SA-MPF: A Status-Aware Mask Prediction Framework for Online Disease Diagnosis

Zefa Hu[1,2] , Linghui Meng[1,2], Yunlong Zhao[1,2], Yuanyuan Zhao[2], Shuang Xu[2], and Bo Xu[1,2,*]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China
{huzefa2018, menglinghui2019, zhaoyunlong2020}@ia.ac.cn
{yuanyuan.zhao, shuang.xu, xubo}@ia.ac.cn

***Abstract*—An increasing number of individuals are turning to online self-diagnosis by matching their symptoms with potential medical conditions. This process involves two primary components: symptom inquiry and disease prediction. Existing works employ two separate modules to learn these tasks individually. Nevertheless, this intuitive approach encounters low data efficiency due to the separate learning of each module. In addition, previous research incorporates symptom statuses solely as part of the input without any additional modeling. However, this oversight neglects the importance of symptom status, which indicates whether the user has experienced the symptom. The status significantly influences both symptom inquiry strategies and disease prediction. To address these challenges, we propose a Status-Aware Mask Prediction Framework for online disease diagnosis, called SA-MPF. SA-MPF formalizes symptom inquiry and disease prediction as a single masked token prediction task, distinguishing them solely through the masked token type. Furthermore, we introduce a masked status prediction task, which unifies the prediction of symptom or disease statuses in a similar manner to masked token prediction, thereby enhancing the modeling of symptom and disease statuses. We evaluate SA-MPF on several datasets collected from various sources. The experimental results demonstrate substantial improvements achieved by SA-MPF. For example, on the GMD-12 dataset, SA-MPF demonstrates a noteworthy 5% improvement in diagnostic accuracy, from 82% to 87%.**[1]

***Index Terms*—online disease diagnosis, self-diagnosis, symptom checking**

I. Introduction

With the widespread adoption of internet technology, an increasing number of users are turning to online searches to address their health concerns and perform self-diagnosis [1]. However, traditional search engines often fail to meet people's demands for accurate medical information [2]. Given that many users lack adequate medical knowledge [3], this compromises the quality of health-related queries, leading to search results that are diverse and may not offer clear and professional medical guidance [4]–[6].

In response to the challenge, symptom checkers have been introduced as an alternative to online self-diagnosis [7]. Prominent among them are tools from Mayo Clinic [8] and WebMD [9], providing users with more focused and targeted medical advice. The workflow of a symptom checker typically consists of three steps [10] : 1) Patients provide their initial symptom information, named explicit symptoms, and $True$ or $False$ is called the status, indicating whether a symptom is present, as shown in Table I; 2) Based on the provided information, the tool further inquires with a series of related questions to obtain more symptom information, the symptoms from the inquires are called implicit symptoms; and 3) The tool gives a possible diagnosis based on all the provided data. In contrast to search engines, symptom checkers excel by not demanding users to create precise health queries, the requirement that users often find challenging [2], [11]. Furthermore, symptom checkers boast higher accuracy in disease diagnosis while demanding less effort and time.

TABLE I
An example of automatic diagnosis data.

Explicit Symptoms:	cough: true fever: true
Implicit Symptoms:	listlessness: false anorexia: true vomit: true
Disease:	upper respiratory tract infection

Online disease diagnosis involves two main components: symptom inquiry and disease prediction. A significant group of existing research [12]–[17] viewed online disease diagnosis as a Markov Decision Process (MDP) [18] and employed Reinforcement Learning (RL) [19], [20] to address it. However, RL-based methods have potential drawbacks in online disease diagnosis. RL requires explicit learning objectives and detailed rewards, making it challenging to strike a balance between symptom inquiry and disease classification. In addition, RL demands a substantial amount of data to perform effectively. Regrettably, the medical field frequently faces a shortage of available data [21]. Another approach is through supervised learning, mainly involving classification methods and generative methods. Recent Transformer-based [22] generative methods such as Diaformer [23] and CoAD [24], as well as classification methods like MTDiag [21], have shown excellent

This work is supported by the National Key Research and Development Program of China (No.2021ZD0201500) and the National Natural Science Foundation of China (No.62206294).

*Corresponding author.

[1] Our codes can be found at https://github.com/FlyingCat-fa/SA-MPF.

performance. Nevertheless, even though these methods share the majority of model parameters, they still treat symptom inquiry and disease diagnosis as separate tasks, posing a challenge in terms of data efficiency. Additionally, existing research overlooks the significance of symptom status, which indicates whether the user has experienced the symptom. This status plays a crucial role in both symptom inquiry strategies and disease prediction. However, current methods merely incorporate it as part of the input without further modeling.

In this work, we propose a Status-Aware Mask Prediction Framework for online disease diagnosis (SA-MPF) to address above challenges. SA-MPF formalizes symptom inquiry and disease prediction as a single masked token prediction task. Specifically, for known symptoms, SA-MPF treats each symptom as an individual token, incorporating its status and a specialized token type S as the symptom input. Additionally, a masked token $[M]$ is introduced as a special input. For symptom inquiry, S is the token type of $[M]$, indicating that the prediction will be a symptom. The status input of $[M]$ is either 'True' or 'False,' representing two inquiry scenarios: confirming a specific disease or excluding similar diseases [25]. For disease prediction, $[M]$ has the status $True$ and type D, signifying the prediction of the disease the user has. By performing the single token prediction task, SA-MPF facilitates joint learning for symptom inquiry and disease prediction more effectively.

