MedKCO: Medical Vision-Language Pretraining via Knowledge-Driven Cognitive Orchestration

Supplementary Material

7. Pretraining and Evaluation Datasets

The proposed model and all baseline models were pretrained using both label-level and description-level data. The datasets used for each modality are summarized in Tab. 8.

For the CFP modality, pretraining data span more than 96 fundus disease categories and include over 190k samples. For the OCT modality, the datasets cover more than 17 disease categories with over 180k samples. For the CXR modality, the pretraining datasets contained more than 14 disease categories and over 380k pretraining samples. During pretraining, we used only the MIMIC-CXR training set for the CXR modality.

For the zero-shot image-to-text retrieval task, we used the entire OpenI dataset for evaluation. For MIMIC-CXR, following CXR-CLIP [42], we evaluated performance solely on its test set.

For the report generation task, we split the OpenI dataset into training, validation, and test sets by a ratio of 70%, 15% and 15%, respectively. For the MIMIC-CXR dataset, we trained the model on its training set and evaluated the results on its test set.

8. Detail of Label-level Curriculum

When it comes to the label-level data of the three modalities used in this work, we first compiled all disease categories present in the datasets listed in Sec. 7, and then grouped these categories into three stages according to the classification criteria defined in Sec. 3.2. The class labels for each stage are presented in Tab. 9.

9. Detail of the Baseline Models

For the **CFP** and **OCT** modality, following the data processing procedures and model architectures adopted in FLAIR [26] and KeepFIT [35], we replaced categorical labels with descriptive phrases generated from domain knowledge templates to construct description-level supervision. In terms of model architecture, consistent with [26, 35], we utilized ResNet50 as the vision encoder and Bio-ClinicalBERT as the text encoder. The models were pretrained using the CLIP InfoNCE loss function and FILIP fine-grained interactive language image pretraining loss. For the **CXR** modality, following MedCLIP [32] and CXR-CLIP [42], we similarly replaced disease labels of varying confidence levels with descriptions generated from domain-knowledge templates. For the model architecture, we followed MedCLIP and CXR-CLIP by adopting ResNet50 as the vision encoder and Bio-ClinicalBERT as the text encoder. Pretraining used both the InfoNCE loss and the FILIP loss.

Additionally, we compared the proposed method with several curriculum learning baselines. Self-paced learning [15, 19] is a training paradigm that dynamically adjusts the training process according to model feedback. In [15], the loss value of each sample indicates its learning difficulty: samples with higher loss are treated as harder and are assigned lower weights in the early stages, with their weights increasing gradually as training progresses. This process is controlled by a loss threshold γ:

$$\mathcal{L} = w_i \mathcal{L}_i(p, y) + \lambda(w_i) \tag{8}$$

$$w_i = \frac{1 + \exp(-\gamma)}{1 + \exp(\mathcal{L}_i - \gamma)}, \tag{9}$$

$$\mu_i = 1 + \exp(-\gamma) - w_i, \tag{10}$$

$$\lambda(w_i) = \mu_i ln(\mu_i) + w_i ln(w_i + \delta) - \gamma w_i, \tag{11}$$

where $\mathcal{L}_i(\cdot)$ represents the loss function for sample i, p and y correspond to the output logits and the target label, δ is a hyperparameter, which we set $1e^{-8}$ here. The threshold γ was gradually increased by a fixed amount at each epoch. For vision-language contrastive pretraining under both the CLIP and FILIP frameworks, $\mathcal{L}_i(\cdot)$ corresponds to symmetric contrastive loss.

In [3], the learning difficulty of each sample is determined by the model's output logits and a hyperparameter γ. A lower logit indicates that the sample is relatively difficult and that it will initially be assigned a lower weight.

$$\mathcal{L} = -|p_t|^\gamma \mathcal{L}_i(p, y), \tag{12}$$

where p_t denotes the logit corresponding to the ground-truth class, with larger values indicating easier samples and smaller values indicating harder ones. The parameter γ was adjusted to modulate the weight of easy and difficult samples throughout training, p and y correspond to the output logits and the target labels, respectively. In vision-language contrastive pretraining for the CLIP and FILIP frameworks, $\mathcal{L}_i(\cdot)$ refers to the symmetric contrastive loss.

Table 8. Pretraining datasets of each modality.

