1 Ablation Study

To validate the effectiveness of the Structured Medical Concept Extraction Module and the Multi-view Representation Disentanglement Module, we conducted a series of ablation studies. As shown in Table III, we designed a comparative experimental setting involving four variants: w/o CM and w/o DM represent our full model without Structured Medical Concept Extraction Module and Multi-view Representation Disentanglement Module.

- w/o CM & DM: BioViL model without either module
- w/o DM: adding the structured concept extraction module only
- w/o CM: adding the multi-view disentanglement module only
- Full Model: integrating both modules

The ablation experimental results demonstrate a stepwise improvement in all evaluation metrics as modules are gradually introduced: the full model achieves an F1-score of 0.290 and AUC of 0.744, outperforming the model without both CM and DM (F1=0.111, AUC=0.495) by a large margin, confirming the effectiveness of both proposed modules.

Using large language models to extract professional medical manual text as prompts offers significant advantages over dataset medical reports, with two core benefits. On one hand, medical manuals are systematically compiled with high specialization and standardization, effectively avoiding individual bias in medical reports, which are influenced by doctors' personal styles and divergent regional medical practices, as proven by the report differences between the IU X-RAY and MIMIC-CXR datasets. On the other hand, medical manuals are authored by domain experts and rigorously reviewed, boasting authority and accuracy to provide a stable, reliable medical knowledge base for the model. In contrast, real-time clinical medical reports often omit basic medical principles and simplify key information due to urgent scenarios and case specificity. This fundamental difference makes model outputs more aligned with industry consensus when using medical manuals as prompts, reducing cognitive biases from individual report variations, which is validated by the sharp performance drop when removing the CM module (F1 from 0.290 to 0.144, AUC from 0.744 to 0.515).

When the multi-view visual feature disentanglement module (DM) is removed, all metrics decline consistently, with F1 dropping from 0.290 to 0.220 and AUC from 0.744 to 0.653. The core reason is that while frontal images provide more central disease-related information, lateral views are not redundant: their view-specific features complement the frontal view, and only full utilization of multi-view information enables comprehensive, 3D feature extraction, avoiding information blind spots from single-view reliance.

Technically, the DM's advantage also lies in its efficient two-view fusion. Most similar models only realize shallow utilization of multi-view data, while our method adopts precise disentanglement and semantic enhancement. Through dual-adapter shared-specific feature disentanglement, the shared adapter extracts cross-view common pathological signals, while view-specific adapters precisely retain unique anatomical diagnostic information for each view, effectively

Table 1. The results were obtained on the MIMIC dataset, where CM denotes the use of the concept extraction module and DM denotes the use of the disentanglement module.

Model Configurations	F1	Recall	Precision	auc
w/o CM & DM	0.111	0.157	0.095	0.495
w/o CM	0.144	0.223	0.141	0.515
w/o DM	0.220	0.386	0.161	0.653
SCD-CXR(Full Model)	0.290	0.400	0.232	0.744

preventing specific features from being obscured by global lung field signals. Furthermore, most two-view models only merge at the visual feature level, while our method semantically binds specific features with medical concepts, upgrading two-view fusion from meaningless pixel feature accumulation to meaningful diagnostic complementary information.