



Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning

Danyang Gao^{b,1}, Ming Kong^{a,1}, Yongrui Zhao^b, Jing Huang^a, Zhengxing Huang^a, Kun Kuang^a, Fei Wu^a, Qiang Zhu^{a,*}

^a College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

^b Computer School, Beijing Information Science and Technology University, Beijing 100005, China

ARTICLE INFO

Keywords:

Medical report generation
Semantic query
Transformer
Computer-aided diagnosis
Deep learning

ABSTRACT

Medical report generation can be treated as a process of doctors' observing, understanding, and describing images from different perspectives. Following this process, this paper innovatively proposes a Transformer-based Semantic Query learning paradigm (TranSQ). Briefly, this paradigm is to learn an intention embedding set and make a semantic query to the visual features, generate intent-compliant sentence candidates, and form a coherent report. We apply a bipartite matching mechanism during training to realize the dynamic correspondence between the intention embeddings and the sentences to induct medical concepts into the observation intentions. Experimental results on two major radiology reporting datasets (i.e., IU X-ray and MIMIC-CXR) demonstrate that our model outperforms state-of-the-art models regarding generation effectiveness and clinical efficacy. In addition, comprehensive ablation experiments fully validate the TranSQ model's innovation and interpretation. The code is available at <https://github.com/zjukongming/TranSQ>.

1. Introduction

Medical imaging technology has been widely used in various diagnosis and treatment scenarios, which helps doctors determine the cause of patients more accurately and conveniently and formulate treatment plans. Composing an accurate and comprehensive medical report is a skillful job that requires sufficient medical knowledge and extensive diagnostic experience. Facing onerous diagnosis needs, writing reports takes up a lot of the energy of the physicians (about 10 min or more on average) (Yang et al., 2022). The experience gap among doctors may result in the misinterpretation or disregard of abnormal medical imaging findings, thereby influencing the reliability of diagnostic conclusions. Therefore, devising an efficient and accurate report-generation approach holds significant practical implications for enhancing the work efficiency and service quality of medical professionals, and has emerged as a critical research focus in the realm of computer-aided diagnosis and treatment in recent times.

We can describe the thinking logic for a doctor to write medical reports by repeating the following steps:

- (1) **Formulating Intentions:** First based on knowledge and experience, the doctor forms diagnostic intentions such as: "Is the patient in a postoperative state (Post Surgery Status)?" or "Is the patient suffering from atelectasis?"

- (2) **Understanding Visual Properties:** Based on each intention, observe the relevant area of the image, analyze and summarize according to its visual features to form a diagnostic cognition;
- (3) **Composing Descriptions:** Compose a textual description of the observation results.

The medical report generation task is similar to the Image Caption (Xu et al., 2015; Huang et al., 2019; Zhou et al., 2020), but with essential differences: Firstly, the medical reports are significantly longer than the single-sentence descriptions, which requires the model to generate coherent long texts that conform to the doctors' thinking logic; Secondly, the generation of medical reports have higher requirements for cross-modal information interaction, and the reasonable correlation between diagnostic intention and observation content can more effectively connect visual and text modalities. The typical report generation approach is to model the above process into a state transition process: simulate changes in observation intentions or describe topics in a specific way, generate corresponding text descriptions according to the current state, and then continuously form new observation intentions based on prior knowledge or generated context content until the image is fully described.

* Corresponding author.

E-mail address: zhuq@zju.edu.cn (Q. Zhu).

¹ These authors contributed equally to this article.

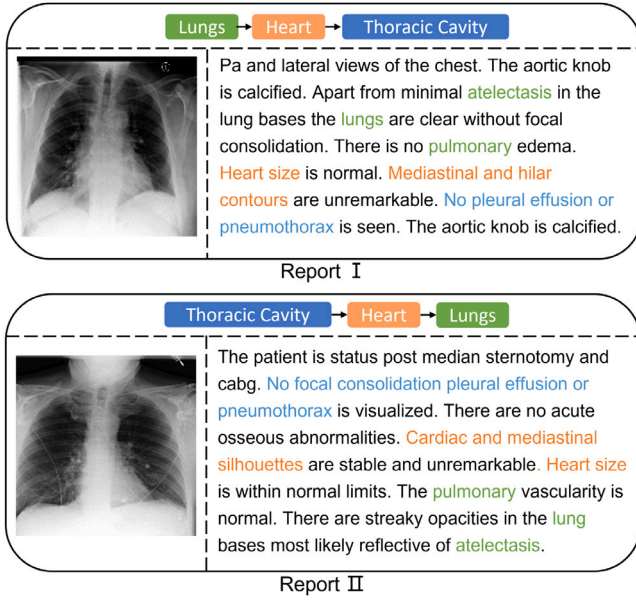


Fig. 1. Difficulties and limitations of the observation intentions transfer learning paradigm.

Most of the early works (Jing et al., 2018; Li et al., 2018) used recurrent neural networks to achieve state transition with a hierarchical structure to alleviate the long-text dependency problem (Li et al., 2018; Nooralahzadeh et al., 2021). However, such methods usually need to introduce artificially extracted templates or prior medical knowledge to prompt the state transition process, which consumes a lot of human costs. Moreover, the hand-crafted prior knowledge also limits the learning ability for latent topics. With the proposal of the Transformer-based language generation models, researchers apply them to the medical report generation field (Alfarghaly et al., 2021). To better model word-level state transition patterns, some works introduce memory modules or medical knowledge graphs to guide the state transition process (Chen et al., 2020; Wang et al., 2022b), or to unify the visual and textual modalities into a shared latent space for the cross-modal state-transition inference (Liu et al., 2021b; Wang et al., 2022c). These works still follow the state transition paradigm. The core assumption is that the doctors' observation sequence of images and thinking logic is serial and consistent, i.e., to deduce the next observation intention from the completed observations.

However, with a thorough study of the medical reports, we found that doctors' diagnostic intentions do not show a clear sequential dependence. As shown in Fig. 1, the medical images associated with the two medical reports exhibit similar visual features, while the doctor's observation intention orders are inconsistent. Report I first observes the lung area, then the heart area, and finally the thoracic cavity, while Report II follows the order of thoracic cavity, heart, and lung. Moreover, although both reports have similar normality/abnormality, such as a normal-sized heart and mild atelectasis, the sentence patterns and wording are also quite different. These findings pose a massive challenge to learn the intention transformation patterns from the weakly labeled reports (i.e., without sentence-level/topic-level labels): on the one hand, the transfer plausibility of observation intentions in medical reports is difficult to define; on the other hand, the inherent differences in the thinking logic of different doctors also make it extremely difficult to learn transfer logic from massive collected medical reports.

Based on the above considerations, this paper proposes a Transformer-based Semantic Query model (TranSQ) to break the modeling paradigm of most existing works based on the state transition. TranSQ learns a set of intention query embeddings to perform semantic queries

on visual features to generate candidate text sets. Specifically, this method can be divided into three steps: (1) to encode the visual features of the input images; (2) to query the visual features through intention embedding to generate semantic features; and (3) to generate a set of candidate sentences from the semantic features, select and sort them to compose the medical report. In particular, this paper conducts a bipartite matching strategy to achieve the dynamic correspondence between intention embeddings and ground-truth sentences during the training process to induct the medical concepts automatically. Moreover, we take further optimization to the method from various aspects, such as multi-view image inputs and retrieval/generation hybrid sentence generation strategy.

On two widely-known MRG benchmarks, IU X-ray and MIMIC-CXR, the TranSQ achieves state-of-the-art results for natural linguistic generation (NLG) and clinical efficacy (CE) metrics, proving the superiority of our proposed approach. Besides, extensive ablation study and visualization results demonstrate its excellent interpretability, reliability, and practicality.

In summary, the major contributions of this work are as follows:

- We conduct a comprehensive rethinking and discussion on medical report generation and propose a novel Transformer-based semantic query model different from the typical intention transition paradigm.
- We propose a bipartite matching-based intention embedding learning strategy to automatically induce medical terminology and description habits without introducing prior knowledge.
- The proposed TranSQ model achieves state-of-the-art performance for both natural linguistic generation and clinical efficacy metrics on two well-known MRG benchmarks.
- The proposed TranSQ model provides candidate sentences from different sentence generation strategies and accurate sentence-level interpretation information, which has the potential to be applied to a computer-aided diagnostic process.

Compared with our conference version (Kong et al., 2022), firstly, we rethink and reorganize the current medical report generation work and our motivation (Section 1). Secondly, we further improve the proposed model, adding the design for multi-view image inputs (Section 3.1) to improve the model performance of the original version. We also introduce the generation model-based sentence generation strategy (Section 3.3) as the alternative strategy, which has potential advantages in specific scenarios. Finally, we extend the experiments, including adding a more detailed ablation study about the key hyperparameters and module designs (Sections 4.5.1–4.5.3), analyzing the retrieval/generation-based report generation strategies (Section 4.5.4), and the effectiveness of introducing domain pre-trained models (Section 4.5.5). We also supplement the correspondence analysis between intention embeddings and medical terminologies (Section 4.6).

