear layers and a GELU activation function between them, as shown in Eq. (4).

$$I_{text} = W_2((W_1 * f_{image}(I_{image}) + b_1) * \Phi(W_1 * f_{image}(I_{image}) + b_1)) + b_2 \quad (4)$$

In the equation, $\Phi(x)$represents the cumulative probability distribution, while $W_n, n = 1, 2$ denotes the weights of the nth linear layer, and b_n represents the bias term of the nth layer. LLaVA-Next maps visual features into an intermediate space through a projection in the first layer, incorporating a nonlinear transformation. Subsequently, it projects the

Table 1
GLCM feature descriptions, the feature extraction method for the GLDM is analogous to that of the GLCM with the corresponding matrix used is changed to $G(i,j)$.

Feature Name	Concept	Formula
Contrast	Local variations in gray levels	$\sum_{i,j}(i-j)^2 P(i,j)$
Correlation	Correlation between gray levels	$\frac{\sum_{i,j}(i-\mu_x)(j-\mu_y)P(i,j)}{\sigma_x\sigma_y}$
Energy	Uniformity of the image	$\sum_{i,j} P(i,j)^2$
Entropy	Complexity of the image	$-\sum_{i,j} P(i,j)\log(P(i,j))$
Homogeneity	Consistency of gray levels	$\sum_{i,j} \frac{P(i,j)}{1+\lvert i-j\rvert}$
Maximum Probability	The highest probability in the GLCM matrix	$\max P(i,j)$
Inverse Difference Moment (IDM)	Smoothness of the image	$\sum_{i,j} \frac{P(i,j)}{1+(i-j)^2}$
Variance	Variance of the gray level distribution	$\sum_{i,j}(i-\mu)^2 P(i,j)$
Standard Deviation	Standard deviation of the gray level distribution	$\sqrt{\text{Variance}}$
Skewness	Skewness of the gray level distribution	$\frac{\sum_{i,j}(i-\mu)^3 P(i,j)}{\sigma^3}$
Kurtosis	Kurtosis of the gray level distribution	$\frac{\sum_{i,j}(i-\mu)^4 P(i,j)}{\sigma^4}$
Total Energy	Total energy of the gray levels	$\sum_{i,j} P(i,j)$
Gray Level Variance	Distribution of gray level variance	$\frac{1}{L^2}\sum_{i=0}^{L-1}(P(i,j)-\bar{P})^2$
Uniformity	Uniformity of the gray level distribution	$\sum_{i,j} P(i,j)^2$

intermediate features into the final embedding space to align with the language model's embedding space, thereby constructing an input mode of {user: images, user: text} as an instruction for the model to achieve the desired generation.

However, this approach treats the mapping of input image features uniformly, lacking the capability for targeted analysis based on the distinct characteristics of different features. To enhance the robustness of the model and improve its representational ability for diverse visual features, we employ a mixture of experts mechanism to construct an enhanced linear layer for forward propagation, which is utilized to establish a multimodal projector, as shown in Fig. 4.

$$p_i(x) = \frac{\exp\left(g_i(x)\right)}{\sum_{j=1}^{E}\exp\left(g_j(x)\right)} \tag{5}$$

$$f_i(x) = w_i(x) + b_i, i = 1, 2, \ldots, n \tag{6}$$

$$g(x) = w_{gate}(x) + b_{gate} \tag{7}$$

$$y = \sum_{i=1}^{n} p_i(x)\cdot f_i(x) \tag{8}$$

In these equations, x represents the input for each expert layer, $p_i(x)$ denotes the confidence weight of the i-th expert, $w_i(x)$ and b_i refer to the weight vector and bias vector of the linear layer for the i-th expert, respectively. $f_i(x)$ signifies the output of the i-th expert, while $w_{gate}(x)$ and b_{gate} represent the weight vector and bias vector of the gating layer. $g(x)$ is the n-dimensional vector output from the gating layer, utilized to compute the confidence weights for the n expert layers. Finally, y denotes the output of the constructed projector.

Specifically, we enhance the constructed linear transfer layer by employing multiple expert layers alongside a gating layer. Each expert layer consists of a linear layer designed to process the input features and perform cross-modal embeddings. The gating layer is utilized to establish a weight distribution for each expert layer, ultimately yielding a weighted sum as the output. This process results in the generation of embedded features mapped to the text modality, which are then integrated with textual instructions to construct the input for the Qwen2 large language model.

3.2. Fuzzy knowledge extraction specialized for organ location information

Based on clinical practice, radiologists use distinct interpretative methods for ultrasound images of different anatomical regions. For example, liver diagnosis involves assessing ultrasound density distribution for fatty liver disease, while breast ultrasound focuses on boundary characteristics and internal features to evaluate tumor malignancy. In thyroid ultrasound, specialists examine borders, shape, and size for hyperthyroidism diagnosis. Accurate identification of anatomical sites is essential for generating precise ultrasound reports.

To capture ultrasound content and enhance model interpretability, we develope a fuzzy knowledge extractor module, consisting of a feature extraction model and a fuzzy multilayer perceptron. The feature extraction module applies statistical methods to analyze and extract meaningful characteristics from the image, enhancing computational efficiency. Specifically, we use gray-level co-occurrence and distribution matrices, as defined in Eqs. (9) and (10), to extract descriptors listed in Table 1 for anatomical site detection validation.

$$P(i,j) = \frac{count(i,j)}{N} \tag{9}$$

$$G(i,j) = \sum_{k} count(i,j,k) \tag{10}$$

In these formulas, $P(i,j)$ represents the GLCM matrix, $G(i,j)$ denotes the GLDM matrix, i and j represent different gray levels, and k indicates the position of neighboring pixels under specific conditions (e.g., above, below, left, or right).

However, ultrasound imaging is challenged by low clarity and ambiguous boundaries, which introduce uncertainty in image content interpretation. Artifacts, such as mirror artifacts, further complicate accurate understanding of ultrasound content. To address the impact of low clarity and artifacts in ultrasound images, we applied fuzzy methods to mitigate uncertainty. Specifically, we constructed an uncertainty representation using a triangular fuzzy membership function as shown in Eq. (11).

$$f_i(x) = \max\left(\min\left(\frac{x-a_i}{b_i-a_i+\epsilon}, \frac{c_i-x}{c_i-b_i+\epsilon}\right), 0\right) \tag{11}$$

Thus,

$$\mathbf{X}_{fuzzy} = \begin{bmatrix} f_1(x_{1,1}) & f_2(x_{1,2}) & \cdots & f_{28}(x_{1,28}) \\ f_1(x_{2,1}) & f_2(x_{2,2}) & \cdots & f_{28}(x_{2,28}) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_{N,1}) & f_2(x_{N,2}) & \cdots & f_{28}(x_{N,28}) \end{bmatrix} \tag{12}$$

