**Supplementary File for “AI-based Differential Diagnosis of Pemphigus Vulgaris and Bullous Pemphigoid: a case study in China”**

S-I. DETAILS TECHNICAL ILLUSTRATION OF THE CO-ATTENTION STRUCTURE

This section presents the detailed illustration of the co-attention structure designed in this paper.

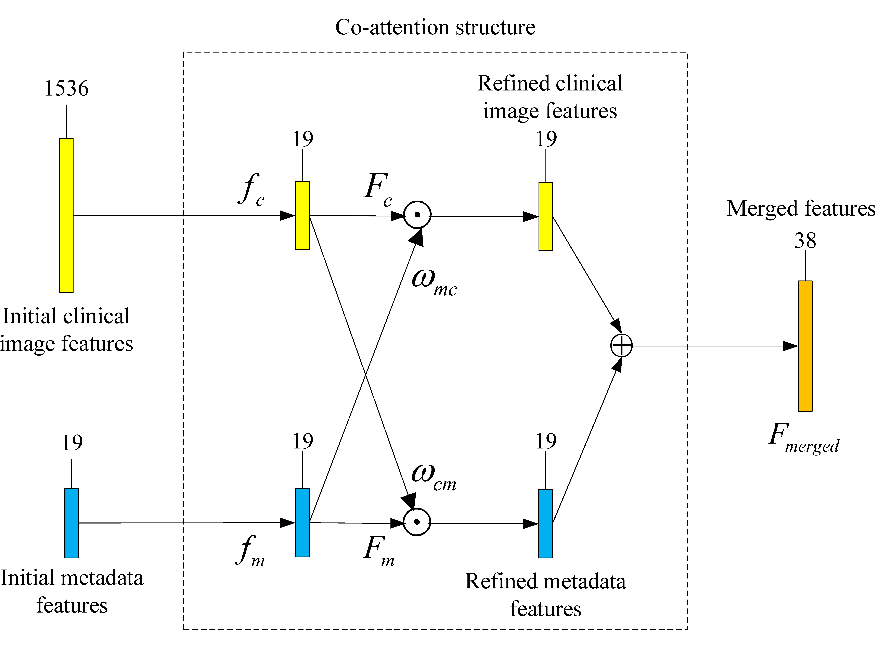


Fig. S-1. Overview of the co-attention structure.

Fig. S-1 presents the proposed co-attention structure. Overall, it includes two branches to process clinical image features and metadata features. In the first channel, the dimension of initial clinical image features  was reduced from 1536 to 19 by a 19-dimensional fully connected (FC) layer. Then, the 19-dimensional feature vector was activated by ReLU and Sigmoid to obtain  and , which served as clinical image features and clinical-image-guided metadata attention distribution, respectively.

 (S-1)

 (S-2)

where  and  represent the activations of ReLU and Sigmoid, respectively, and represent the parameters of the 19-dimensional FC layer in the clinical image branch. Similarly, in the second channel, initial metadata features  also went through a 19-dimensional FC layer and output metadata features  and metadata-guided clinical image attention distribution  as follows:

 (S-3)

 (S-4)

where  represent the parameters of the 19-dimensional FC layer in the metadata branch. By performing the pixel-wise product between the features and the corresponding attention distribution, the refined clinical image features and metadata features are obtained. Subsequently, the two refined features are concatenated to form the final merged features :

 (S-5)

where  denotes the pixel-wise product operation and  represent the feature concatenation operation. It can be seen from Equations (S-1)-(S-5) that  will be updated with the learnable parameters  and  in the training process and finally meet the goal of joint learning of modalities.