2. Related work

The medical report generation task has been derived from the image captioning task, so most early works adopt the CNN-RNN architecture (You et al., 2016; Liang et al., 2017; Gale et al., 2019; Zhang et al., 2020; Gajbhiye et al., 2020). Afterward, some works applied Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the visual extractor for its superior performance and incorporate prior knowledge to improve the visual feature extraction (Liu et al., 2021a; You et al., 2021). Despite these advancements in visual understanding, how to handle long-text dependency effectively remains the most crucial problem.

Many works propose innovative solutions by modeling a state transition approach to deal with this challenge. We re-organize these works from the perspective of observation intention transfer and observation content description. In general, the previous works can be categorized into branches: topic-level and word-level intention granularity.

The topic-level intention state transition comprises the medical report generation into two stages. The first stage is to generate the current observation and description topics with temporal encoding models (Xue et al., 2018; Yuan et al., 2019), such as RNN (Elman, 1990), LSTM (Hochreiter and Schmidhuber, 1997), and Transformer (Vaswani et al., 2017); the second stage is to generate textual descriptions based on the current topic state (Jing et al., 2020) with linguistic retrieval or generation strategies (Ma et al., 2018; Han et al., 2018; Harzig et al., 2019). For example, Jing et al. (2018) fed the feature of fusing the visual and medical label information into the hierarchical LSTM model to generate the medical report, including two LSTM networks. The Sentence LSTM was for the topic-level intention state prediction, while the Word LSTM was for generating a sentence word-by-word based on the current topic-level intention. Wang et al. (2021) proposed another hierarchical LSTM framework to enhance the MRG task with image-report pair matching. Wang et al. (2022a) introduced Task Distillation Module that employed a knowledge graph to group the descriptions to several entities for each structure and generated sentences based on the task-specific entities. Due to the difficulties of training a strong sentence generator, some works considered retrieval of a database to get the description sentences (Sun et al., 2019; Endo et al., 2021; Yang et al., 2021). Ni et al. (2020) employed a cross-modal retrieval technique, which leveraged the correspondence between medical images and the visual and observation intentions in the report to retrieve relevant abnormal representations from the dataset. This retrieval method was then utilized to measure the similarity between the image and the report sentence in the database. On the other hand, Han et al. (2018) proposed a symbol synthesis module that utilized prior knowledge to convert the observation intention vector into a dictionary, which was used in conjunction with a pre-written rule base to rewrite the template. Other works proposed hybrid models that combined both generation and retrieval strategies (Li et al., 2018; Biswal et al., 2020; Li et al., 2023). Li et al. (2018) proposed a hybrid model incorporating reinforcement learning in the network, in which the agent determined whether to generate the current sentence through retrieval or network generation and used CIDEr score as a reward to train the agent. Although these methods promote the progress of the MRG problem, the topic-level intention transition usually relies on additional knowledge or labeling, which incurs extra costs of human labeling.

According to the long-text dependency problem, the CNN-RNN-based methods are hard to generate the entire report word-by-word with a word-level intention transition (Wang et al., 2018; Ma et al., 2021). With the transformer-based text generation model proposed, the above problem is relieved, and many word-level intention transition methods with Transformer for the MRG problem are proposed (Yang et al., 2022). Alfarghaly et al. (2021) proposed a transformer-based model with pre-training with GPT (Brown et al., 2020) to learn word-level state transition patterns and applied an auxiliary task similar to that in Vinyals et al. (2015). This work demonstrated the effectiveness of the approach in word-level state transition. Wang et al. (2022b) employed three loss functions, namely Term-Weighted Report Generation Loss, Temporal-Weighted ITM Loss, and Multi-Label Classification Loss, to force the model generate accurate medical term and sentences that closely aligned with the input medical images. Liu et al. (2021a) introduced prior domain knowledge, such as graph structure and disease entities, to enhance the accuracy of generated reports. Chen et al. (2020) introduced the Relational Memory module in Transformer to capture long text dependencies, focusing on templating and long text characteristics. MSAT (Wang et al., 2022c) introduced a Memory-augmented Sparse Attention block to capture the higher-order interactions between the input fine-grained image features, optimizing report generation through reinforcement learning. Although the transformer-based model helps alleviate the challenge of long-text dependency, instability in long-range text generation still suffers the solutions. Besides, most of these methods keep relying on human-defined prior knowledge. Interpretation is another problem. Word-level

intention transition methods just provide word-level interpretations, which are hard to correlate with the intuitive diagnostic concern.

Considering the ambiguity and uncertainty of the current state transition-based methods, we propose a novel approach different from the typical intention state transition. Specifically, we advocate directly predicting a set of intention embeddings to generate description candidates and selecting the most appropriate ones to compose the report. This innovative approach circumvents the need to model intention state transitions and significantly reduces the uncertainty in the report generation process.

3. Method

According to the doctors' thinking logic mentioned in the Introduction, the process of writing a medical report can be divided into three main tasks:

- (1) **Visual properties Understanding**, to observe and understand the visual representation of images;
- (2) **Intention-based observation**, to form the observation intention of the images and find the corresponding visual properties and their region from the image;
- (3) **Description generation**, to describe the observation results into sentences to form the diagnosis report.

Corresponding to the above task decomposing, as shown in Fig. 2, the proposed TranSQ model includes three modules: visual extractor, semantic encoder, and report generator.

3.1. Visual extractor

First, we need to extract and encode the visual features of the input images, i.e., to convert inputs into a set of visual feature sequences that describe the local features of the image. Supposing the inputs contain M images, denoted as $X = \{X_1, X_2, \dots, X_M\}$, $X_i \in \mathbb{R}^{C \times H \times W}$, we resize all the images and divide them into a series of $P \times P$ sized patches and map them with a linear projection. Thus, the input image is converted into a patch embedding sequence with a length of $N = M \times \frac{H \times W}{P^2}$. Next, a set of learnable spatial position embeddings F_{pos}^i and type embeddings $F_{type}^{m_i}$ with the same dimensions are added to the patch embeddings to distinguish their positions and sources. Spatial position embeddings are used to express the correlation between the local regions of the same image, while type embeddings represent the relationship of different images and strengthen the correlation of the visual features of the same image. Therefore, the input sequence of the visual encoder can be expressed as $z_0 = [z_0^1, \dots, z_0^N]$, where the visual feature z_0^i corresponding to each patch can be recorded as:

$$z_0^i = x_i E_v + F_{pos}^i + F_{type}^{m_i}, i = 1, 2, \dots, N \quad (1)$$

where $E_v \in \mathbb{R}^{(P^2 \cdot C) \times D}$ is a linear projection, $F_{pos}, F_{type} \in \mathbb{R}^{N \times D}$ are the spatial position embeddings and type embeddings, respectively.

In particular, we try two definitions of type embeddings: order-based and view-based. Order-based embeddings simply distinguish the source images of different patches, and view-based embeddings are to distinguish the view direction of the image. According to the experiment, we found that the combination of two kinds of type embeddings achieves the best results, i.e., $F_{type}^m = F_{view}^m + F_{order}^m$. We provide a detailed discussion on this in the experimental Section 4.5.2.

Considering the need to analyze multiple images, we adopt a visual encoder based on ViT (Dosovitskiy et al., 2021) for its outstanding multi-modal adaptation, which contains of L_V transformer encoder layers. Each layer consists of a multi-head self-attention block and a multi-layer perceptron. The calculating process of the visual extractor is as follows:

$$\begin{aligned} z_l' &= MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1, \dots, L \\ z_l &= MLP(LN(z_l')) + z_l', \quad l = 1, \dots, L \end{aligned} \quad (2)$$