To enhance SA-MPF's modeling of symptom and disease statuses, we introduce a masked status prediction task, aiming to infer the statuses of symptoms or diseases in a manner similar to masked token prediction. The key distinction between this task and masked token prediction is that, in this task, the actual token is known, but what needs to be predicted is the token's status. Specifically, for masked symptom status prediction, we leverage known symptoms and the actual disease to infer the status of an unknown symptom related to them. For masked disease status prediction, we expect the model to predict the status of a disease based on known symptoms, indicating whether the user has the disease. Through masked status prediction, SA-MPF comprehensively learns the status relationships between diseases and symptoms, thereby further enhancing its diagnostic accuracy and efficiency. During the inference phase, we diagnose diseases from both disease status and token prediction perspectives to further improve diagnostic accuracy. An overview of SA-MPF is shown in Figure 1.

We conduct experiments on several medical diagnosis datasets from diverse sources, including online medical websites, offline hospitals, and the knowledge base SymCAT [26]. These datasets vary in terms of data scale, as well as the number of diseases and symptoms they include. Extensive experimental results demonstrate that the proposed SA-MPF attains state-of-the-art performance on these datasets, thereby highlighting the effectiveness of our approach.

Our primary contributions are as follows:

- We introduce a Status-Aware Mask Prediction Framework for online disease diagnosis (SA-MPF), which unifies symptom inquiry and disease prediction into a masked token prediction task, facilitating joint learning for symptom inquiry and disease prediction more effectively.
- We introduce a masked status prediction task that simultaneously addresses masked disease status prediction and masked symptom status prediction. By incorporating these tasks into the masked prediction framework, the model's ability to capture status information is enhanced.
- In extensive experiments across multiple datasets of varying sources and scales, the proposed SA-MPF consistently demonstrated state-of-the-art performance, validating the effectiveness of the mask prediction framework.

II. Related Work

Early studies often viewed automatic diagnosis as sequential decision process problem and employed reinforcement learning (RL) to address it. [27] views automatic diagnosis task as a combination of symptom inquiry and disease classification and firstly leverages reinforcement learning (RL) to slove the problem. Subsequent studies focus on enhancing the RL framework's efficiency and accuracy [13], [15]–[17], [28], [29]. The comprehensive review by [20] sheds light on the trajectories and milestones of RL in automated medical diagnosis. Nonetheless, a recurring criticism has been the inefficiency of data associated with RL-based techniques. Attaining optimal results continues to pose a significant challenge, especially given the data constraints in the medical domain.

To mitigate the challenges associated with exploration and sparse rewards in RL, many supervised-base works have been proposed, including generated-based method [23], [24] and classification-based methods [21], [30]. However, both of these methods treat symptom inquiry and disease diagnosis as two separate tasks. This hinders the comprehensive modeling of symptom inquiry and disease diagnosis, consequently reducing learn efficiency. Moreover, to our best knowledge, all existing models, only take the status as the part of input and lack further status modeling.

III. Methodology

In this section, we first introduce the proposed mask prediction framework. Subsequently, we introduce the prediction of masked symptom tokens and masked disease tokens, which are employed for symptom inquiry and disease prediction, respectively. Furthermore, we expound on the prediction of masked symptom status and masked disease status, which serve to enhance the modeling of status dependencies between symptom and disease. Lastly, we present the joint prediction based on masked token and status prediction of the disease during the inference phase, facilitating disease discrimination.

A. Mask Prediction Framework

Figure 1 illustrates the mask prediction framework. In this framework, we utilize a series of stacked Transformer encoder blocks to model online disease diagnosis via mask prediction. Each Transformer block consists of a feed-forward layer and a multi-head attention layer, where the parameters are shared among all input tokens through self-attention [22].

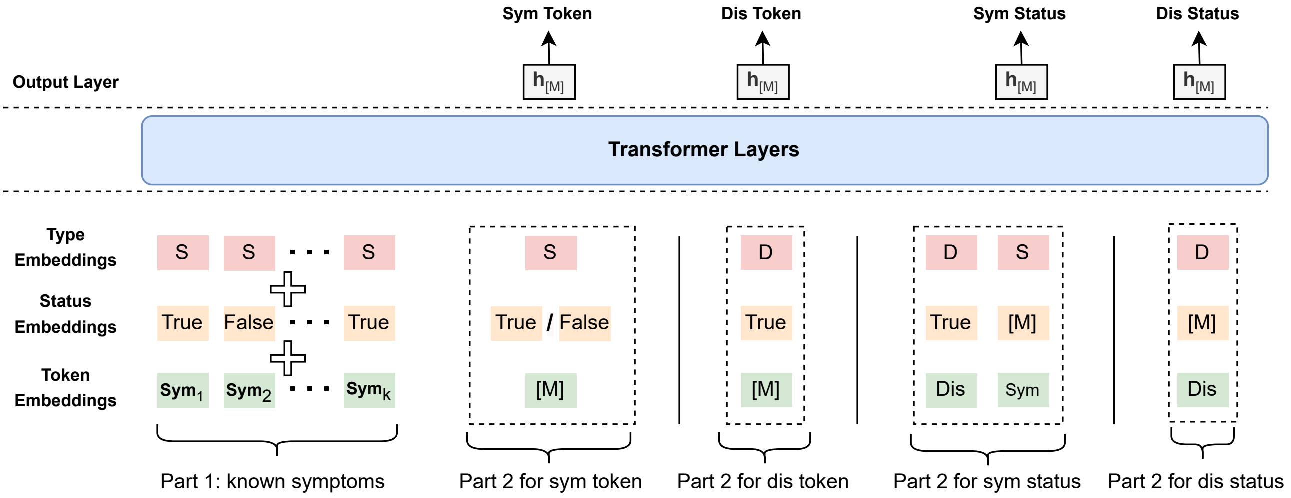

Fig. 1. The Overview of proposed status-aware mask prediction framework. sym and Dis are short forms for symptom and disease, with S and D representing the types of symptom and disease, respectively.