Here, a_i and c_i denote the minimum and maximum values of feature x_i, respectively, with b_i representing the midpoint of the ith feature within the constructed set of 28 combined features. The parameter ϵ is a small constant added to prevent division by zero, enhancing numerical stability. The function $f_i(x)$ denotes the fuzzification function applied to the ith feature. $\mathbf{X}_{fuzzy}$ represents the matrix of fuzzified features, where each element corresponds to the membership degree of a specific feature, while $x_{i,j}$ indicates the jth feature value for the ith sample. Noise caused by low clarity and artifacts typically resides in the extreme regions of feature distributions, proximate to the boundaries a_i or c_i. Moreover, the triangular function assigns linearly higher weights to intermediate feature values within [c-a,c + a]and rapidly attenuates extremes. This mechanism Suppresses Outlier Noise: Gaussian

and half-Gaussian, who tails decay more slowly, retaining spurious extremes; in contrast, the triangular function's sharper linear decay effectively down-weights long-tail noise in statistical feature distributions, enhancing robustness. In ultrasound imaging, anatomical boundaries often correspond to mid-range gray-level co-occurrences; the triangular mapping accentuates these boundary-related features, improving their discriminability in downstream classification. Through the membership function, the membership degrees of such points are suppressed to near-zero values, whereas the central region (vicinal to b_i retains membership degrees approaching unity. Simultaneously, the membership degrees for normal feature values exhibit a smooth transition, thereby preserving the semantic integrity of the features. We constructed the anatomical site extraction using a two-layer MLP with Relu as the activation function to obtain the anatomical region information **Region**. Additionally, we developed a novel instruction modality to serve as the prompt information, which was constructed with the next four sentences. We determined the decision to store the weights of the constructed MLP solely based on the classification accuracy achieved on the validation set.

- user: {**images**}
- user: The anatomical region corresponding to this ultrasound image is {**Region**};
- user: The texture features of the input ultrasound image, processed through fuzzy knowledge are X_{fuzzy}.
- user: Provide the ultrasound report.

3.3. Fact-based evaluation metrics based on entity extraction

The existing evaluation methods based on cosine similarity effectively model the consistency between generated content and text. However, the process of directly embedding sentences into vectors can lead to high similarity scores when words or phrases are structurally similar, making it difficult for cosine similarity to distinguish these subtle yet important factual differences. As a result, this approach lacks the capability for deep semantic understanding or factual verification. The BLEU metric, which is constructed based on n-gram matching precision, assesses translation accuracy by calculating the overlap of n-grams between the generated text and reference translations. Although this method allows for a more granular comparison of matched texts, it still lacks deep semantic understanding and matching capabilities, thereby affecting the precision of the evaluation.

To address these limitations, we employ an entity extraction method to structure the content. Leveraging the deep understanding capabilities of Qwen 2.5 (Hui et al., 2024), we first utilized the method proposed and combined it with clinical experience to extract nine entity types from the report labels and generated reports, including location, shape, aspect ratio, cystic-solid nature, boundary, echogenicity, calcification, posterior acoustic shadowing, and blood flow. We then performed linking and recognition using Qwen2.5 and constructed the evaluation metrics through a comprehensive scoring system. This evaluation effectively identifies and assesses the deep semantics of the generated reports, enhancing the reliability of the evaluation. The overall workflow is presented in Algorithm 1.

Algorithm 1 Ultrasound report generation.

```
Pseudo code: {input:raw_image; output:report}
function Ultrasound_Report_Generation(images):
    # 1. Visual Encoding
    vision_feats = SiglipEncoder(images)
    # 2. Fuzzy Knowledge Extraction
    stats_feats = ExtractGLCM/GLDM(images)
    fuzzy_feats = TriangularFuzzification(stats_feats)
    record hyper parameters: a,b,c
    # 3. Anatomical Region Identification
    region = Fuzzy_MLP(fuzzy_feats)
    #4. Mixture-of-Experts Projection
    projected_feats = MOEProjector(vision_feats)
    #5. Prompt Assembly
    Prompts=[
        user: {images}
        user: The anatomical region corresponding
        to this ultrasound image is {Region};
        user: The texture features of the input
        ultrasound image, processed through fuzzy knowledge
        are .
        user: Provide the ultrasound report.
    ]
    # 6. LLM Decoding
    report = Qwen2Decoder(concatenate(projected_feats,
             prompts))
    # 7. Fact-based Consistency Scoring
    entities = ExtractEntities(report)
    consistency = ComputeFactMetric(entities)
    return report
```

4. Experiments and results

4.1. Datasets

4.1.1. Public dataset

The public dataset used in this study was constructed by Li et al. (2024b) and consists of 7390 cases, categorized into three main types of ultrasound scans: breast ultrasound (3,521 cases), thyroid ultrasound (2,474 cases), and liver ultrasound (1,395 cases). These cases span a wide range of disease types and patient demographics, including diverse age groups and clinical conditions. Each case is associated with two representative ultrasound images, which were carefully selected by clinical experts. The corresponding ultrasound report provides detailed diagnostic information. The dataset is split into training, validation, and test sets in a 7:1:2 ratio, ensuring that each set contains a representative sample of all case types.

4.1.2. Proprietary dataset

In addition to the public dataset, we have curated a proprietary dataset by collecting ultrasound cases from the First Affiliated Hospital of Sun Yat-sen University. This dataset includes 15,000 ultrasound cases, with 5000 cases each for breast, liver, and thyroid ultrasound scans. All proprietary cases were pre-screened by clinical experts, and any instances with missing data were excluded. The cases were selected to ensure diversity in terms of disease types, patient age groups, and clinical conditions, making the dataset highly representative of real-world scenarios.

Each case is accompanied by a corresponding ultrasound report, providing rich diagnostic information. The ultrasound images for each case range from 2 to 12 images per case, stored in PNG format to preserve image quality. To facilitate data processing, we applied standard encoding techniques to the textual components of the ultrasound reports, similar to the approach employed by Li et al. (2024b). These data are derived entirely from clinical case records, meanwhile removing Patient-identifiable information, and the study protocol received ethical approval under ChiCTR2300073307. These reports are all supported by corresponding disease diagnoses, thus confirming their validity as a gold standard. The age, gender, and disease information of the patients in the dataset are summarized in the Fig. 5.