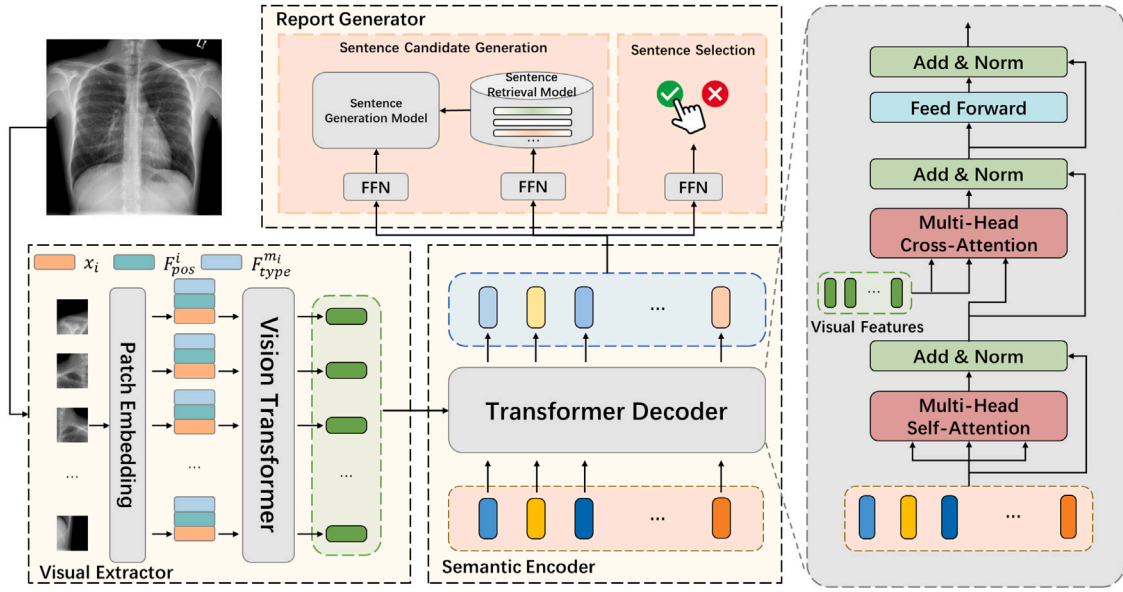


Fig. 2. The framework overview of TranSQ model consists of three major modules: Visual extractor, semantic encoder, and report generator.

where MSA is a multi-head self-attention layer, MLP is a multi-layer perceptron, and LN is a layer normalization. The output of the last layer is the visual feature sequence f^v , i.e., $f^v = z_L$.

3.2. Semantic embedding

The primary task of medical report generation is to observe images based on a specific intention to form a semantic representation of the findings. Thus, the model can establish the relevance between the visual features of the local region in the image and the semantics of the sentence. Unlike related works that model the state transition patterns, this paper innovatively proposes a Transformer-based semantic encoder, which makes semantic queries to the visual features with the intention embeddings to generate the semantic features.

Specifically, given a set of learnable observation intention embeddings $\mathbf{q} = \{q_1, \dots, q_K\}$, each of which corresponds to a specific implicit intention for medical image observation. By querying the extracted visual feature sequence f^v , the critical visual information related to the observation intention is transformed into the semantic domain, generating semantic features $f^S = \{f_1^S, f_2^S, \dots, f_K^S\}$.

The semantic encoder module consists of L_S -layers of transformer encoder blocks. Each block contains a *multi-head self-attention layer* (MSA), a *multi-head cross-attention layer* (MCA), and a *multi-layer perceptron* (MLP). MCA extracts the visual features related to the observation intention by calculating the attention weight between the visual features and the intention embeddings. Simultaneously, MSA realizes the fusion of observation information according to the association and difference between observation intention representations. The semantic embedding process can be summarized as follows:

$$\begin{aligned} h_0 &= \mathbf{q}E_q \\ h_l' &= MSA(LN(h_{l-1})) + h_{l-1}, \quad l = 1, \dots, L \\ h_l'' &= MCA(LN(f^v), LN(h_l')) + h_l', \quad l = 1, \dots, L \\ h_l &= MLP(LN(h_l'')) + h_l'', \quad l = 1, \dots, L \\ f^S &= LN(h_L) \end{aligned} \quad (3)$$

where $E_q \in \mathbb{R}^{K \times D}$ is a linear projection and LN is a layer normalization layer.

Noting that although the relevance and co-occurrence between different intention embeddings are considered, the generation of semantic features is only related to the intention representation without the sentence order.

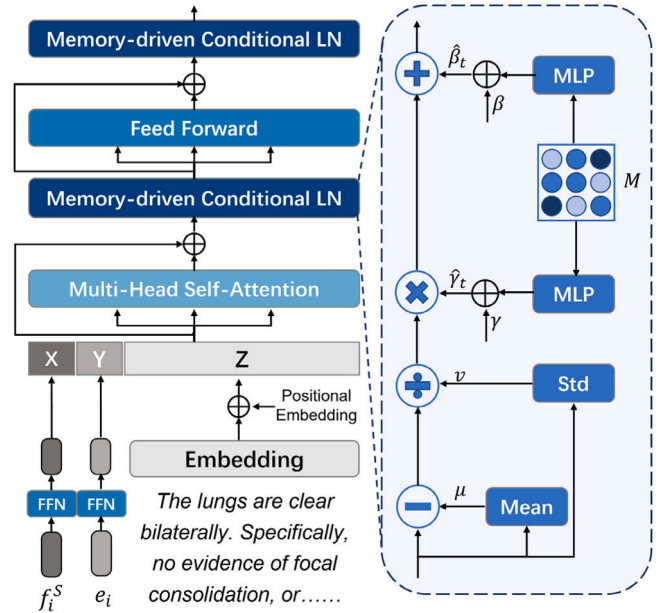


Fig. 3. The framework of the sentence generation module, where μ and v correspond with the mean and standard deviation for input normalization, β and γ are the mean and standard deviation for memory normalization.

3.3. Report generator

The semantic feature set f^S denotes the semantic encoding outcomes pertinent to the intentions. The ensuing process involves transforming the semantic features into sentences and selecting and sorting the most valuable sentences to compose the medical report. Precisely, this module encompasses the fulfillment of three sub-tasks: text generation, text selection, and text sorting. We will now elaborate on the specific realization of these sub-tasks.

3.3.1. Text generation

To obtain a collection of candidate sentences, we need to transform the semantic features f^S in the set into sentences. Generally, we can consider two strategies: retrieval-based and generation-based.

Retrieval-based Strategy: In the context of the retrieval strategy, a straightforward approach is to retrieve the semantically closest sentence from a pre-built dataset of sentences using a similarity measure. Firstly, we employ a pre-trained sentence embedding model to semantically encode all the sentences \hat{y}_i in the training set, obtaining corresponding sentence vectors. Thus, the sentence dataset $D = \{(v_1, v_1), \dots, (v_{N_D}, v_{N_D})\}$ contains N_D text-sentence vector pairs. For each semantic vector in f^S , we use a feedforward network to obtain a set of candidate sentence vectors $e \in \mathbb{R}^{K \times D}$. Then, for each sentence vector e_i in e , we can measure its similarity with each sentence vector v_j in the descriptive sentence set D based on cosine similarity:

$$\text{sim}(v_i, e_j) = \frac{v_i \cdot e_j}{\|v_i\| \cdot \|e_j\|} \quad (4)$$

For each predicted sentence vector e_i , the most similar sentence in terms of cosine similarity can be retrieved from the description sentence set D , forming a set of predicted candidate sentences $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K\}$.

Generation-based Strategy: Given the standardized and structured nature of medical reports, a retrieval-based strategy can generally meet the requirements for generating textual descriptions. However, the quality of the generated descriptions using a retrieval-based strategy is highly dependent on the quality and diversity of the sentences set D , and it cannot effectively handle situations that require flexible expressions, such as quantitative descriptions. Therefore, we adopt a conditional generation model to generate candidate sentences as an upgrade and extension to the retrieval-based strategy.

We employ a memory-driven Transformer-based text generation model proposed by R2Gen (Chen et al., 2020), as shown in Fig. 3. Specifically, based on the standard DistilGPT (Sanh et al., 2019) architecture, we replace the standard normalization in the Transformer Decoder with a Memory-driven Conditional Layer Normalization to enhance the modeling of common structured descriptions in medical reports.

For the i th observed intention query, we perform linear transformations on the semantic features f_i^S and predicted sentence vectors e_i using a fully connected network, which serves as the conditional input for the text generation model to generate descriptive sentences word by word. The visual features relevant to the observation intention query q_i are integrated into f_i^S , and the sentence vector e_i provides the pattern information of the text modality.

3.3.2. Text selection

The text generation component produces a candidate set consisting of K sentences. However, not all information is worth presenting in a medical report. Therefore, we use a simple linear projector to predict the selection probability of candidate sentences \hat{p}_i based on the semantic vector f_i^S :

$$\hat{p}_i = \text{MLP}(f_i^S) \quad (5)$$

where MLP represents a multilayer perceptron. Our model determines the candidate sentence selection by a predefined threshold based on the selection probability of the predicted observation intention.

3.3.3. Text sorting

It is observed that sentence occurrence orders have great uncertainty and can be affected by various factors, such as different doctors' writing habits. Therefore, this paper proposes a simple and intuitive way to sort sentences: We calculate the average position of ground truth sentences matched by each observation intention query in the training set for the entire report. This average position is used as a reference to rank the candidate sentences. This simple strategy ensures that the medical concepts mentioned in the report roughly conform to recognized observation and description habits.