1) Input Representation: As shown in Figure 1, known symptoms are transformed into specific tokens, which we refer to as the part 1 of our input. Correspondingly, we design distinct part 2 of the input for different sub-tasks. Differing from the original Transformer [22], we have omitted positional encoding. In addition to symptoms as inputs, diseases may also serve as inputs. To differentiate between diseases and symptoms, we introduce type embedding, where D and S indicate disease and symptom types, respectively. Furthermore, status information is incorporated through status embedding. Here, $True$ and $False$ indicate whether a patient exhibits a specific symptom or has a disease. Both status embedding and type embedding are added to the token embedding. Additionally, we design a special token M for masked inputs.

2) Mask Prediction: After multiple layers of Transformer-based attention modeling, we obtain the representation of the final layer for the masked token, denoted as $h_{[M]}$. This representation serves as the input for our tailored mask prediction tasks, encompassing masked token prediction and masked status prediction. Specifically, masked token prediction encompasses both masked symptom token prediction and masked disease token prediction, differing only in their input, yet sharing the same model parameters entirely. Similarly, masked status prediction comprises masked symptom status prediction and masked disease status prediction, also differing in input but sharing all model parameters. Notably, the sole difference between masked token prediction and masked status prediction lies in the parameters of their respective task heads. To map the hidden state $h_{[M]}$to probabilities across the corresponding vocabularies, we employ two task heads, each composed of a feed forward network and a softmax function:

$$\begin{aligned} p_T &\leftarrow Token\_Head(\mathbf{h}_{[M]}), \\ p_S &\leftarrow Status\_Head(\mathbf{h}_{[M]}), \end{aligned} \tag{1}$$

The token vocabulary includes symptom tokens and disease tokens, while the status vocabulary consists of $True$ and $Flase$. We utilize the negative log-likelihood loss for the two tasks, denoted $\mathcal{L_T}$ and $\mathcal{L_S}$. The final loss is the sum of these two losses:

$$\mathcal{L} = \mathcal{L_T} + \mathcal{L_S} \tag{2}$$

During the inference phase, the framework initiates a symptom inquiry, leveraging masked symptom token prediction. Once the symptom inquiry phase is completed, disease prediction is carried out, relying on the gathered symptoms. Additional details will be elaborated in the following sections.

B. Masked Symptom Token Prediction

In this section, we reformulate symptom inquiry into a masked symptom token prediction task. Formally, within a dataset for automatic diagnosis, each example has explicit symptoms denoted as $S_{exp} = \{es_1, ..., es_n\}$, implicit symptoms as $S_{imp} = \{is_1, ..., is_m\}$, and a corresponding disease tag, Dis. Initially, only the explicit symptoms obtained from a patient's self-report are available. During each symptom inquiry, the patient simulator responds with $True$ or $False$ for a positive or negative symptom, and "not sure" for symptoms not appear in the user goal $S_{exp} \cup S_{imp}$. The objective of symptom inquiry is to maximize the likelihood $P(S_{imp}|S_{exp})$. We denoted the symptoms obtained through inquiries as $S_{im\_known} \subseteq S_{imp}$, while those not acquired as $S_{im\_unknown} = S_{imp} - S_{im\_known}$.Taking into account the disorderliness of symptoms in the datasets, the learning objective can be formalized as follows:

$$\prod_{S_{im\_known} \subseteq S_{imp}} P(S_{im\_unknown} \mid S_{exp}, S_{im\_known}) \tag{3}$$

To reformulate the multi-step reasoning of symptom inquiry into token prediction, we decompose the multi-turn diagnostic dialogue into multiple independent one-step token prediction data examples, covering all possible cases in the dialogue. To maximize the $p(S_{imp}|S_{exp})$, we could maximize each $P(S_{im\_unknown} \mid S_{exp}, S_{im\_known})$, which corresponds to an intermediate state before a dialogue inquiry: given explicit symptoms S_{exp} and known implicit symptoms S_{im_known}, the objective is to predict the remaining implicit symptoms $S_{im\_unknown}$. Since $S_{im\_unknown}$ may contain multiple symptoms, we treat each symptom as a label, and construct independent training examples for each symptom label to perform masked symptom token prediction.