4.2. Experimental setup

We evaluated the performance of our method on the public dataset using four Nvidia A6000 GPUs. Additionally, we conducted experiments on our proprietary dataset using a server equipped with two Nvidia

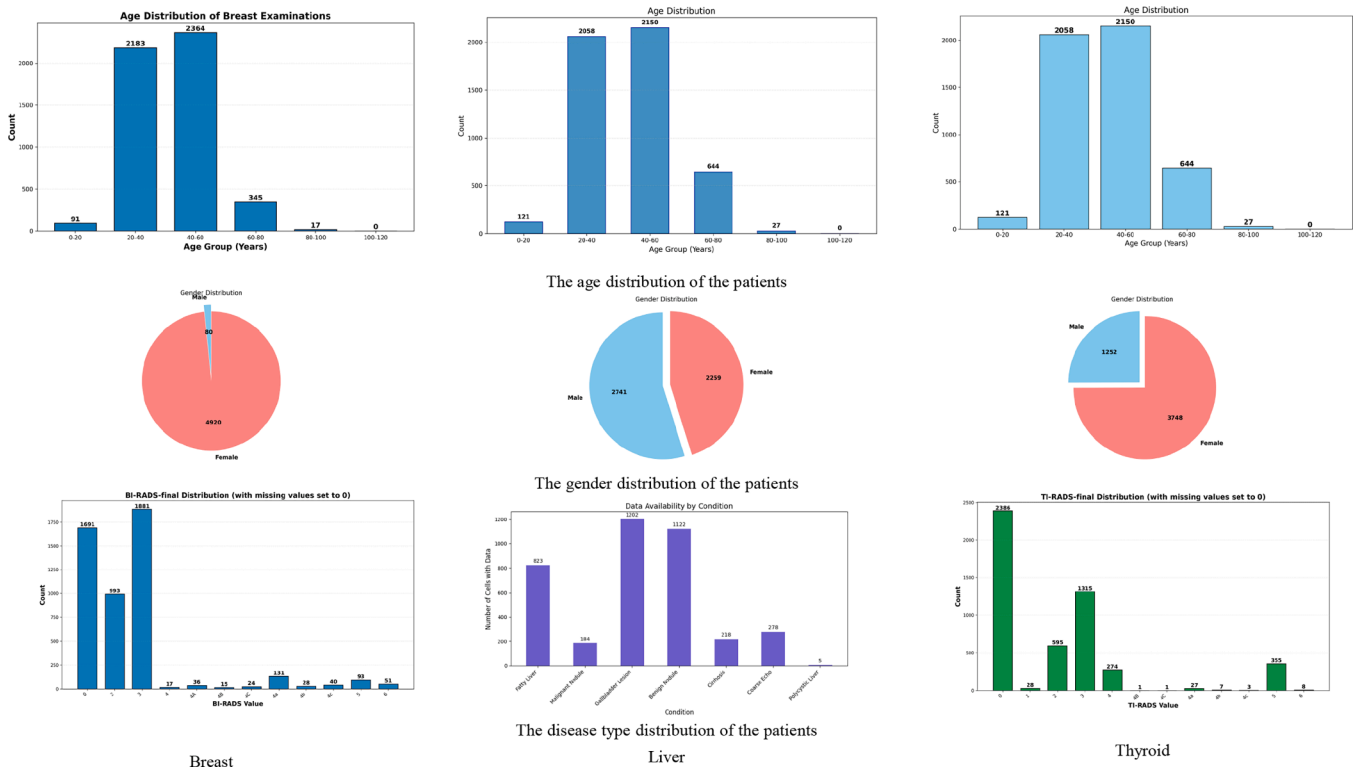


Fig. 5. The age, gender, and disease information of the patients.

A100 80 GB GPUs, one Nvidia A100 40 GB GPU, and one Nvidia H100 GPU. We configured the sparse mapping layer with eight experts, activating the top two experts for each input. We configured the fine-tuning of our large language model and other large language models with a per-device batch size of 1 and trained it over 10 epochs. During inference, with the size of 8,109.53 M parameters in total,when the number of input images is 2, the 5.5 s latency (≈12.5 tokens/s) approaches real-time clinical reporting requirements. Moreover, the 8.11 B-parameter model uses on average 22.9 GB of GPU memory-feasible on modern single-GPU systems.

We employed an autoregressive cross-entropy loss combined with a teacher-forcing strategy to finetune our multimodal LLM. At each generation step t, the model receives the tokenized instructing input I ground-truth prefix tokens $x_{1:t-1}$ and predicts the next token x_t, computing the loss L_t

$$L_t = \sum_{T}^{t} log P(x_t | x_{1:t-1}; I) \tag{13}$$

We back-propagates the gradient at each step, ensuring the loss remains a smooth, differentiable function of the model logits-thus supporting exact gradient computation for every token prediction-and confines each loss calculation to the current time step and its given prefix, obviating any dependence on future tokens or full-sequence context. By employing teacher forcing-supplying the ground-truth token at step $1 : t-1$ rather than the model's own sampled output, we avert error accumulation and enable efficient, parallel computation of token-level losses across the training batch.

4.3. Evaluation metrics

To construct comprehensive evaluation metrics for ultrasound report generation, we utilized bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE-L). These metrics allow for a holistic assessment of the similarity between the generated reports and the reference reports.

BLEU (Papineni, Roukos, Ward, & Zhu, 2002) is a widely used metric for evaluating the overlap of words between generated text and reference text. It measures the degree of overlap at different n-gram levels, including BLEU-1, BLEU-2, BLEU-3, and BLEU-4, thereby capturing various levels of linguistic similarity between the generated and reference reports.

ROUGE-L (Lin, 2004) is based on the longest common subsequence (LCS) algorithm. It considers sentence-level structural similarity and identifies the longest co-occurring n-grams in the sequence. This metric effectively captures the overall structural similarity between the generated report and the reference report.

4.4. Comparison models

CNN-RNN (Vinyals et al., 2015): This approach is one of the earliest methods to apply deep learning for image captioning. We employ a pre-trained ResNet-50 as the feature extraction algorithm for images and use a 3-layer LSTM for text generation, with a hidden state dimension of 256.

R2Gen (Chen, Song, Chang, & Wan, 2020): This method proposes a memory-driven unit that integrates memory into the Transformer architecture to enhance the performance of radiology report generation.

Knowledge-CNN-Transformer(KCT) (Li et al., 2024b): This approach utilizes clustering to enhance case knowledge. Specifically, it clusters reports to extract knowledge about diseases and incorporates this knowledge into the report generation process as guidance. The model uses a CNN for image feature extraction and a Transformer-based encoder-decoder structure for report generation.

Qwen2-VL (Wang et al., 2024): This model employs a pure Transformer-based ViT (Vision Transformer) as the image feature extraction algorithm and introduces multimodal rotational positional embeddings to enhance visual feature understanding. It aligns visual and textual features for report generation.