3.4. Model training

3.4.1. Intention query-sentence correspondence

Based on the previous discussion, this paper proposes a method of generating sentences with intention embeddings. However, the training data does not explicitly label the observation intentions and medical terminologies implied by the ground-truth sentences. In other words, the intention embedding that the sentence should correspond to is not specified, and it is not clear which text candidates of intention embedding should be selected.

To address this issue, a dynamic decision-making strategy is adopted to establish the correspondence between observation intentions and ground truth sentences based on a binary matching approach.

It should be noted that we aim to ensure that the intention embedding set contains all possible observation intentions as much as possible, in order to correspond to different types of medical image representations and abnormal regions. Therefore, the size of the intention embedding set K is much larger than the number of sentences contained in a single medical report. The training task can be described as follows: define the set of sentences in the ground truth medical report as y , and the model needs to select the best subset from the prediction set $\hat{y} = \{\hat{y}_i\}_{i=1}^K$, so that the sentences in the subset are most similar to the sentences in the ground truth set. Assuming that the ground truth set is padded with ϕ to the same size as the prediction set, the selection problem of the optimal subset can be transformed into a bipartite graph matching problem. In other words, we are searching for the matching strategy $\sigma \in \Theta_K$ of K elements to find the optimal strategy that minimizes the corresponding matching loss:

$$\hat{\sigma} = \underset{\sigma \in \Theta_K}{\text{argmin}} \sum_i^K L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \quad (6)$$

where $L_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is the matching loss of ground truth y_i and $\hat{y}_{\sigma(i)}$, and it is calculated as:

$$L_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = \mathbb{1}_{y_i \neq \phi} L_{\text{sim}}(v_i, \hat{v}_{\sigma(i)}) - \mu \mathbb{1}_{y_i \neq \phi} \hat{p}_{\sigma(i)} \quad (7)$$

$$L_{\text{sim}} = 1 - \frac{v_i \cdot \hat{v}_{\sigma(i)}}{\|v_i\| \cdot \|\hat{v}_{\sigma(i)}\|}$$

The calculation of the matching loss involves two aspects as shown in the formula: on the one hand, the similarity loss L_{sim} describes whether the candidate sentence vector $\hat{v}_{\sigma(i)}$ is semantically similar to the ground truth sentence vector v_i ; on the other hand, the selection probability of the candidate sentence vector $\hat{p}_{\sigma(i)}$ determines whether the sentence should be selected or not. μ is the balance coefficient between semantic similarity and candidate text selection probability. The optimal matching of the model based on the matching loss can be efficiently obtained using the Hungarian algorithm.

3.4.2. Optimization

During training, the model is optimized based on the optimal matching results obtained through the aforementioned matching strategy. The final training loss is calculated by weighting the semantic consistent loss L_{sim} of the matching results and the text selection loss L_{cls} :

$$L(y, \hat{y}) = \sum_i^K \mathbb{1}_{y_i \neq \phi} L_{\text{sim}}(v_i, \hat{v}_{\hat{\sigma}(i)}) - \lambda \mathbb{1}_{y_i \neq \phi} L_{\text{cls}}(c_i, \hat{p}_{\hat{\sigma}(i)}) \quad (8)$$

where λ is the balance coefficient between semantic consistent loss and text selection loss, and $c_{\sigma(i)}$ is the text selection label. Specifically, we assign a text selection label to each predicted sentence in the candidate set according to the binary matching result. For a prediction sentence $\hat{y}_{\sigma(i)}$, if its corresponding ground truth sentence $y_i \neq \phi$, its corresponding text selection label $c_{\sigma(i)} = 1$, and vice versa $c_{\sigma(i)} = 0$.

It is noteworthy that the model utilizes a matching strategy that generates sentences for the observation intention query that are closer in meaning to the ground truth sentence while simultaneously increasing the probability of the candidate sentence being selected. At the

early stage of training, the matching relationship between the ground truth sentence and the observation intention query is nearly random. However, as the training process progresses, the correspondence between the ground truth sentence and the observation intention query gradually becomes certain, allowing each observation intention query to align with the visual query target and text description content that possess clear medical significance.

Subsequently, we explicate the semantic consistent loss L_{sim} and text selection loss L_{cls} in detail:

Semantic Consistent Loss: By calculating the semantic similarity between the predicted sentence vector and the ground truth sentence, the model gradually generates sentence vectors that contain precise semantic information and are more similar to the ground truth.

Specifically, the cosine similarity between the ground truth sentence vector v_i and the predicted sentence vector $\hat{v}_{\hat{\sigma}(i)}$ is used as the evaluation metric for semantic consistent. As shown in the formula, when the cosine similarity value is closer to 1, indicating that the two sentence vectors are more similar, the corresponding loss value is smaller.

$$L_{sim} = \frac{1}{k} \sum_{i=1}^K \mathbb{1}_{y_i \neq \emptyset} \left(1 - \frac{v_i \cdot \hat{v}_{\hat{\sigma}(i)}}{\|v_i\| \cdot \|\hat{v}_{\hat{\sigma}(i)}\|} \right) \quad (9)$$

Text Selection Loss: By calculating the text selection loss between candidate sentence labels and ground-truth sentence labels, the model gradually and accurately selects the predicted sentence vector that corresponds to the current sample, thereby optimizing the process of sentence vector selection.

As label selection dynamically adjusts with matching results during training, we use a multi-label classification loss called DB loss (Wu et al., 2020) to address the long-tail distribution of labels, considering that text descriptions corresponding to different intentions appear with significant frequency differences in medical reports.

Specifically, for K intention queries, assuming each category i contains n_i samples, the sample's label-containing probabilities need to be re-weighted by sampling. For a specific sample x , the actual sampling probability is the sum of the contribution from each category, i.e:

$$P^I(x) = \frac{1}{K} \sum_{y_k=1}^K \frac{1}{n_i} \quad (10)$$

The contribution of each category to the sampling rate can then be represented by the ratio of $P_i^K(x)$ and $P^I(x)$ as a weighting factor for the sample sampling. A smoothing function is also introduced to map the weighting coefficients to a reasonable range, i.e:

$$\hat{r}_i = \alpha^{DB} + \frac{1}{1 + \exp(-\beta^{DB} \times (\frac{P_i^K(x)}{P^I(x)} - \mu^{DB}))} \quad (11)$$

where α^{DB} , β^{DB} and μ^{DB} are hyperparameters.

It is worth noting that for general problems, the number of samples in each category n_i can be calculated by statistical analysis of the distribution of samples in the training set. However, in this task, the number of samples corresponding to each intention query is unknown and variable. To address this issue, a dynamic statistical strategy is proposed in this paper to update the correspondence between sentences and intentions. At the end of each training epoch, the number of samples corresponding to each intention query is periodically updated. As the training process progresses, the sample distribution gradually stabilizes.

4. Experiment

4.1. Datasets

To evaluate the effectiveness of TranSQ, we make comprehensive experiments on two well-known medical report generation benchmarks, IU X-RAY and MIMIC-CXR.

IU X-RAY (Demner-Fushman et al., 2016) is a Chest X-ray dataset collected and organized by Indiana University in the United States. It includes 7470 medical images and 3955 radiology reports, in which each medical report strictly corresponds to two Chest X-ray images. Consistent with previous work (Li et al., 2018, 2019; Chen et al., 2020; Jing et al., 2020), the dataset is split into training, validation, and testing sets with a ratio of 7:1:2.

MIMIC-CXR (Johnson et al., 2019) contains a total of 65,379 medical images and corresponding reports of patients, including 377,110 chest X-ray images and 227,835 radiology reports. Each report may correspond to one or more images. According to the official division, the training set contains 222,758 reports and 368,960 medical images, the validation set contains 2991 medical images corresponding to 1808 reports, as well as the test set contains 5159 medical images and 3269 reports.

For all the experiments, we follow the conventional method of splitting the datasets in related work (Li et al., 2018). The private information of all patients in the medical reports in the two datasets has been desensitized during data sorting, and the reports in both datasets are lower-cased. The words that occur less than the pre-defined threshold are filtered out and replaced with a [unk] token.

4.2. Baselines and evaluation metrics

To verify the language generation accuracy and clinical efficacy of the TranSQ model, we introduce representative methods for the performance comparison, including the retrieval-related methods, such as HRGR (Li et al., 2018), KERP (Li et al., 2019), CCR (Liu et al., 2019), MedWriter (Yang et al., 2021), PPKED (Liu et al., 2021a), KnowMT (Yang et al., 2022), and DCL (Li et al., 2023), as well as the generation-based methods, such as CoAtt (Jing et al., 2018), M^2 Trans (Cornia et al., 2020), CMAS-RL (Jing et al., 2020), GDGPT (Alfarghaly et al., 2021), Transformer (Chen et al., 2020), CMCL (Liu et al., 2021b), R2Gen (Chen et al., 2020), AlignTransformer (You et al., 2021), KGAE (Liu et al., 2021b), ITA (Wang et al., 2022a), Multicriteria (Wang et al., 2022b), and MSAT (Wang et al., 2022c) (MSAT-RL represents the MSAT method with Reinforcement Learning). In addition, to prove that the performance advantage of TranSQ is mainly due to the improvement of the accuracy of single-sentence semantic rather than ordering strategy, we provide the results of TranSQ with the randomly sorted sentence orders, namely TranSQ-RS.