Specifically, we take S_{exp} and S_{im_known} as the part 1 of the token input, and incorporate the corresponding symptom status and S as the status and type of tokens. For the part 2 of the input, we use a special token, $[M]$, as the token input. Since we know the masked token is a symptom, S is input as the token type. Regarding the status input, we incorporate both $True$ and $False$ to inquire about the presence or absence of symptoms among the unknown implicit symptoms, respectively. This approach mirrors the logic of medical inquiry, where during the diagnosis process, certain symptoms are inquired about to increase the confidence in diagnosing a specific disease, while the absence of symptoms is queried to eliminate the possibility of similar diseases. The problem can be converted to the masked token prediction:

$$Input(S_{exp} \cup S_{im\_known}, part_2) \xrightarrow{\text{predict}} \text{Label}(Sym), \quad part_2 = \{[M], Status, S\} \tag{4}$$

where $part_2$ represents the part 2 of the input. $[M]$, $Status$ and S denote the masked token, the status and type. The status can be either $True$ or $False$, and $Label(Sym) \in S_{im\_unknown}$ corresponds the symptoms associated with the specific status. For the cases where a specific status corresponds to multiple symptoms, we create multiple training examples, with each example corresponding to a specific symptom label. We partially demonstrate the construction of training data based on the example in Table 1. When only explicit symptoms are known, we can obtain three training data samples shown in the upper part of Table II. When ”anorexia” is added as a new known symptom, we can obtain two additional training data samples in the lower part of Table II. For a dialogue with k implicit symptoms, where for each of the i known implicit symptoms, there are $\binom{k}{i}$ cases, and each case have $k-i$ unknown implicit symptoms, a total of $\sum_{i=0}^{k}(k-i) * \binom{k}{i}$ training samples can be constructed. This also increases the scale of the training data, alleviating the issue of data sparsity in the medical field.

TABLE II
A DECOMPOSING DEMONSTRATION OF A MULTI-TURN DIALOGUE EXAMPLE FROM TABLE I INTO MASKED TOKEN PREDICTION.

Part 1	Part 2	Label
cough, fever	[M], True, S	anorexia
cough, fever	[M], True, S	vomit
cough, fever	[M], False, S	listlessness
cough, fever, anorexia	[M], True, S	vomit
cough, fever, anorexia	[M], False, S	listlessness

We follow the multi-turn setting to imitate the medical dialogue scenario during the inference. Taking the part 1 and part 2 as inputs in each turn, we obtain the corresponding probabilities for all symptom inquiries of ”true” and ”false”, separately. We then choose, based on the highest probability, whether to inquire about ”true” or ”false,” as well as which symptom to inquire about. If the patient responds with ”True” or ”False”, the symptom and the corresponding status will be added to the known implicit symptoms S_{im_known}. The part 1 and $S_{im\_unknown}$ will also be updated. Otherwise, the next probability symptom will be the inquired symptom until finding a symptom in $S_{im\_unknown}$ or stopping. A stop threshold $\delta \in (0, 1)$ serves as the minimum probability boundary. After the symptom inquiry is finished, all obtained symptoms S_{exp} and S_{im_known} will be utilized for disease prediction.

C. Masked Disease Token Prediction

Masked disease token prediction is similar to masked symptom token prediction but focuses on modeling disease diagnosis. We primarily design the input for masked disease token prediction in a specific manner. The part 1 of input consists of explicit symptoms and all implicit symptoms $S_{exp} \cup S_{imp}$, rather than explicit symptoms and known implicit symptoms $S_{exp} \cup S_{im\_known}$, where $S_{im\_known} \subseteq S_{imp}$. It helps to avoid extra noise and performance degradation resulting from predicting diseases solely based on a partial set of symptoms. We also take $[M]$ as the token input for part 2. To distinguish it from symptom prediction, we use D as the type input, indicating that what needs to be predicted is a disease. The input for the status is simply $True$, as we are concerned with which disease the patient has been diagnosed with. Masked disease token prediction can be formulated as follows:

$$Input(S_{exp} \cup S_{imp}, part_2) \xrightarrow{\text{predict}} \text{Label}(Dis), \quad part_2 = \{[M], True, D\} \tag{5}$$

where $Label(Dis)$ corresponds the disease label.

In this way, symptom inquiry and disease diagnosis are effectively unified as a single masked token prediction task, allowing them to mutually enhance each other more comprehensively. During the inference, we set all the predicted symptom token probabilities to 0, considering only the disease probabilities for disease prediction. A similar approach is employed for symptom prediction as well.

D. Masked Symptom Status Prediction

Masked symptom status prediction focuses on modeling the status relationships between diseases and symptoms from the perspective of symptom reasoning. After training on this task, we aim for the model to learn how to infer the corresponding status of unknown implicit symptoms $S_{im\_unknown}$ when given known symptoms $S_{exp} \cup S_{im\_known}$ and informed the disease label. For a multi-turn dialogue with k

implicit symptoms, we decompose it into $\sum_{i=0}^{k}\binom{k}{i}$ independent one-step cases, and get corresponding S_{exp}, S_{im_known} and $S_{im\_unknown}$ tuples, as described in Section III-B. The part 1 of the input for masked symptom status prediction is $S_{exp} \cup S_{im\_known}$, identical to that of masked symptom token prediction. The part 2 has two tokens, including the disease label and a unknown implicit symptom $Label(Sym) \in S_{im\_unknown}$. We take $True$ and D as the status and type of the disease input. $Label(Sym)$ is input as a token, we input $[M]$ and S as the status and type of the symptom token. Based on this input, the model predicts the corresponding state for a given symptom from unknown implicit symptoms. Masked symptom status prediction can be formulated as follows:

$$\begin{aligned}&Input(S_{exp} \cup S_{im\_known}, part_2) \xrightarrow{\text{predict}} \text{Label}(Status),\\&part_2 = \{\text{Label}(Dis), True, D\} \cup \{\text{Label}(Sym), [M], S\}\end{aligned} \tag{6}$$

where $Label(Sym)$ is a unknown implict symptom, $Label(Status)$ and $[M]$ denote the actual status and the status input of $Label(Sym)$.