LLAVA-llama3 Liu et al. (2023): This model uses CLIP as the visual encoder and a two-layer fully connected network as the

Table 2
Quantitative evaluation of report generation performance on the public ultrasound dataset using BLEU and ROUGE-L metrics. We have highlighted the best results in bold for emphasis.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CNN-RNN	0.050	0.024	0.015	0.000	0.142
R2Gen	0.500	0.470	0.448	0.431	0.340
KCR	0.743	0.702	0.672	0.643	0.609
Llava-OV	0.814	0.737	0.709	0.667	0.655
Llava-llama3	0.781	0.731	0.692	0.660	0.640
Qwen2-VL	0.779	0.732	0.694	0.663	0.640
Qwen2.5-VL	0.808	0.772	0.750	0.719	0.670
Llava-med	0.787	0.746	0.715	0.688	0.610
Ours	**0.836**	**0.803**	**0.777**	**0.731**	**0.695**

Table 3
Quantitative evaluation of report generation performance on the proprietary ultrasound dataset using BLEU and ROUGE-L metrics. We have highlighted the best results in bold for emphasis.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Llava-OV	0.722	0.667	0.625	0.588	0.410
Qwen2-VL	0.621	0.558	0.513	0.472	0.312
Qwen2.5-VL	0.639	0.571	0.521	0.478	0.359
Llava-med	0.701	0.644	0.602	0.554	0.406
Ours	**0.751**	**0.699**	**0.659**	**0.624**	**0.441**

mapping layer to map visual features into the text space. Llama3.1 is then used to process the text space and aligned features to generate textual responses.

Llava-OV Li et al. (2024a): This model utilizes Siglip as the visual encoder and a two-layer fully connected network to map visual features into the text space. It then uses Qwen2 to process the text space and aligned features for generating textual responses.

Qwen2.5-VL Bai et al. (2025): This model employs a pure ViT backbone with multimodal rotary positional embeddings and dynamic-resolution support to achieve state-of-the-art results on diverse vision-language benchmarks.

LLaVA-Med Li et al. (2023a): This model extends the LLaVA framework to biomedicine by first aligning domain vocabulary on a large figure-caption corpus and then self-instructing on GPT-4 questions to excel at biomedical VQA.

4.5. Comparative experiment result

Tables 2 and 3 present the quantitative evaluation comparisons of report generation performance using our model and the comparison models on the public ultrasound dataset and the proprietary dataset, respectively. Figs. 6 and 7 provide a quantitative comparison of different methods on both the public dataset and our proprietary dataset. As observed, compared to the current state-of-the-art CNN/Transformer deep learning methods and multimodal large language models, our method generates diagnostic reports that more closely resemble the real labels. Specifically, compared to Llava-OV, which is the second-best performing method, our method shows an improvement of 2.2 %, 6.6 %, 6.8 %, and 6.4 % in BLEU-1 to BLEU-4, respectively. This indicates that our method achieves superior performance in token-level matching evaluation. Additionally, in terms of the ROUGE-L metric, our method outperforms Llava-OV by 1.3 %, demonstrating better performance in sentence-level matching evaluation.

Tables 4 and 5 present the quantitative evaluation comparisons of report generation performance using our model and the comparison models on the public ultrasound dataset and the proprietary dataset, respectively. In terms of advanced semantic cognition, on the public dataset, our method achieves state-of-the-art results in generating entities for six types: Location, Shape, Aspect Ratio, Cystic-Solid Nature, Posterior Acoustic Shadowing, and Blood Flow. Particularly, in terms of Location recognition, our method reduces the error count by 30 compared to the second-best method, Llava-OV. For the remaining three entity types, the recognition performance of our method is very close to the state-of-the-art methods, which proves that our approach generates ultrasound reports with higher content accuracy, better meeting the clinical needs of physicians.

Table 4
Error entities by category in report generation from different methods on the public ultrasound dataset. We have highlighted the best results in bold for emphasis.

Model Boundary	Location Echogenicity	Shape Calcification	Aspect Ratio Posterior Acoustic Shadowing	Cystic-solid Nature Blood Flow
CNN-RNN	2951	294	610	131
170	130	363	188	839
R2Gen	812	410	381	272
40	**85**	64	106	331
KCR	341	211	322	42
105	90	119	79	238
Llava-OV	240	184	260	41
63	96	96	**64**	176
Llava-llama3	272	152	271	32
100	**61**	124	60	176
Qwen2-VL	307	299	304	49
96	84	128	74	218
Qwen2.5-VL	274	250	280	30
90	70	130	62	200
Llava-med	262	145	310	47
89	73	69	79	192
Ours	**201**	**142**	**243**	**28**
70	82	85	**56**	**162**

Table 5
Error entities by category in report generation from different methods on the proprietary ultrasound dataset. We have highlighted the best results in bold for emphasis.

Model Boundary	Location Echogenicity	Shape Calcification	Aspect Ratio Posterior Acoustic Shadowing	Cystic-solid Nature Blood Flow
Llava-OV	627	344	178	110
300	46	262	466	567
Qwen2-VL	654	325	278	**93**
266	45	244	405	624
Qwen2.5-VL	582	317	240	95
240	60	234	386	603
Llava-med	550	302	185	100
233	44	220	353	568
Ours	**531**	**255**	**132**	118
216	**42**	**205**	**338**	**540**

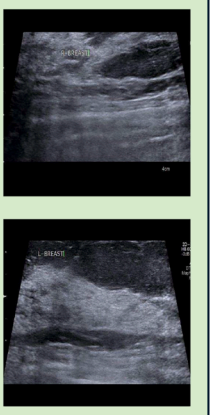
Table 6
Quantitative evaluation of report generation performance on the cross-institution ultrasound dataset using BLEU and ROUGE-L metrics. We have highlighted the best results in bold for emphasis.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Llava-OV	0.440	0.271	0.198	0.160	**0.211**
Qwen2-VL	0.326	0.159	0.076	0.030	0.074
Qwen2.5-VL	0.336	0.174	0.09	0.04	0.07
Llava-med	0.370	0.184	0.104	0.06	0.122
Ours	**0.500**	**0.351**	**0.288**	**0.255**	0.202

4.6. Cross-institution validation

To assess the generalizability of our proposed method compared to existing models, we have incorporated external center evaluations using data unseen during training:

case1:



Ultrasound Images

Knowledge CNN-Transformer :
肝脏形态饱满，包膜光滑，于肝右叶可见一斑状强回声，直径约_SCM_。余肝实质回声细密增强，门静脉系统显示欠清晰，肝肾回声对比增强。肝内外胆管未见扩张。门静脉主干内径正常范围。胆囊大小形态如常，壁不厚，光滑，腔内未见明显异常回声。胰腺大小形态如常，实质回声均匀，胰管不宽，内未见明确占位性病变。脾脏大小形态如常，实质回声均匀，内未见明显占位性病变。

The liver appears full in shape with a smooth capsule. A patchy area of strong echo, approximately SCM in diameter, is observed in the right hepatic lobe. The remaining liver parenchyma shows finely enhanced echotexture. The portal venous system is somewhat unclear, with increased contrast between liver and renal echoes. No dilation is observed in the intrahepatic or extrahepatic bile ducts. The portal vein trunk diameter is within the normal range. The gallbladder is normal in size and shape, with a smooth, non-thickened wall and no obvious abnormal echoes within the cavity. The pancreas is normal in size and shape, with homogeneous parenchymal echotexture, a non-dilated pancreatic duct, and no clear space-occupying lesions. The spleen is normal in size and shape, with uniform parenchymal echotexture and no apparent space-occupying lesions.