To evaluate the models' natural language generation (NLG) ability, we follow the principles of existing work and the standard evaluation protocol (Chen et al., 2015) and apply the widely-used NLG metrics, i.e., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004) for the evaluation. Furthermore, to fairly evaluate the models' clinical efficacy (CE) of whether the generated reports accurately describe the key medical terminologies, we follow previous works (Chen et al., 2020; Liu et al., 2021b) to use an open-sourced tool named CheXpert (Irvin et al., 2019) labeler to extract the categorization results on 14 medical terminologies from the reports and evaluate the Precision, Recall, and F1-score metrics of the category prediction.

4.3. Experimental settings

We adopt the Vision Transformer model ViT-B/32 pre-trained on ImageNet as the visual encoder, where the input image size is 384×384 , the patch size is 32×32 , and the hidden size is 768. The semantic query module adopts a standard Transformer decoder architecture with 12 layers and 12 attention heads. The size of the hidden states of the module is set to 768. All the retrieval candidates in the sentence gallery are from the training sets, and the sentence embeddings are generated by a pre-trained MPNet model (Reimers and Gurevych, 2019).

Table 1

Comparisons of the TranSQ model with previous studies on IU X-RAY with NLG metrics.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
CoAtt (Jing et al., 2018)	0.455	0.288	0.205	0.154	–	0.369
[†] HRGR (Li et al., 2018)	0.438	0.298	0.208	0.151	–	0.322
[†] KERP (Li et al., 2019)	0.482	0.325	0.226	0.162	–	0.339
\mathcal{M}^2 Trans (Cornia et al., 2020)	0.437	0.290	0.205	0.152	0.176	0.353
CMAS-RL (Jing et al., 2020)	0.464	0.301	0.210	0.154	–	0.362
GDGPT (Alfarghaly et al., 2021)	0.387	0.245	0.166	0.111	0.164	0.289
Transformer (Chen et al., 2020)	0.396	0.254	0.179	0.135	0.164	0.342
CMCL (Liu et al., 2021b)	0.473	0.305	0.217	0.162	0.186	0.378
R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.187	0.371
[†] MedWriter (Yang et al., 2021)	0.471	0.336	0.238	0.166	–	0.382
[†] PPKED (Liu et al., 2021a)	0.483	0.315	0.224	0.168	0.190	0.376
AlignTransformer (You et al., 2021)	0.484	0.313	0.225	0.173	0.204	0.379
KGAE (Liu et al., 2021b)	0.519	0.331	0.235	0.174	0.191	0.376
[†] KnowMT (Yang et al., 2022)	0.496	0.327	0.238	0.178	–	0.381
ITA (Wang et al., 2022a)	0.505	0.340	0.247	0.188	0.208	0.382
[†] DCL (Li et al., 2023)	–	–	–	0.163	0.193	0.383
Multicriteria (Wang et al., 2022b)	0.496	0.319	0.241	0.175	–	0.377
TranSQ-RS	0.516	0.365	0.269	0.198	0.209	0.398
TranSQ	0.516	0.365	0.272	0.205	0.210	0.409

Table 2

Comparisons of the TranSQ model with previous studies on MIMIC-CXR with respect to NLG and CE metrics.

Method	NLG metrics						CE metrics		
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	Precision	Recall	F1-score
[†] CCR (Liu et al., 2019)	0.313	0.206	0.146	0.103	–	0.306	–	–	–
\mathcal{M}^2 Trans (Cornia et al., 2020)	0.238	0.151	0.102	0.067	0.110	0.249	0.331	0.224	0.228
Transformer (Chen et al., 2020)	0.314	0.192	0.127	0.090	0.125	0.265	0.197	0.145	0.133
CMCL (Liu et al., 2021b)	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
[†] PPKED (Liu et al., 2021a)	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
AlignTransformer (You et al., 2021)	0.378	0.235	0.156	0.112	0.158	0.283	–	–	–
KGAE (Liu et al., 2021b)	0.369	0.231	0.156	0.118	0.153	0.295	0.389	0.362	0.355
[†] KnowMT (Yang et al., 2022)	0.363	0.228	0.156	0.115	–	0.284	0.458	0.348	0.371
ITA (Wang et al., 2022a)	0.395	0.253	0.170	0.121	0.147	0.284	–	–	–
[†] DCL (Li et al., 2023)	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
Multicriteria (Wang et al., 2022b)	0.351	0.223	0.157	0.118	–	0.287	–	–	–
MSAT (Wang et al., 2022c)	0.373	0.235	0.162	0.120	0.143	0.282	–	–	–
MSAT-RL (Wang et al., 2022c)	0.413	0.266	0.186	0.136	0.170	0.298	–	–	–
TranSQ-RS	0.423	0.259	0.168	0.112	0.167	0.263	0.482	0.563	0.519
TranSQ	0.423	0.261	0.171	0.116	0.168	0.286			

We set the size of the intention embedding set K to 50 for the MIMIC-CXR dataset and 25 for the IU X-ray dataset. For both datasets, we set the scale factor of bipartite matching loss $\mu = 0$ and the scale factor of loss function $\lambda = 1$. The hyper-parameters of L_{cls} follow the settings of DB loss (Wu et al., 2020) on the COCO-MLT dataset, where $\alpha^{DB} = 0.1$, $\beta^{DB} = 10$, $\mu^{DB} = 0.2$, $\lambda^{DB} = 0.5$, and $\nu^{DB} = 0.5$.

In the training process of both datasets, the batch size is set to 64, and the optimizer is Adamw (Loshchilov and Hutter, 2017), with a learning rate of $1e-4$ for parameter optimization with linear decay. All the experiments run on an Nvidia A100 GPU.

4.4. Comparison with baselines

We evaluate the model performance and compare it with existing works on IU X-RAY and MIMIC-CXR datasets, where Table 1 reports the results of NLG metrics on the IU X-ray dataset, and Table 2 reports the results of both NLG and CE metrics on the MIMIC-CXR dataset. [†] indicates the retrieval-related methods, while the remaining indicates the generation-based methods. BL- n denotes the BLEU score using up to n -grams, and MTR and RG-L denote METEOR and ROUGE-L respectively.

The results show that TranSQ achieves state-of-the-art results on most of the NLG metrics on both the IU X-RAY dataset and the MIMIC-CXR dataset, with the exception of MSAT-RL, which harnesses the power of reinforcement learning. These results effectively prove the ability of TranSQ to generate accurate report text. Besides the performance of TranSQ-RS, which randomly sorts the selected sentences, the

NLG metrics only show a slight decrease and remain comparable to the baseline methods. This indicates that the outperformance of the TranSQ model comes from the precise predictions of single sentences rather than a well-designed sentence-ordering strategy, which proves the ability of the TranSQ model to generate accurate sentence candidates and make effective selection decisions.

As for the clinical efficacy aspect, TranSQ achieves an astonishing improvement over the state-of-the-art method DCL (Li et al., 2023) in both the Recall and F1-score metrics. The outstanding performance on the clinical efficacy metrics demonstrates the capability of the TranSQ model to generate high-quality descriptions for clinical abnormalities compared to existing models, which again indicates that the model can learn the intention queries related to medical terminologies, especially without introducing human-crafted medical knowledge definitions.

4.5. Ablation experiments

We conducted ablation experiments in various aspects to compare and analyze the effectiveness of key hyperparameters, module design, and strategy selection of the TranSQ model.

4.5.1. Size of intention query set

The size of the intention query set is a major hyperparameter of the TranSQ model. Intuitively, a large number of intent queries can provide rich and refine definitions of observation intentions, as well as diversity and redundancy of semantic expression. However, increasing

Table 3

The results of TranSQ with different observation intention size K on IU X-RAY and MIMIC-CXR.

	K	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-RAY	25	0.516	0.365	0.272	0.205	0.210	0.409
	50	0.504	0.332	0.252	0.191	0.201	0.383
	75	0.455	0.316	0.226	0.164	0.209	0.396
	100	0.468	0.319	0.219	0.151	0.202	0.374
MIMIC-CXR	25	0.403	0.244	0.157	0.106	0.174	0.286
	50	0.423	0.261	0.171	0.116	0.168	0.286
	75	0.383	0.233	0.151	0.103	0.155	0.280
	100	0.330	0.207	0.139	0.097	0.140	0.283

Table 4

The results of TranSQ with different positional embeddings on IU X-RAY.