E. Masked Disease Status Prediction

Masked disease status prediction focuses on modeling the status relationships between diseases and symptoms from the perspective of disease reasoning. This task aims for the model to learn whether the patient may have a disease when given known symptoms $S_{exp} \cup S_{im\_known}$. Here $True$ and $False$ indicate that the patient may have or must not have a disease, respectively.

For the labeled disease, given $S_{exp} \cup S_{im\_known}$, we can know the status is $true$. However, there are some unknown implicit symptoms, it is possible that some diseases cannot be ruled out. But we can determine that when all implicit symptoms are known, the patient cannot have any diseases other than the labeled one. Therefore, for disease labels, the task can be formulated as follows:

$$Input(S_{exp} \cup S_{im\_x}, part_2) \xrightarrow{\text{predict}} \text{Label}(Status). \tag{7}$$

we can construct the training examples of the disease label:

$$\begin{aligned}&S_{im\_x} = S_{im\_known},\\&part_2 = \{\text{Label}(Dis), [M], D\},\\&\text{Label}(Status) = True.\end{aligned} \tag{8}$$

The training examples of the non-labeled disease can be constructed as follows:

$$\begin{aligned}&S_{im\_x} = S_{imp},\\&part_2 = \{\text{Other}(Dis), [M], D\},\\&\text{Label}(Status) = False,\end{aligned} \tag{9}$$

where $Other(Dis)$ denotes a non-labeled disease.

TABLE III
Statistics of experimental datasets.

Dataset	Type	Disease	Symptom	Train	Test
Dxy	Web	5	41	423	104
Muzhi	Web	4	66	568	142
GMD-12	Hospital	12	118	2151	239
SymCAT-90	Synthetic	90	266	24000	6000
SymCAT-200	Synthetic	200	328	20000	10000
SymCAT-300	Synthetic	300	349	20000	10000
SymCAT-400	Synthetic	400	355	20000	10000

F. Joint Token and Status Disease Inference

Given that masked disease token prediction and masked disease status prediction contribute to disease reasoning from both token and status perspectives, we integrate them during inference for disease prediction. Initially, we derive the probability distribution of diseases through masked disease token prediction. Subsequently, we select the top-k diseases and obtain their statuses through masked disease status prediction. Diseases with a predicted status of $False$ are eliminated. Finally, we choose the disease with the highest remaining probability as the ultimate diagnostic disease.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate our method on Dxy [15], Muzhi [12], GMD-12 [25] and the synthetic SymCAT datasets including four versions. Both Dxy and Muzhi are obtained from online medical websites. GMD-12 is constructed from hospital medical records. [16] constructed a synthetic dataset based on a symptom-disease database known as SymCAT, referred to as SymCAT-90. [14] created three more versions of SymCAT, namely SymCAT-200, SymCAT-300, and SymCAT-400. The characteristics of these datasets are shown in Table III.

2) Baselines: We initially choose the widely used traditional classifier SVM [31] as a baseline. **SVM-ex&im** utilizes both explicit and implicit symptoms for disease prediction. The RL-based baseline models consist of **Basic DQN** [12], **REFUEL** [14], **PPO** [32], **HRL** [16], **KR-DS** [15], and **BR-Agent** [25]. Several Transformer-based supervised learning baselines include **Diaformer** [23], **CoAD** [24], and **MTDiag** [21]. We also consider existing methods tailored for extensive disease spaces, such as **BSODA** [33], **GAMP** [28], **MMF-AC** [29], and **V-IP** [30].

3) Metrics.: Following previous settings [12], [21], [23], we use disease diagnosis accuracy, implicit symptom recall, and average inquiry turns as evaluation metrics for Dxy, Muzhi, GMD-12, and SymCAT-90. Accuracy was the primary metric for disease diagnosis, while recall and average turns served as indicators of inquiry efficiency. For SymCAT-200, SymCAT-300, and SymCAT-400, we employed accuracy metrics for the top 1, top 3, and top 5 predictions, along with the average inquiry turns for disease diagnosis, following the settings of previous works [14], [33].

TABLE IV
RESULTS ON DXY, MUZHI, GMD-12 AND SYMCAT-90 DATASETS.

	Dxy			MuZhi			GMD-12			SymCAT-90		
	DAcc	SRec	ATurn	DAcc	SRec	ATurn	DAcc	SRec	ATurn	DAcc	SRec	ATurn
SVM-ex&im	77.9	-	-	71.0	-	-	-	-	-	73.2	-	-
Basic DQN	73.1	32.2	2.9	65.0	30.1	**3.1**	62.0	5.0	-	35.6	2.0	**2.0**
HRL	69.5	16.1	**2.4**	69.4	27.6	3.5	-	-	-	49.6	33.8	8.4
KR-DS	74.0	-	3.4	73.0	-	3.4	69.0	21.0	-	-	-	-
GAMP	76.9	-	3.3	73.0	-	6.3	-	-	-	-	-	-
PPO	74.6	-	3.3	73.2	-	6.3	-	-	-	61.8	-	12.6
Diaformer	82.9	82.7	13.1	74.2	75.2	15.3	-	-	-	73.3	90.6	13.7
BR-Agent	84.6	48.6	-	76.0	67.0	-	82.0	50.0	-	-	-	-
CoAD	85.0	**93.0**	10.5	75.0	83.0	13.4	-	-	-	-	-	-
MTDiag	85.4	91.3	12.5	75.9	79.4	17.9	-	-	-	75.4	90.7	15.1
SA-MPF	**87.5**	91.3	16.2	**77.5**	**85.0**	19.3	**87.0**	**89.0**	15.4	**77.8**	**93.8**	15.4

TABLE V
RESULTS ON SYMCAT-200, SYMCAT-300 AND SYMCAT-400 DATASETS.