Ours:
肝脏形态饱满，包膜光滑，实质回声细密增强，门静脉系统显示欠清晰，后场回声衰减，肝肾回声对比增强。肝内外胆管未见扩张。门静脉主干内径正常范围。胆囊大小形态如常，壁不厚，光滑，腔内未见明显异常回声。胰腺大小形态如常，实质回声均匀，胰管不宽，内未见明确占位性病变。脾脏大小形态如常，实质回声均匀，内未见明显占位性病变。

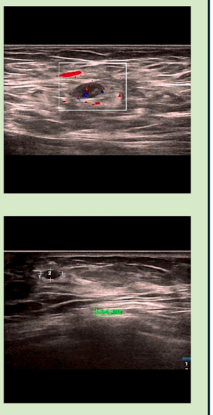
The liver appears full in shape with a smooth capsule and finely enhanced parenchymal echotexture. The portal venous system is somewhat unclear, with echo attenuation in the posterior field and increased contrast between liver and renal echoes. No dilation is observed in the intrahepatic or extrahepatic bile ducts. The portal vein trunk diameter is within the normal range. The gallbladder is normal in size and shape, with a slightly rough, non-thickened wall and no obvious abnormal echoes within the cavity. The pancreas is normal in size and shape, with homogeneous parenchymal echotexture, a non-dilated pancreatic duct, and no clear space-occupying lesions. The spleen is normal in size and shape, with uniform parenchymal echotexture and no apparent space-occupying lesions.

Ground Truth:
肝脏大小形态如常，包膜光滑，实质回声细密增强，门静脉系统显示欠清晰，肝肾回声对比增强。肝内外胆管未见扩张。门静脉主干内径正常范围。胆囊大小形态如常，壁毛糙，欠光滑，腔内未见明显异常回声。胰腺大小形态如常，实质回声均匀，胰管不宽，内未见明确占位性病变。脾脏大小形态如常，实质回声均匀，内未见明显占位性病变。

The liver is normal in size and shape, with a smooth capsule and finely enhanced parenchymal echotexture. The portal venous system appears somewhat unclear, with increased contrast between the liver and renal echoes. No dilation is observed in the intrahepatic or extrahepatic bile ducts. The portal vein trunk diameter is within the normal range. The gallbladder is normal in size and shape, with a slightly rough, less smooth wall and no obvious abnormal echoes within the cavity. The pancreas is normal in size and shape, with homogeneous parenchymal echotexture, a non-dilated pancreatic duct, and no clear space-occupying lesions. The spleen is normal in size and shape, with uniform parenchymal echotexture and no apparent space-occupying lesions.

(a) Case 1 on public dataset

case2:



Ultrasound Images

Knowledge CNN-Transformer :
甲状腺全切术后，原区域未见明确占位性病变。双侧颈部扫查，右颈部III区可探及一低回声结节，大小约_2DS_，边界欠清晰，形态尚规整，“淋巴门”结构显示欠清，CDFI示可探及血流信号。

Post-thyroidectomy, no obvious space-occupying lesions were observed in the original area. Examination of the bilateral neck revealed a hypoechoic nodule in the right neck region III, approximately _2DS_ in size, with poorly defined borders and relatively regular shape. The 'lymph node gate' structure was poorly defined, and CDFI showed detectable blood flow signals.

Qwen2-VL:
双侧乳腺腺体结构稍紊乱，乳导管不扩张，于右侧乳腺_Loc_区距乳头_SCM_处可见一低回声结节，大小约_2DS_，边界清晰，形态规整，CDFI示未探及血流信号。双侧腋下未见明显肿大淋巴结。

The bilateral breast glandular structure is slightly disordered, and the milk ducts are not dilated. A hypoechoic nodule, approximately _2DS_ in size, with clear borders and regular shape, is observed at the right breast in the Loc area, near the nipple _SCM_. CDFI shows no detectable blood flow signals. No obvious enlarged lymph nodes are seen in both axillae.

Ours:
双侧乳腺腺体结构稍紊乱，乳导管不扩张，于左侧乳腺_Loc_区距乳头_SCM_处可见一低回声结节，大小约_2DS_，于右侧乳腺可见多个低回声结节，大者位于_Loc_区乳头旁，大小约_2DS_，边界清晰，形态规整，CDFI示未探及血流信号。左侧腋下扫查可探及一低回声结节，大小约_2DS_，边界清晰，形态规整，双侧腋下未见明显肿大淋巴结。

The bilateral breast glandular structure is slightly disordered, and the milk ducts are not dilated. A hypoechoic nodule, approximately 2DS in size, is observed in the left breast at the Loc area near the nipple _SCM_. Multiple hypoechoic nodules are observed in the right breast, with the largest located near the nipple in the Loc area, approximately _2DS_ in size, with clear borders and regular shape. CDFI shows no detectable blood flow signals. A hypoechoic nodule, approximately _2DS in size, with clear borders and regular shape, is found in the left axilla. No obvious enlarged lymph nodes are seen in the bilateral axillae.

Ground Truth:
双侧乳腺腺体结构稍紊乱，乳导管不扩张，于左侧乳腺_Loc_区乳头旁可见一低回声结节，大小约_2DS_，于右侧乳腺可见多个低回声结节，大者位于_Loc_区乳头旁，大小约_2DS_，边界清晰，形态规整，CDFI示未探及血流信号。左侧腋下扫查可探及一低回声结节，大小约_2DS_，边界清晰，形态规整，皮质增厚，“淋巴门”结构显示不清晰，CDFI示可探及较丰富血流信号。

The bilateral breast glandular structure is slightly disordered, and the milk ducts are not dilated. A hypoechoic nodule, approximately 2DS in size, is observed near the nipple in the left breast at the Loc area. Multiple hypoechoic nodules are observed in the right breast, with the largest located near the nipple in the Loc area, approximately 2DS in size, with clear borders and regular shape. CDFI shows no detectable blood flow signals. A hypoechoic nodule, approximately 2DS in size, with clear borders and regular shape, is found in the left axilla, with cortical thickening. The "lymph node gate" structure is unclear, and CDFI shows relatively abundant blood flow signals.

(b) Case 2 on public dataset

Fig. 6. Examples of ultrasound report generation. Red denotes sections where explicit errors are generated, while underlined text indicates areas where imprecisions are introduced.