	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
w/o TypeEmb.	0.491	0.336	0.244	0.181	0.203	0.370
w/ order	0.505	0.357	0.263	0.196	0.204	0.398
w/ view	0.500	0.348	0.256	0.196	0.208	0.411
w/ order & view	0.516	0.365	0.272	0.205	0.210	0.409

Table 5

Comparisons with different visual encoders on IU X-RAY and MIMIC-CXR.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-RAY						
ITA	0.505	0.340	0.247	0.188	0.208	0.382
TranSQ-ResNet	0.479	0.332	0.239	0.176	0.210	0.406
TranSQ	0.516	0.365	0.272	0.205	0.210	0.409
MIMIC-CXR						
ITA	0.395	0.253	0.170	0.121	0.147	0.284
TranSQ-ResNet	0.406	0.245	0.157	0.104	0.163	0.281
TranSQ	0.423	0.261	0.171	0.116	0.168	0.286

Table 6

The results of TranSQ with different report generation strategies on IU X-RAY and MIMIC-CXR.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-RAY						
Generation	0.506	0.355	0.255	0.187	0.210	0.405
Retrieval	0.516	0.365	0.272	0.205	0.210	0.409
Hybrid	0.511	0.363	0.278	0.208	0.209	0.407
MIMIC-CXR						
Generation	0.341	0.221	0.154	0.111	0.140	0.292
Retrieval	0.423	0.261	0.171	0.116	0.168	0.286
Hybrid	0.410	0.258	0.172	0.120	0.164	0.292

Table 7

The impact of pre-trained visual encoders in the medical domain on TranSQ.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
ResNet50-ImageNet	0.479	0.332	0.239	0.176	0.210	0.406
ResNet50-MedKLIP	0.490	0.336	0.241	0.176	0.225	0.405

Table 8

The impact of pre-trained text encoders in the medical domain on TranSQ.

Method	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
biobert-nli	0.405	0.258	0.177	0.124	0.193	0.354
S-BioBERT	0.467	0.314	0.228	0.169	0.209	0.377
MPNet	0.516	0.365	0.272	0.205	0.210	0.409

the number of intention queries also increases the difficulty of sentence candidate selection. Therefore, the size of the intent query set needs to consider a balance between these two factors. We compare the NLG metrics of the models with different numbers of intention queries, denoted as $K = \{25, 50, 75, 100\}$. The results in Table 3 show that as the value of K increases, the NLG performance of the TranSQ model on

the two datasets shows a trend of first increasing and then decreasing, which indicates that an appropriate K can achieve the performance of text generation and selection balance. On the IU X-ray dataset, $K = 25$ achieves the best performance, and on the mimic dataset, $K = 50$ achieves the best performance. These results align with our understanding of the two datasets. Compared to the IU X-ray dataset, the MIMIC-CXR dataset has more diverse and complex sentence patterns and more abundant description contents, necessitating a larger-scaled intention query set.

4.5.2. Type embedding definition

In some cases, a medical report is generated by analyzing multiple images, so it is necessary to add a type embedding visual extractor input to distinguish visual features from different images. This paper considers two ways of type embedding definitions: order-based and view-based. The Order-based type embedding simply represents which image the patch comes from, and the view-based type embedding represents the view of the corresponding image. Both types of encodings are set as learnable parameters. In particular, to obtain the perspective labels (frontal or lateral views) of medical images, we randomly selected 1000 images from the IU X-ray dataset to train a view classifier based on the ResNet-50 (He et al., 2016) model.

Table 4 shows that applying both kinds of type embedding helps improve the model performance. The results indicate that type embedding can effectively enhance the correlation of visual features within the same image and express the relationship between different images and views.

4.5.3. CNN-based visual extractor adaptation

To prove that the improvement of TranSQ mainly comes from the design of the semantic query mechanism rather than the image encoding ability of ViT, we replace the visual extractor of the original model with a ResNet-50, denoted as TranSQ-ResNet. By comparing with the original TranSQ, TranSQ-ResNet, and KGAE, the baseline model with ResNet-50 backbone (see Table 5), we observe a slight decrease in performance when replacing the visual extractor, but it is still comparable to the baseline method. The result indicates that the semantic query strategy is adaptable to a CNN-based visual extractor.

4.5.4. Report generation strategy

We compare and analyze the retrieval and generation strategies for sentence generation, and the results on the two datasets are shown in Table 6. It can be found that the retrieval-based sentence generation still outperforms the generation-based method, which benefits from the highly templated characteristics of medical descriptions.

Furthermore, We try to propose a hybrid strategy that combines the complementary strengths of retrieval and generation methods. Specifically, for the trained TranSQ model, we compare the NLG metrics (e.g., BLEU-4) of the sentences obtained by the retrieval and generation strategies on the validation set to decide the sentence generation strategy corresponding for each intention embedding. The hybrid strategy conforms to the intuition: certain intentions that aim to capture macroscopic observations require stable descriptions, while others corresponding to specific features or speculations need to prioritize flexibility. We find that hybrid strategies outperform retrieval-based text generation strategies on some high-order metrics, which further expands the potential of our method.

4.5.5. Domain pre-trained models integration

There are obvious differences between medical and general images, and medical reports contain a bunch of domain-specific terminologies and acronyms. It is reasonable to speculate that models pre-trained using specific domain data may exhibit superior performance in feature extraction and encoding. Recently, many works have been proposed on visual and linguistic pre-training using medical datasets. We explored the impact of applying these works to the TranSQ model.

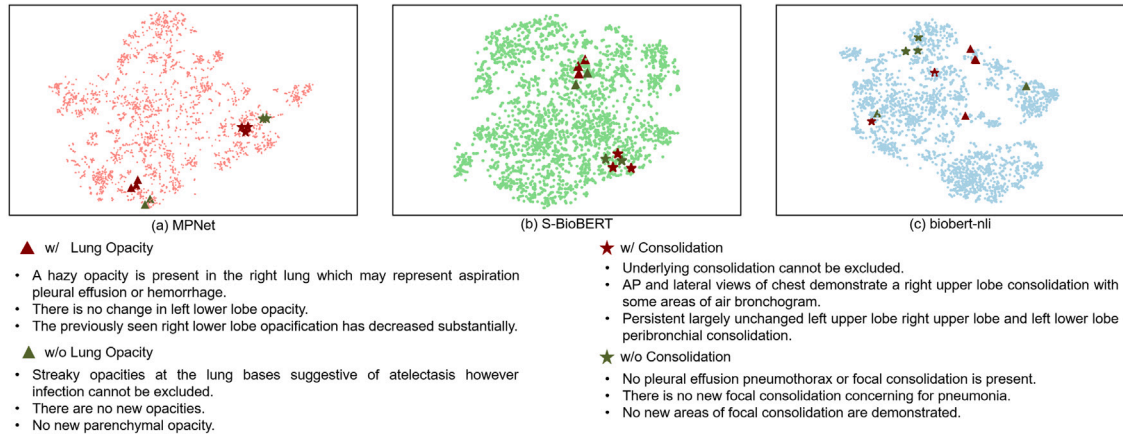


Fig. 4. T-SNE visualization of sentence embeddings generated by three text encoders, where MPNet pre-trained on the general domain datasets, while S-BioBERT and biobert-nli pre-trained on medical domain datasets.