	SymCAT-200				SymCAT-300				SymCAT-400			
	Top1	Top3	Top5	ATurn	Top1	Top3	Top5	ATurn	Top1	Top3	Top5	ATurn
REFUEL	53.76	73.12	79.53	**8.24**	47.65	66.22	71.79	**8.39**	43.01	59.65	68.89	**8.92**
BSODA	55.65	80.71	89.32	12.02	48.23	73.82	84.21	13.10	44.63	69.22	79.54	14.42
V-IP	68.10	-	-	-	59.90	-	-	-	51.30	-	-	-
MMF-AC	59.00	84.66	92.45	12.54	51.91	78.24	87.55	13.80	45.80	72.14	82.43	14.73
SA-MPF	**86.55**	**96.41**	**97.82**	12.14	**84.77**	**94.64**	**96.74**	12.79	**82.50**	**93.33**	**95.95**	13.14

4) Implementation Details: The main differences in model configurations among the various datasets are primarily related to the number of transformer layers and the hidden size. Specifically, for Dxy, MuZhi, GMD-12, and SymCAT, SA-MPF has 4, 4, 6, and 6 transformer layers, respectively, with hidden sizes of 512, 512, 512, and 768, respectively. We employ the AdamW optimizer with a learning rate of 3e-5, a weight decay of 0.01, and warm up step to 20. We set the stop threshold δ to 0.009, and follow previous works [21], [23], [24] in limiting the maximum number of dialogue turns to 20. During the inference, Top-k is set to Top-20 for disease prediction. diseases is selected after masked disease token prediction during the inference. In the synthetic datasets from SymCAT, since there is no dialogue, the examples only include symptoms with the $True$ status. Consequently, we do not perform masked status prediction within the these datasets.

B. Results

We report the results of baselines from previous works if available. Table IV presents experimental results for Dxy, Muzhi, GMD-12 and SymCAT-90. Overall, our proposed SA-MPF achieves superior or competitive performance in diagnostic accuracy and symptom recall.

Compared to the competitive RL-based approach, BR-Agent, we achieve superior results in diagnostic accuracy, such as an absolute improvement of over 5% on the GMD-12. Furthermore, SA-MPF demonstrates significantly higher symptom recall, which constitutes a key factor contributing to our advantage in diagnostic accuracy over RL-based approaches. Compared to the classification-based approach, MTDiag, SA-MPF demonstrates significant performance improvements, achieving better diagnostic accuracy, such as an absolute improvement of 2.1% on the Dxy dataset. Additionally, SA-MPF achieves superior symptom recalls except on the Dxy dataset, where SA-MPF and MTDiag exhibit the same symptom recalls. We attribute these improvements to our unified modeling of symptom inquiry and disease diagnosis as a single masked token prediction task, which enables more effective mutual learning between symptom inquiry and disease prediction, while MTDiag requires handling them with two separate task heads. Our proposed status modeling tasks also contribute to these advantages, which will be further analyzed in Section IV-C. We observed that, although CoAD achieves high symptom recalls, such as 93% on the Dxy dataset, its disease diagnosis accuracy falls behind our SA-MPF. This is likely due to CoAD's sequence modeling introducing symptom order, which interferes with disease diagnosis, a problem not present in SA-MPF.

Additionally, we observed that our method tends to have slightly more dialogue turns. This is pragmatic and justifiable in real-world scenarios since having access to more symptoms can significantly help the doctor make precise diagnoses. Further analysis about symptom inquiry efficiency will be discussed in Section IV-D.

Table V presents the results for datasets with a larger number of diseases. Several baselines, including the non-RL method BSODA and the RL-based method MMF-AC, have been enhanced to improve diagnostic performance in a

TABLE VI
ABLATION STUDY OF STATUS MODELING.

	Dxy			GMD-12		
	DAcc	SRec	ATurn	DAcc	SRec	ATurn
SA-MPF	**87.5**	91.3	**16.2**	**87**	89	**15.4**
w/o status train	85.6	92.3	16.7	86.6	89.9	16.2
w/o-sym status train	86.5	**92.9**	16.9	86.2	**90.4**	16.3
w/o-dis status infer	86.5	91.3	**16.2**	85.8	89.7	16.6

larger disease space. Compared to these baselines, our SA-MPF achieves significant performance improvements. the top-1 diagnostic accuracy of SA-MPF even surpasses the top-3 prediction accuracy of the baselines. For instance, SA-MPF achieves a top-1 accuracy of 86.55% on the SymCAT-200 dataset, while MMF-AC's top-3 accuracy is only 84.66%. Furthermore, as the disease scale increases, SA-MPF's performance decline is slower compared to the baselines, resulting in a more significant performance advantage. This indicates that SA-MPF can effectively adapt to larger disease spaces. Additionally, the average turns of SA-MPF are similar to or even lower than those of BSODA and MMF-AC. This suggests that SA-MPF maintains high efficiency of symptom inquiry in large disease spaces.