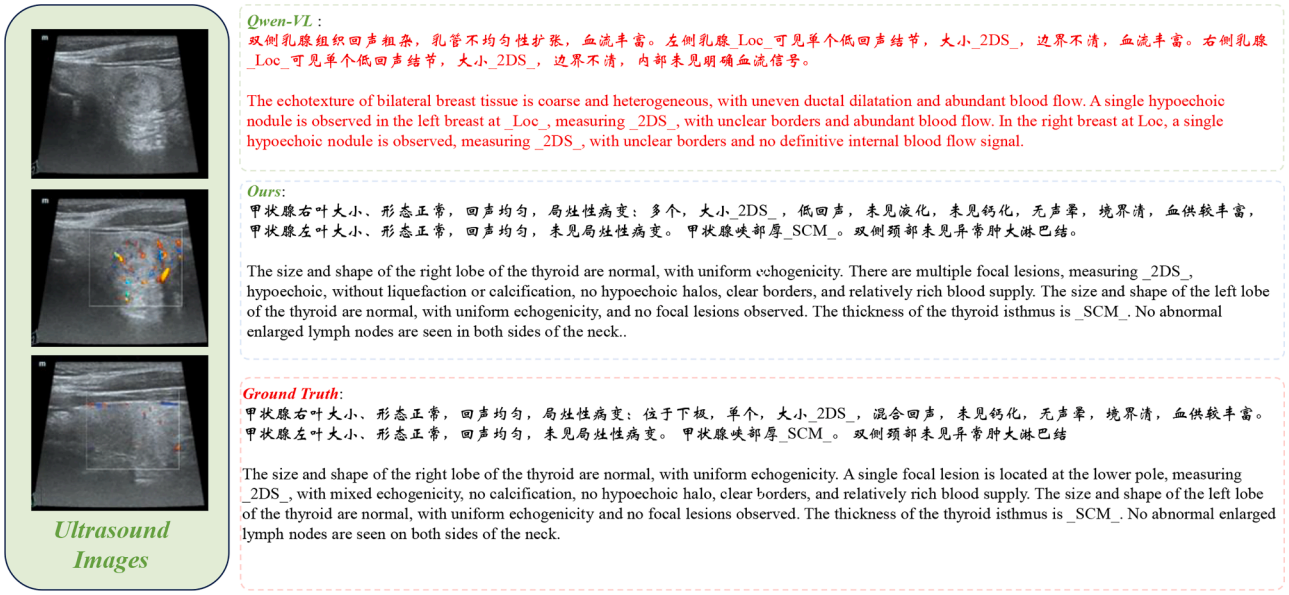


Fig. 7. Example of ultrasound report generation. Red denotes sections where explicit errors are generated.

Fig. 8. Examples of ultrasound report generation. Red denotes sections where explicit errors are generated, while underlined text indicates areas where imprecisions are introduced.

1) Foshan Sanshui Hospital Thyroid Set: 499 cases
2) Foshan Sanshui Hospital Liver Set: 498 cases
3) The First Affiliated Hospital of Guangzhou Breast Set: 424 cases

Table 6 presents the evaluation comparisons of the report generation performance using our model and the comparison models on the cross-institution dataset.

In scenarios where significant inter-center data heterogeneity exists, our model demonstrates superior adaptability by producing more contextually appropriate outcomes. This underscores its enhanced generalization capabilities across diverse clinical environments. The visualization result of a representative case is shown in Fig. 8.

4.7. Statistical significance tests

To validate that our method achieves statistically significant improvements over alternative approaches, we conducted paired t-tests on the collected dataset. The paired *t*-test is a statistical procedure designed to assess whether the mean difference between two related measurements is significantly different from zero. It is most appropriate for “before-after” or matched-subject designs, where each subject contributes a pair of observations. The test operates on the within-pair differences, assumes those differences are approximately normally distributed, and computes a t-statistic from the sample mean and standard deviation of the differences with $n - 1$ degrees of freedom. Deci-

Table 7
Statistical significance tests (paired *t*-test).

Model A	Model B	Metric	Mean A	Mean B	T-statistic	P-value	Significance	Better Model
Llava-OV vs Qwen2-VL								
Llava-OV	Qwen2-VL	bleu1	0.722670	0.621158	21.242530	2.19×10^{-93}	significant	Ltava-OV
Llava-OV	Qwen2-VL	bleu2	0.666707	0.558552	21.299335	7.66×10^{-94}	significant	Ltava-OV
Llava-OV	Qwen2-VL	bleu3	0.625424	0.513131	21.316711	5.55×10^{-94}	significant	Ltava-OV
Llava-OV	Qwen2-VL	bleu4	0.588796	0.473284	21.328314	4.47×10^{-94}	significant	Ltava-OV
Llava-OV	Qwen2-VL	rouge_1	0.410226	0.312751	19.429554	2.75×10^{-79}	significant	Ltava-OV
Llava-OV vs Qwen2.5-VL								
Llava-OV	Qwen2.5-VL	bleu1	0.722670	0.639058	17.394686	1.24×10^{-64}	significant	Ltava-OV
Llava-OV	Qwen2.5-VL	bleu2	0.666707	0.571544	18.223207	1.91×10^{-78}	significant	Ltava-OV
Llava-OV	Qwen2.5-VL	bleu3	0.625424	0.521923	18.743029	3.35×10^{-74}	significant	Ltava-OV
Llava-OV	Qwen2.5-VL	bleu4	0.588796	0.478853	19.181914	1.95×10^{-77}	significant	Ltava-OV
Llava-OV	Qwen2.5-VL	rouge_1	0.410226	0.359139	7.593211	4.14×10^{-14}	significant	Ltava-OV
Llava-OV vs Ours								
Llava-OV	Ours	bleu1	0.722670	0.751157	-11.931011	4.31×10^{-32}	significant	Ques
Llava-OV	Ours	bleu2	0.666707	0.608577	-11.912979	5.30×10^{-32}	significant	Ques
Llava-OV	Ours	bleu3	0.625424	0.659497	-11.894821	6.53×10^{-32}	significant	Ques
Llava-OV	Ours	bleu4	0.588796	0.624617	-11.883921	7.39×10^{-32}	significant	Ques
Llava-OV	Ours	rouge_1	0.410226	0.441809	-11.893417	9.28×10^{-28}	significant	Ques
Qwen2-VL vs Qwen2.5-VL								
Qwen2-VL	Qwen2.5-VL	bleu1	0.621158	0.639058	-4.491421	7.34×10^{-6}	significant	Quen2.5-VL
Qwen2-VL	Qwen2.5-VL	bleu2	0.558552	0.571544	-3.196811	1.40×10^{-3}	significant	Quen2.5-VL
Qwen2-VL	Qwen2.5-VL	bleu3	0.513131	0.521923	-2.129523	3.33×10^{-2}	significant	Quen2.5-VL
Qwen2-VL	Qwen2.5-VL	bleu4	0.473284	0.478853	-1.137898	2.55×10^{-1}	not significant	Quen2.5-VL
Qwen2-VL	Qwen2.5-VL	rouge_1	0.312751	0.359139	-8.268130	2.02×10^{-16}	significant	Quen2.5-VL
Qwen2-VL vs Ours								
Qwen2-VL	Ours	bleu1	0.621158	0.751157	-25.204136	2.86×10^{-127}	significant	Ques
Qwen2-VL	Ours	bleu2	0.558552	0.608577	-25.297918	4.05×10^{-128}	significant	Ques
Qwen2-VL	Ours	bleu3	0.513131	0.659497	-25.327081	2.20×10^{-128}	significant	Ques
Qwen2-VL	Ours	bleu4	0.473284	0.624617	-25.345395	1.50×10^{-128}	significant	Ques
Qwen2-VL	Ours	rouge_1	0.312751	0.441809	-23.065939	1.78×10^{-108}	significant	Ques
Qwen2.5-VL vs Ours								
Qwen2.5-VL	Ours	bleu1	0.639058	0.751157	-22.660612	4.73×10^{-105}	significant	Ques
Qwen2.5-VL	Ours	bleu2	0.571544	0.608577	-23.361456	5.17×10^{-111}	significant	Ques
Qwen2.5-VL	Ours	bleu3	0.521923	0.659497	-23.747890	2.36×10^{-114}	significant	Ques
Qwen2.5-VL	Ours	bleu4	0.478853	0.624617	-24.065363	3.99×10^{-117}	significant	Ques
Qwen2.5-VL	Ours	rouge_1	0.359139	0.441809	-11.843199	1.18×10^{-31}	significant	Ques