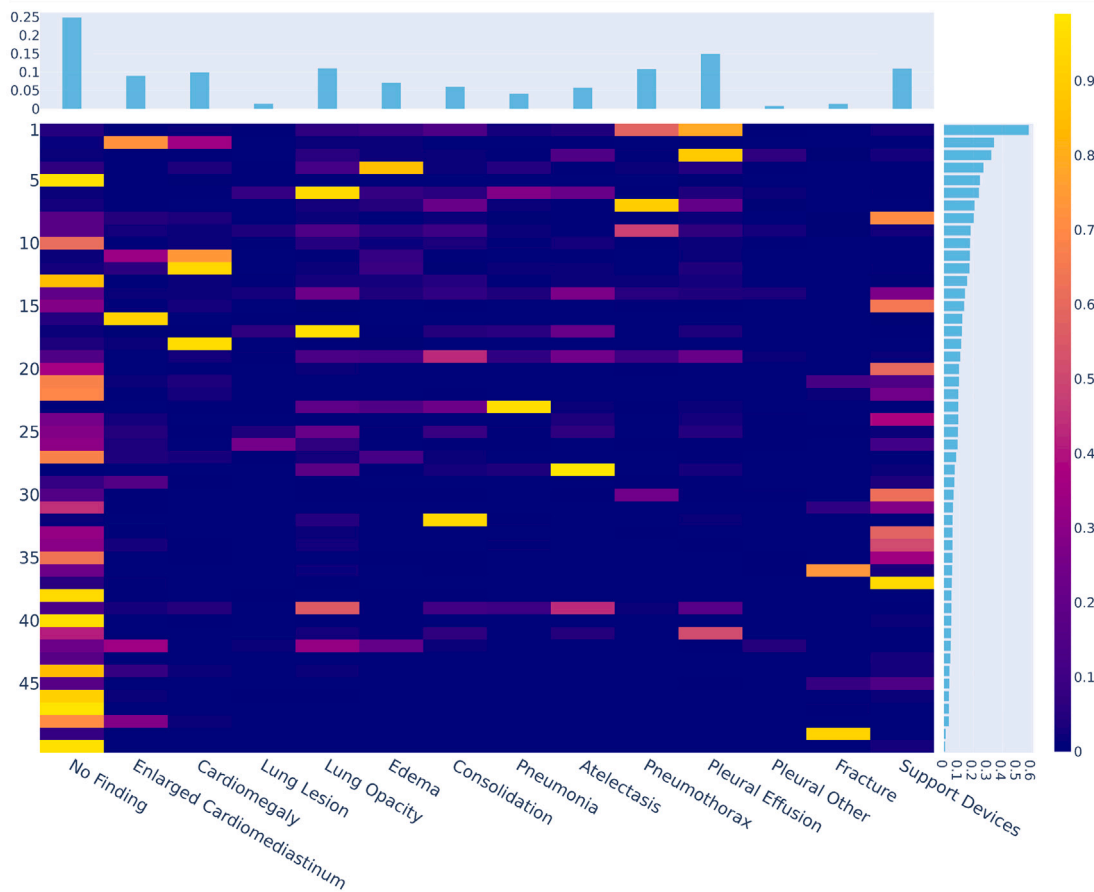


Fig. 5. The correspondence between intention embeddings and medical terms and their co-occurrence frequency.

Visual Encoder We consider MedKLIP, a self-supervised pre-training method for the visual extractor using contrastive vision-language learning from medical domain data. Table 7 shows the NLG performance comparison of the TranSQ model with ImageNet and MedKLIP pre-trained ResNet-50 visual extractors on the IU X-ray dataset. The results show that the MedKLIP pre-trained visual extractor performs better on most NLG indicators, proving that a specific domain pre-trained visual extractor is more effective in generating accurate medical reports.

Text encoder We apply two well-known text encoding models pre-trained on medical datasets, namely S-BioBERT, and biobest-nli, to

generate TranSQ-required sentence embeddings. We hope the models in the specific domain can encode medical terminologies and acronyms more accurately and improve model performance. However, as shown in Table 8, the sentence embeddings generated by S-BioBERT and biobert-nli did not improve performance compared to the MPNet pre-trained on general data, which is contrary to our initial expectations.

We perform the dimensional reduction to the sentence embeddings generated by the three models using t-SNE and visualize the result in Fig. 4. It can be seen that MPNet-generated embeddings can effectively distinguish sentences with different meanings and group similar sentences into clusters. We further inspect some samples, where the



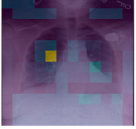

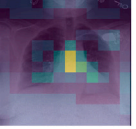



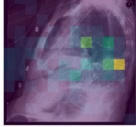
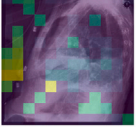
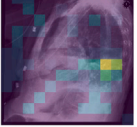
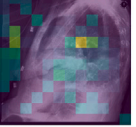
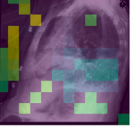

Original Images	Predicted Reports and Visualization					
						
	The patient is status post median sternotomy along with CABG	The lungs appear clear	There is a dual lead left-sided pacemaker again seen with leads extending to the expected positions of the right atrium and right ventricle	The cardiomeastinal silhouette is stable normal	There is no pleural effusion or pneumothorax	Intact median sternotomy wires are again noted
Ground-Truth: Patient is status post median sternotomy and cardiac valve replacement. Dual lead left-sided pacemaker is seen with leads extending to the expected position of the right atrium and right ventricle. There may be minimal basilar atelectasis. No focal consolidation is seen. There is no pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are stable and unremarkable.						
						
	An area of slightly increased lung density in the lateral aspects of the right upper lobe also persists	The cardiac silhouette is stable in appearance and non-enlarged	Moderate bilateral pleural effusions have increased more so on the right with increasing adjacent atelectasis on the left	There is mild pulmonary vascular congestion and edema	There is no pneumothorax	The aorta demonstrates calcifications particularly at the aortic knob
Ground-Truth: There is a moderate left pleural effusion increased since the prior exam. There is a stable small right pleural effusion. The pulmonary vasculature is prominent consistent with pulmonary edema. Opacity in the left lung most likely represents atelectasis. The heart size is top normal and there are aortic knob calcifications. There is no pneumothorax.						
<div> <div>Aorta Calcifications</div> <div>Heart Size</div> <div>Pulmonary Vascular Congestion</div> </div> <div> <div>Atelectasis</div> <div>Pleural Effusion</div> <div>Support Device</div> </div> <div> <div>Cardiomeastinum</div> <div>Pneumothorax</div> <div>Edema</div> </div> <div> <div>Post Median Sternotomy Status</div> </div>						

Fig. 6. Illustrations of the generated reports compared to the ground-truth reports and interpreted with the sentence-level visualizations. To better distinguish the content in the reports, different colors highlight the matched medical terms.

shape (such as Δ , \star) represents descriptions with the same medical terms (such as *Lung Opacity*, *Consolidation*), and the color represents the presence of abnormalities (red for Yes and green for No). MPNet with general pre-training distinguishes the sentence with abnormality type and presence better than the other two models with domain pre-training, indicating that specific domain pre-trained text encoders need further improvements in semantic accuracy.

4.6. Correspondence analysis

To ascertain that the observation intention embeddings encompass relevant medical concepts, we visualize the correspondence between 50 observation intention embeddings and medical terminologies on the MIMIC-CXR dataset. The medical terminologies are obtained by CheXPert, which consists of 14 labels. The resulting visualization, depicted in Fig. 5, displays a heatmap representing the correspondence between the observed intention embeddings and medical definitions. The top and right histograms indicate the frequency of the medical terminologies and intention embeddings, respectively. The visualization result indicates that:

- The majority of medical definition labels have a unique corresponding observation intention embedding, such as *Pneumonia* corresponding to the 23rd intention embedding, *Atelectasis* corresponding to the 28th intention embedding, and *Consolidation* corresponding to the 32nd intention embedding. This provides compelling evidence of the close alignment between intention embeddings and medical definitions. It indicates the capacity of intention embeddings to capture reasonable and distinctive descriptive themes, thereby generating rich and precise textual outputs.
- As observed, certain intention embeddings are associated with multiple medical definitions. For example, the 2nd intention embedding is linked to both *Enlarged Cardiomeastinum* and *Cardiomegaly* medical definition labels. This is because the underlying meanings of these medical definition labels are primarily

similar, both referring to abnormal phenomena caused by cardiac hypertrophy, thus resulting in similar feature distributions in medical images. Additionally, a physician usually observes and describes one of the two in such cases. It can be seen that the proposed model captures the collinearity between medical definitions and the observation and description levels. This strongly validates that, compared to state transition approaches, the observation intention embedding prediction method can better simulate the description and writing habits of physicians in real-life situations.

- In some cases, certain medical definitions are associated with multiple intention embeddings. For instance, both the 1st and 3rd intention embeddings are closely related to the medical terminology of *Pleural Effusion*. This phenomenon arises because the observation intention modeling further refines the corresponding medical label definitions. With further analysis, we find out that the 1st intention embedding primarily describes pulmonary vessels, while the 3rd intention embedding mainly describes the glenohumeral joints region. The reason is that, in realistic scenarios, physicians must observe the local vessels and joints to determine whether a patient has pleural effusion and its cause.
- Upon examining the corresponding sentences, we found that the 44th intention embedding is closely related to the description of consolidation, such as “there is no large effusion or definite consolidation” and “no lobar consolidation is seen”. According to the category, the 44th intention embedding should be classified as a medical definition label for consolidation, but CheXPert mistakenly labeled it as “No finding”. Therefore, a thorough analysis of the semantics of intention embeddings can help identify annotation errors and omissions.
- The heatmap also reveals that a few labels have no clear correspondence with any observation intention, such as *Lung Lesion* and *Pleural Other*. The corresponding histogram shows that they have a low occurrence frequency, indicating that their training may still be affected by the long-tailed distribution of the data.

4.7. Visualization and interpretable analysis

Most previous works struggle to generate explanations that directly relate medical reports to images. Some works generate reports and provide explanations word-by-word, which is unintuitive for humans to judge the accuracy of the description logic. Some works generate reports using the auxiliary multi-label prediction task to obtain the correspondence between medical labels and sentences. This approach allows for the use of the class activation map to visualize the labels and establish the association of report and image features. However, the semantic information in the text description far exceeds that in the label. Additionally, the visualization of the label classification fails to provide a complete explanation of the sentence generation process. It thus cannot be considered equivalent to the explanation of the sentence.