C. Ablation Study

In this study, we introduce masked symptom status prediction and masked disease status prediction to model symptom and disease statuses during the training phase. During inference, we used masked disease status prediction to exclude certain unreasonable diseases. As a result, we established three model variants for comparison , as shown in Table VI.

”w/o status train” indicates the exclusion of all masked status predictions, relying solely on masked token prediction. When comparing the results of ”w/o status train” with the baselines in Table IV, we observe that SA-MPF outperforms all baselines in diagnostic accuracy using only masked token prediction. This validates the effectiveness of unifying symptom inquiry and disease prediction into masked token prediction. From Table VI, we can observe that SA-MPF achieves not only higher diagnostic accuracy but also requires fewer average dialogue turns compared to ”w/o status train”. This demonstrates that status modeling can further enhance diagnostic accuracy and efficiency of masked token prediction. ”w/o-sym status train” means the exclusion of masked symptom status prediction. Without the modeling of symptom status, ”w/o-sym status train” necessitates more dialogue turns to gather additional symptoms. However, its diagnostic accuracy experiences a slight decrease. This implies that masked symptom status prediction can also aid in modeling the relationship between symptoms and diseases statuses, thereby improving the diagnostic accuracy of SA-MPF. ”w/o-dis status infer” means the removal of masked disease status prediction for disease exclusion during the inference phase, relying solely on masked disease token prediction to obtain the final disease prediction. Experimental results indicate that the combination of both status and token predictions for disease during inference results in higher diagnostic accuracy compared to token prediction alone.

TABLE VII
RESULTS WITH SMALLER DIFFERENT LIMITED TURNS.

Turn	Model	Dxy			SymCAT-90		
		DAcc	SRec	ATurn	DAcc	SRec	ATurn
5	Basic DQN	64.7	31.1	2.5	35.6	2.0	**2.0**
	HRL	70.2	15.2	**1.9**	44.3	2.4	4.3
	Diaformer	76.6	54.5	4.8	49.4	**46.1**	4.9
	MTDiag	76.1	**58.1**	5.0	51.1	44.1	5.0
	SA-MPF	**78.8**	57.4	5.0	**51.4**	45.8	5.0
10	Basic DQN	71.5	32.2	2.7	35.6	2.0	**2.0**
	HRL	71.8	15.9	**2.3**	48.8	30.7	7.4
	Diaformer	80.6	77.8	9.6	63.2	73.6	9.6
	MTDiag	81.9	**82.7**	9.6	63.6	72.5	10.0
	SA-MPF	**84.6**	78.7	10.0	**65.6**	**74.3**	9.9
15	Basic DQN	71.2	32.0	2.7	35.6	2.0	**2.0**
	HRL	71.8	15.9	**2.3**	49.9	32.2	8.3
	Diaformer	82.8	82.6	12.4	71.1	86.6	12.6
	MTDiag	85.4	89.8	11.9	73.3	87.9	14.0
	SA-MPF	**86.5**	**90.2**	14.8	**74.3**	**88.5**	13.6

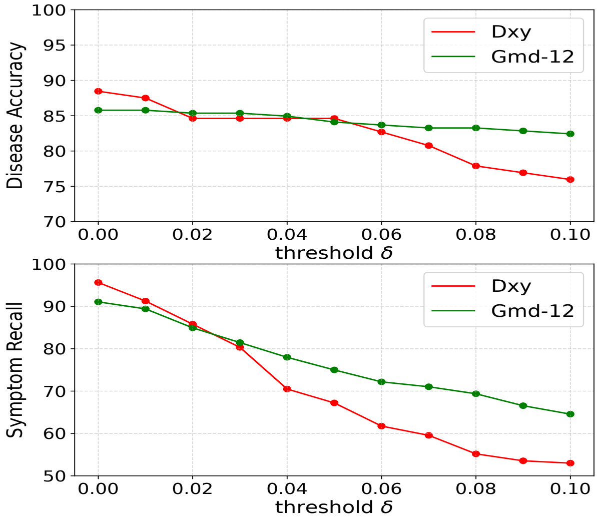

Fig. 2. The effect of the threshold δ.

D. Further Analysis

1) Effect of Smaller Different Limited Turns: As depicted in Table VII, our experiments involve a maximum of 5/10/15 dialogue turns. In comparison to RL-based methods, supervised learning approaches like the generative method Diaformer, the classification method MTDiag, and our SA-MPF tend to engage in more turns to achieve higher symptom recall. Despite this, SA-MPF still attains the highest diagnostic accuracy and competitive symptom recall within a limited number of turns. This underscores the capability of the proposed SA-MPF to deliver satisfactory performance in scenarios with a restricted number of dialogue turns.

2) Effect of Stopping Criterion Threshold δ: In Figure 2, the threshold value δ in SA-MPF is crucial for achieving a balance between diagnostic accuracy and efficiency. We observe that as the threshold δ increases, the recall of implicit symptoms tends to decrease, resulting in a reduction in diagnostic accuracy. This aligns with our intuition, as a higher threshold leads to an earlier termination of the inquiry process, potentially missing important implicit symptoms. However, it's important to note that the decrease in recall is more significant than the decrease in diagnostic accuracy, implying that our method prioritizes inquiring about key implicit symptoms in the early stages of the dialogue. In practical applications, the selection of the threshold δ should be guided by the desired trade-off between accuracy and efficiency.