sions are made by comparing this statistic against the Student's t distribution, yielding a p-value that guides inference. This method enjoys widespread application in biomedical, psychological, and engineering studies, provided its assumptions are met and sample size is sufficient. The experimental results are presented in Table 7. Specifically, this analysis confirms that our approach yields a statistically significant breakthrough.

4.8. Validation of the proposed fuzzy MLP method for location recognition

To validate the performance of the constructed Fuzzy Knowledge Extraction Network for ultrasound region recognition, we implemented region recognition tasks using ViT (Dosovitskiy et al., 2021), MLP, Fuzzy C-Means (Bezdek, Ehrlich, & Full, 1984), Monte-Carlo dropout (Gal & Ghahramani, 2016), Deep-Ensemble Learning (Ganaie, Hu, Malik, Tanveer, & Suganthan, 2022) and Eviditial Deep Learning methods. The accuracy of region recognition achieved by these methods was compared with that of our proposed fuzzy knowledge extraction network. For the Fuzzy C-Means method, training was performed directly on the entire dataset. For ViT, MLP, Monte-Carlo dropout, Deep-Ensemble Learning, Evidential Deep Learning (Sensoy, Kaplan, & Kandemir, 2018) and our proposed Fuzzy MLP approach, we split the training set of a publicly available dataset to construct a region recognition dataset, using a 7:1:2 ratio for the training, testing, and validation sets. The experimental results presented in Table 8 demonstrate that our method outperforms existing fuzzy-based approaches and Vision Transformer models in anatomical region classification tasks.

5. Limitations

This study demonstrates that our model-trained on mixed ultrasound imaging and report data from three anatomical regions (breast, liver, and thyroid)-achieves high accuracy in report generation, thereby evidencing its cross-site cognitive precision. However, several notable limitations temper the generalizability and clinical applicability of our findings

1) Explainability and Transparency

While the fuzzy-MLP prompts improve anatomical guidance, the internal reasoning chain remains latent. In future work, we will explore RLHF/GRPO-based reinforcement learning and Chain-of-Thought (CoT) prompting to expose intermediate inference steps and boost clinical trustworthiness.

2) Image Quality Robustness

Our training used key high-quality frames; although fuzzy-MLP mitigates noise, large volumes of redundant or poor-quality frames in live scanning may degrade performance. We plan to integrate temporal models (e.g., Video Transformers) or frame-selection networks to filter such frames.

Table 8
Model performance metrics across organs and overall accuracy.

Model	Precision	Recall	F1-score	Accuracy
Liver				
Fuzzy-C-Means	0.146	0.216	0.177	0.216
ViT	0.959	0.986	0.972	0.986
MC_dropout	0.960	0.955	0.958	0.955
Deep Ensemble Learning	0.956	0.974	0.965	0.974
Evidential DL	0.9606	0.9525	0.9565	0.9829
MLP	0.740	0.677	0.707	0.677
Ours	0.962	0.989	0.976	0.989
Mammary				
Fuzzy-C-Means	0.952	0.659	0.775	0.659
ViT	0.900	0.973	0.935	0.973
MC_dropout	0.793	0.835	0.813	0.835
Deep Ensemble Learning	0.866	0.824	0.845	0.824
Evidential DL	0.8246	0.8378	0.8311	0.8385
MLP	0.801	0.960	0.876	0.960
Ours	0.958	0.915	0.937	0.915
Thyroid				
Fuzzy-C-Means	0.420	0.478	0.448	0.478
ViT	0.958	0.824	0.887	0.824
MC_dropout	0.766	0.810	0.787	0.810
Deep Ensemble Learning	0.810	0.766	0.787	0.766
Evidential DL	0.7732	0.7591	0.7661	0.8515
MLP	0.760	0.591	0.666	0.591
Ours	0.901	0.945	0.923	0.945
Overall Accuracy				
Fuzzy-C-Means				0.451
ViT				0.928
MC_dropout				0.867
Deep Ensemble Learning				0.855
Evidential DL				0.891
MLP				0.743
Ours				0.950

3) Cross-lingual Capability

Currently, our study is conducted within a monolingual framework. To improve the generalizability and applicability of the approach, future research should consider extending it to additional languages, such as English and French.

6. Conclusion

In this work, we present a novel multimodal large language model for multi-region ultrasound report generation. Our approach leverages radiomics texture statistical features and addresses the challenges of low image clarity and noise inherent in ultrasound imaging by designing a fuzzy knowledge extraction framework to serve as a zero-shot learning prompt. Additionally, we construct a hybrid expert mapping layer for multi-region recognition, further enhancing the accuracy of report generation by large models. Extensive experimental setups demonstrate the reliability and superiority of our method. However, similar to existing approaches, our method is not sensitive to specific numerical values, and further research is needed to focus on precise numerical prediction and computation to improve clinical reliability and practical applicability.

CRediT authorship contribution statement

Ziming Li: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing - original draft; **Mingde Li:** Validation; **Wei Wang:** Investigation; **Qinghua Huang:** Supervision, Writing - review & editing.