TranSQ can provide a comprehensive and intuitive depiction of sentence-level interpretation by visualizing the attention weight map corresponding to each observation intention embedding. As shown in Fig. 6, we offer two case studies in both the front and side views of MIMIC-CXR, comprising a comparison of the predicted reports against the ground truth and visual-semantic attention maps. The medical terms and their corresponding sentences are color-coded in the text. For instance, the medical keyword “*Cardio mediastinal Silhouette*” is identified by the pale green mark referenced in the cases. Notably, the predicted results in each case are highly aligned with the semantics of the ground truth.

Besides, it is observed that TranSQ can accurately predict some rare professional terminologies, such as *Post Median Sternotomy Status* and *Aortic Calcification*.

The visual-semantic attention map fully demonstrates the effectiveness of TranSQ. The attention map corresponding to each observation intention embedding visualizes the feature importance of the medical image. For example, in *Case #1*, the attention map of the 1st sentence “*The cardiomeastinal silhouette is stable normal*” focuses on the heart region, and the attention map of the 3rd sentence about the *pacemaker* accurately focuses on the device region. Moreover, the attention maps of the sentences with similar clinical concerns are also focused on similar areas, such as the descriptions about “*status post median sternotomy*” and “*Intact median sternotomy wires*” in *Case #1* and descriptions about *compression deformation* and *degenerative changes* in *Case #2*, which indicates the strong correlation between visual and semantic representations.

In essence, the comprehensive and precise visual interpretation of the visual features we provide offers conclusive evidence for the robust association between visual and semantic representations. Additionally, it also facilitates a straightforward and intuitive assessment of the integrity of the information conveyed by the sentence, which is particularly advantageous for medical practitioners.

5. Conclusion and discussion

This paper proposes a *Transformer-based Semantic Query* (TranSQ) approach to address the medical report generation problem. By simulating the thinking logic of human doctors of “Formulating Intentions → Understanding Visual Properties → Composing Descriptions”, we consider the method to learn a set of intention embeddings and predict sentence candidates with a semantic query process. In particular, we propose a bipartite matching-based strategy to achieve the dynamic correspondence between intention embeddings and ground-truth sentences during the training process to induct the medical terminology concepts automatically.

The experiments on two chest X-ray report generation benchmarks verify the effectiveness of our approach, especially the significant improvements in the clinical efficacy metrics. Besides, we conduct comprehensive research on the proposed method. Specifically, we propose a retrieval/generation hybrid strategy to generate sentence candidates,

extending the flexibility of our model’s descriptions. We also explore the integration of domain pre-trained models to improve our model and discuss the benefits of each. Finally, we demonstrate the visualization of intention-terminology correspondence and sentence-level interpretations, showcasing the potential clinical application as an auxiliary diagnostic approach.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in or the review of the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by National Key R & D Program of China (2022YFF1202400), the Key Laboratory for Corneal Diseases Research of Zhejiang Province, Key R & D Projects of the Ministry of Science and Technology, China (2020YFC0832500), Project by Shanghai AI Laboratory, China (P22KS00111), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study, China (SN-ZJU-SIAS-0010).

References

- Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A., 2021. Automated radiology report generation using conditioned transformers. *Inform. Med. Unlocked* 24, 100557.
- Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. pp. 65–72.
- Biswal, S., Xiao, C., Glass, L.M., Westover, B., Sun, J., 2020. Clara: clinical report auto-completion. In: *Proceedings of the Web Conference*. pp. 541–550.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, Vol. 33. pp. 1877–1901.
- Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L., 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Chen, Z., Song, Y., Chang, T.-H., Wan, X., 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R., 2020. Meshed-memory transformer for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10578–10587.
- Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J., 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* 23 (2), 304–310.
- 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*.
- Elman, J.L., 1990. Finding structure in time. *Cogn. Sci.* 14 (2), 179–211.
- Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P., 2021. Retrieval-based chest X-ray report generation using a pre-trained contrastive language-image model. In: *Machine Learning for Health*. PMLR, pp. 209–219.
- Gajbhiye, G.O., Nandedkar, A.V., Faye, I., 2020. Automatic report generation for chest X-Ray images: A multilevel multi-attention approach. In: *Computer Vision and Image Processing*. Singapore, pp. 174–182.
- Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L.J., Bradley, A.P., 2019. Producing radiologist-quality reports for interpretable deep learning. In: *IEEE 16th International Symposium on Biomedical Imaging*. IEEE, pp. 1275–1279.
- Han, Z., Wei, B., Leung, S., Chung, J., Li, S., 2018. Towards automatic report generation in spine radiology using weakly supervised framework. In: *Medical Image Computing and Computer Assisted Intervention*. Cham, pp. 185–193.

- Harzig, P., Einfalt, M., Lienhart, R., 2019. Automatic disease detection and report generation for gastrointestinal tract examination. In: *Proceedings of the 27th ACM International Conference on Multimedia*. pp. 2573–2577.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9 (8), 1735–1780.
- Huang, L., Wang, W., Chen, J., Wei, X.-Y., 2019. Attention on attention for image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4634–4643.
- Irvine, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. pp. 590–597.
- Jing, B., Wang, Z., Xing, E., 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.
- Jing, B., Xie, P., Xing, E., 2018. On the automatic generation of medical imaging reports. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. pp. 2577–2586.
- Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S., 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Kong, M., Huang, Z., Kuang, K., Zhu, Q., Wu, F., 2022. Transq: Transformer-based semantic query for medical report generation. In: *Medical Image Computing and Computer Assisted Intervention*. Cham, pp. 610–620.
- Li, Y., Liang, X., Hu, Z., Xing, E.P., 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Adv. Neural Inf. Process. Syst.* 31.
- Li, C.Y., Liang, X., Hu, Z., Xing, E.P., 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. pp. 6666–6673.
- Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X., 2023. Dynamic graph enhanced contrastive learning for chest X-ray report generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3334–3343.
- Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P., 2017. Recurrent topic-transition gan for visual paragraph generation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3362–3371.
- Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. pp. 74–81.
- Liu, G., Hsu, T.-M.H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., Ghassemi, M., 2019. Clinically accurate chest x-ray report generation. In: *Machine Learning for Healthcare Conference*. PMLR, pp. 249–269.
- Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y., 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 13753–13762.
- Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al., 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Adv. Neural Inf. Process. Syst.* 34, 16266–16279.
- Loshchilov, I., Hutter, F., 2017. Fixing weight decay regularization in adam. *CoRR abs/1711.05101*. *arXiv:1711.05101*.
- Ma, X., Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., Sun, X., 2021. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*.
- Ma, K., Wu, K., Cheng, H., Gu, C., Xu, R., Guan, X., 2018. A pathology image diagnosis network with visual interpretability and structured diagnostic report. In: *Neural Information Processing*. Cham, pp. 282–293.
- Ni, J., Hsu, C.-N., Gentili, A., McAuley, J., 2020. Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. *arXiv preprint arXiv:2010.02467*.
- Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M., 2021. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*.
- 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.
- Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Sun, L., Wang, W., Li, J., Lin, J., 2019. Study on medical image report generation based on improved encoding-decoding method. In: *Intelligent Computing Theories and Application: 15th International Conference*. Springer, pp. 686–696.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems*, Vol. 30.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3156–3164.
- Wang, Z., Han, H., Wang, L., Li, X., Zhou, L., 2022b. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Trans. Med. Imaging* 41 (10), 2803–2813.
- Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., Chen, J., 2022a. An inclusive task-aware framework for radiology report generation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 568–577.
- Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M., 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9049–9058.
- Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L., 2022c. A medical semantic-assisted transformer for radiographic report generation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 655–664.
- Wang, Z., Zhou, L., Wang, L., Li, X., 2021. A self-boosting framework for automated radiographic report generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2433–2442.
- Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D., 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In: *European Conference on Computer Vision*. Springer, pp. 162–178.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning*. PMLR, pp. 2048–2057.
- Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X., 2018. Multimodal recurrent model with attention for automated radiology report generation. In: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 457–466.
- Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L., 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Med. Image Anal.* 80, 102510.
- Yang, X., Ye, M., You, Q., Ma, F., 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. *arXiv preprint arXiv:2106.06471*.
- You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4651–4659.
- You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X., 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 72–82.
- Yuan, J., Liao, H., Luo, R., Luo, J., 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: *Medical Image Computing and Computer Assisted Intervention*. Springer, pp. 721–729.
- Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D., 2020. When radiology report generation meets knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. pp. 12910–12917.
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J., 2020. Unified vision-language pre-training for image captioning and vqa. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. pp. 13041–13049.