V. Conclusion

In this paper, we propose SA-MPF, a masked prediction framework for automatic medical diagnosis. We reformulate symptom inquiry and disease prediction into a single masked token prediction task, with the primary difference being the input. It facilitates mutual learning between symptom inquiry and disease prediction, and alleviates the data scarcity issue simultaneously. Moreover, we introduce a masked status prediction task for disease status prediction and symptom status prediction, to enhance the modeling of status between symptoms and diseases. The experimental results on multiple datasets confirmed the effectiveness of SA-MPF.

References

[1] S. Fox and M. Duggan, "Health online 2013," *Health*, 2013.

[2] G. Zuccon, B. Koopman, and J. Palotti, "Diagnose this if you can," in *European Conference on Information Retrieval*. Springer, 2015, pp. 562–567.

[3] A. Keselman, A. C. Browne, and D. R. Kaufman, "Consumer health information seeking as hypothesis testing," *Journal of the American Medical Informatics Association*, vol. 15, no. 4, pp. 484–495, 2008.

[4] Y. Hu and J. Haake, "Search your way to an accurate diagnosis: Predictors of internet-based diagnosis accuracy," *Atlantic Journal of Communication*, vol. 18, no. 2, pp. 79–88, 2010.

[5] C. T. Lopes and C. Ribeiro, "Query behavior: The impact of health literacy, topic familiarity and terminology," in *International Conference on Human Factors in Computing and Informatics*. Springer, 2013, pp. 212–223.

[6] I. Puspitasari, "The impacts of consumer's health topic familiarity in seeking health information online," in *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2017, pp. 104–109.

[7] H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra, "Evaluation of symptom checkers for self diagnosis and triage: audit study," *bmj*, vol. 351, p. h3480, 2015.

[8] "Mayo clinic symptom checker," https://www.mayoclinic.org/symptom-checker/select-symptom/itt-20009075, 2015.

[9] "Webmd symptom checker," https://symptoms.webmd.com/, 2020.

[10] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, no. 3366, pp. 9–21, 1959.

[11] Q. T. Zeng, S. Kogan, R. M. Plovnick, J. Crowell, E.-M. Lacroix, and R. A. Greenes, "Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval," *International journal of medical informatics*, vol. 73, no. 1, pp. 45–55, 2004.

[12] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X.-J. Huang, K.-F. Wong, and X. Dai, "Task-oriented dialogue system for automatic diagnosis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 201–207.

[13] H.-C. Kao, K.-F. Tang, and E. Chang, "Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] Y.-S. Peng, K.-F. Tang, H.-T. Lin, and E. Chang, "Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis," *Advances in neural information processing systems*, vol. 31, 2018.

[15] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Lin, "End-to-end knowledge-routed relational dialogue system for automatic diagnosis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7346–7353.

[16] K. Liao, Q. Liu, Z. Wei, B. Peng, Q. Chen, W. Sun, and X. Huang, "Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning," *arXiv preprint arXiv:2004.14254*, 2020.

[17] C. Zhong, K. Liao, W. Chen, Q. Liu, B. Peng, X. Huang, J. Peng, and Z. Wei, "Hierarchical reinforcement learning for automatic disease diagnosis," *Bioinformatics*, vol. 38, no. 16, pp. 3995–4001, 2022.

[18] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[19] H. Cuayáhuitl, S. Keizer, and O. Lemon, "Strategic dialogue management via deep reinforcement learning," *arXiv preprint arXiv:1511.08099*, 2015.

[20] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.

[21] Z. Hou, Y. Cen, Z. Liu, D. Wu, B. Wang, X. Li, L. Hong, and J. Tang, "Mtdiag: an effective multi-task framework for automatic diagnosis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 241–14 248.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] J. Chen, D. Li, Q. Chen, W. Zhou, and X. Liu, "Diaformer: Automatic diagnosis via symptoms sequence generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 4432–4440.

[24] H. Wang, W. C. Kwan, K.-F. Wong, and Y. Zheng, "Coad: Automatic diagnosis through symptom and disease collaborative generation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 6348–6361.

[25] W. Liu, Y. Cheng, H. Wang, J. Tang, Y. Liu, R. Zhao, W. Li, Y. Zheng, and X. Liang, "" my nose is running."" are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics," in *Ijcai*, 2022.

[26] A. R. Inc, "Symcat: Symptom-based, computer assisted triage." *http://www.symcat.com*, 2017.

[27] K.-F. Tang, H.-C. Kao, C.-N. Chou, and E. Y. Chang, "Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning," in *NIPS Workshop on Deep Reinforcement Learning*, 2016.

[28] Y. Xia, J. Zhou, Z. Shi, C. Lu, and H. Huang, "Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 1062–1069.

[29] W. He and T. Chen, "Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4695–4703.

[30] A. Chattopadhyay, K. H. R. Chan, B. D. Haeffele, D. Geman, and R. Vidal, "Variational information pursuit for interpretable predictions," in *The Eleventh International Conference on Learning Representations*, 2023.

[31] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[32] M. S. Teixeira, V. Maran, and M. Dragoni, "The interplay of a conversational ontology and ai planning for health dialogue management," in *Proceedings of the 36th annual ACM symposium on applied computing*, 2021, pp. 611–619.

[33] W. He, X. Mao, C. Ma, Y. Huang, J. M. Hernàndez-Lobato, and T. Chen, "Bsoda: a bipartite scalable framework for online disease diagnosis," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2511–2521.