Code availability

The custom code used to perform the analyses in this study will be publicly available at https://github.com/75486210/Ultrasound-Report-Generation-with-Fuzzy-Knowledge-and-Multi-modal-Large-Language-Model.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants 12326609, 82030047, 82371983, 82402386 and 82272076.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). Flamingo: A visual Llanguage model for few-shot learning. 2204.14198v2.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. 2308.12966v3.

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304. https://doi.org/10.1016/j.neucom.2018.05.080

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J. et al. (2025). Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy *c*-means clustering algorithm. *Computers & Geosciences*, *10*(2), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1439–1449). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.112

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *2020 IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10575–10584). Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPR42600.2020.01059

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net.

Du, X., Pan, H., Zhang, K., He, S., Bian, X., & Chen, W. (2023). Automatic report generation method based on multiscale feature extraction and word attention network. In *Web and Big Data* (pp. 520–528). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25198-6_40

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059). PMLR.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151.

Gao, L., Guo, Z., Zhang, H., Xu, X., & Shen, H. T. (2017). Video captioning with attention-based LSTM and semantic consistency. *IEEE Transactions on Multimedia*, *19*(9), 2045–2055. https://doi.org/10.1109/TMM.2017.2729019

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: one embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15180–15190).

Huang, H., Oh, S.-K., Fu, Z., Wu, C.-K., Pedrycz, W., & Kim, J.-Y. (2024). FSCNN: Fuzzy channel filter-Based separable convolution neural networks for medical imaging recognition. *IEEE Transactions on Fuzzy Systems*, *32*(10), 5449–5461. https://doi.org/10.1109/TFUZZ.2024.3450000

Huang, Q., Jia, L., Ren, G., Wang, X., & Liu, C. (2023a). Extraction of vascular wall in carotid ultrasound via a novel boundary-delineation network. *Engineering Applications of Artificial Intelligence*, *121*, 106069. https://doi.org/10.1016/j.engappai.2023.106069

Huang, Q., Wang, D., Lu, Z., Zhou, S., Li, J., Liu, L., & Chang, C. (2023b). A novel image-to-knowledge inference approach for automatically diagnosing tumors. *Expert Systems with Applications*, *229*, 120450. https://doi.org/10.1016/j.eswa.2023.120450

Huh, J., Park, H. J., & Ye, J. C. (2023). Breast ultrasound report generation using langchain. arXiv preprint arXiv:2312.03013.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., Dang, K., Fan, Y., Zhang, Y., Yang, A., Men, R., Huang, F., Zheng, B., Miao, Y., Quan, S., Feng, Y., Ren, X., Ren, X., Zhou, J., & Lin, J. (2024). Qwen2.5-coder technical report. 2409.12186v3.

Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying Visual-semantic embeddings with multimodal neural language models.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). BabyTalk: understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903. https://doi.org/10.1109/TPAMI.2012.162

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., & Li, C. (2024a). LLaVA-OneVision: Easy visual task transfer. 2408.03326v3.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023a). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. 2306.00890v1.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023b). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2301.12597v3.

Li, J., Su, T., Zhao, B., Lv, F., Wang, Q., Navab, N., Hu, Y., & Jiang, Z. (2024b). Ultrasound report generation with cross-modality feature alignment via unsupervised guidance.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Lin, Q., Chen, X., Chen, C., & Garibaldi, J. M. (2023). A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty. *IEEE Transactions on Fuzzy Systems*, *31*(8), 2532–2544. https://doi.org/10.1109/TFUZZ.2022.3228332

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. 2304.08485v2.

Liu, Y., Chen, B., Wang, S., Lu, G., & Zhang, Z. (2024). Deep fuzzy multiteacher distillation network for medical visual question answering. *IEEE Transactions on Fuzzy Systems*, *32*(10), 5413–5427. https://doi.org/10.1109/TFUZZ.2024.3402086

Mall, U., Phoo, C. P., Liu, M. K., Vondrick, C., Hariharan, B., & Bala, K. (2023). Remote Sensing vision-language foundation models without annotations via ground remote alignment. https://arxiv.org/abs/2312.06960v1.

Pang, T., Li, P., & Zhao, L. (2023). A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, *22*(1), 48. https://doi.org/10.1186/s12938-023-01113-y

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Qin, H., Zhang, L., & Guo, Q. (2023). Computer-aided diagnosis system for breast ultrasound reports generation and classification method based on deep learning. *Applied Sciences*, *13*(11), 6577. https://doi.org/10.3390/app13116577

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. 2103.00020v1.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. 1910.10683v4.

Roy, S., & Maji, P. (2020). Medical image segmentation by partitioning spatially constrained fuzzy approximation spaces. *IEEE Transactions on Fuzzy Systems*, *28*(5), 965–977. https://doi.org/10.1109/TFUZZ.2020.2965896

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, *31*.

Sharma, D., Dhiman, C., & Kumar, D. (2023). Evolution of visual data captioning methods, datasets, and evaluation metrics: a comprehensive survey. *Expert Systems with Applications*, *221*, 119773. https://doi.org/10.1016/j.eswa.2023.119773

Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., & Shen, H. T. (2019). From deterministic to generative: multimodal stochastic RNNs for video captioning. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(10), 3047–3058. https://doi.org/10.1109/TNNLS.2018.2851077

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B. et al. (2025). Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: a neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3156–3164). Boston, MA, USA: IEEE. https://doi.org/10.1109/CVPR.2015.7298935

Wang, N., Xie, J., Luo, H., Cheng, Q., Wu, J., Jia, M., & Li, L. (2023). Efficient image captioning for edge devices. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(2), 2608–2616. https://doi.org/10.1609/aaai.v37i2.25359

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. 2409.12191v2.

Xu, R., Zhao, H., Li, Z.-Y., & Wang, C.-D. (2023). ALGCN: Accelerated light graph convolution network for recommendation. In *Database systems for advanced applications: 28th international conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, proceedings, Part II* (pp. 221–236). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-031-30672-3_15

Yang, S., Niu, J., Wu, J., Wang, Y., Liu, X., & Li, Q. (2021). Automatic ultrasound image report generation with adaptive multimodal attention mechanism. *Neurocomputing*, *427*, 40–49.

Yang, Y., Teo, C., Daumé III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 444–454). Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models.

Yoon, Y. H., Khan, S., Huh, J., & Ye, J. C. (2019). Efficient B-Mode ultrasound image reconstruction from sub-sampled RF data using deep learning. *IEEE Transactions on Medical Imaging*, *38*(2), 325–336. https://doi.org/10.1109/TMI.2018.2864821

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 11941–11952). Paris, France: IEEE. https://doi.org/10.1109/ICCV51070.2023.01100