Modality	Label-level	Description-level
CFP	EYEPACS, IDRID, RFMid, DEN, LAG, ODIR, PAPILA, PARAGUAY, STARE, ARIA, AGAR300, APTOS, FUND-OCT, JICHI, DiaRetDB1, DRIONS-DB, Drishti-GS1, E-ophta, G1020, HRF, ORIGA, ROC, BRSET, OIA-DDR, AIROGS, SUSTech-SYSU, CHAKSU, DR1-2, Cataract, ScarDat	MM-Retinal CFP
OCT	RetinalOCT_C8, Large_Dataset_of_Labeled_OCT, GAMMA1, STAGE1, STAGE2, glaucoma_detection, GOALS, OIMHS, OCTA_500, DUKE_DME, BIOMISA_Retinal_Image_Database_for_Macular_Disorders	MM-Retinal OCT
CXR	CheXpert	MIMIC-CXR

Table 9. The disease at different stage of each modality in label-level curriculum

Modality	Stage	Label
CFP	Stage 1	hard exudates, soft exudates, microaneurysms, haemorrhages, media haze, drusens, tessellation, laser scar, optic disc cupping, tortuous vessels, asteroid hyalosis, optic disc pallor, exudates, cotton wool spots, colobomas, preretinal haemorrhage, myelinated nerve fibers, tilted disc, vitreous haemorrhage, large optic cup, optic atrophy, fibrosis, silicon oil, scar, nevus, red small dots
	Stage 2	no diabetic retinopathy, mild diabetic retinopathy, moderate diabetic retinopathy, severe diabetic retinopathy, proliferative diabetic retinopathy, age-related macular degeneration, pathologic myopia, macular scar, shunt, branch retinal vein occlusion, epiretinal membrane, central retinal vein occlusion, optic disc edema, retinal traction, retinitis, retinal pigment epithelium changes,retinitis pigmentosa, haemorrhagic retinopathy, central retinal artery occlusion, post traumatic choroidal rupture, choroidal folds, vasculitis, plaque, branch retinal artery occlusion, collaterals, maculopathy, severe hypertensive retinopathy, dragged disk, disc swelling and elevation, congenital disk abnormality, yellow-white spots flecks, abnormal macula, peripheral retinal degeneration and break, no proliferative diabetic retinopathy, hypertensive retinopathy, geographical age-related macular degeneration, abnormal optic disc, abnormal vessels, macular edema, increased cup disc, a disease, intraretinal microvascular abnormalities, retina detachment, normal
	Stage 3	diabetic macular edema, no referable diabetic macular edema, non clinically significant diabetic macular edema, central serous retinopathy, anterior ischemic optic neuropathy, parafoveal telangiectasia, chorioretinitis, macular hole, optic disc pit maculopathy, haemorrhagic pigment epithelial detachment, Vogt-Koyanagi syndrome, glaucoma, Bietti crystalline dystrophy, neoplasm, no glaucoma, neovascular age-related macular degeneration, cataract, no cataract, macroaneurysm, cystoid macular edema, acute central serous retinopathy, chronic central serous retinopathy, neovascularisation
OCT	Stage 1	macular hole stage3, macular hole stage4, vitreomacular Interface Disease, epiretinal membrane
	Stage 2	age related macular degeneration, drusen, diabetic macular edema, macular hole stage1, macular hole stage2, normal, macular hole, central serous retinopathy, choroidal neovascularization
	Stage 3	glaucoma, diabetic retinopathy, retinal artery occlusion, retinal vein occlusion
CXR	Stage 1	Lung Opacity, Consolidation, Support Devices
	Stage 2	No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Edema, Pneumonia, Atelectasis, Pneumothorax, Hernia, Pleural Effusion, Emphysema, Infiltration, Mass
	Stage 3	Lung Lesion, Pleural Other, Fracture, Fibrosis, Nodule, Pleural_Thickening

10. Implementation Detail

FILIP framework. FILIP[40] computes fine-grained image–text and text–image similarities by leveraging sequences of local image and text features. The original FILIP adopts a ViT-based vision encoder. In this paper, we replaced it with ResNet50 for consistency with FLAIR, KeepFit, KeepFit v2, Med-CLIP and CXR-CLIP. Specifically, in the output part of the vision encoder, we removed the global average pooling layer and the corresponding normalization layer, obtaining a feature map of size $B \times H \times W \times C$.

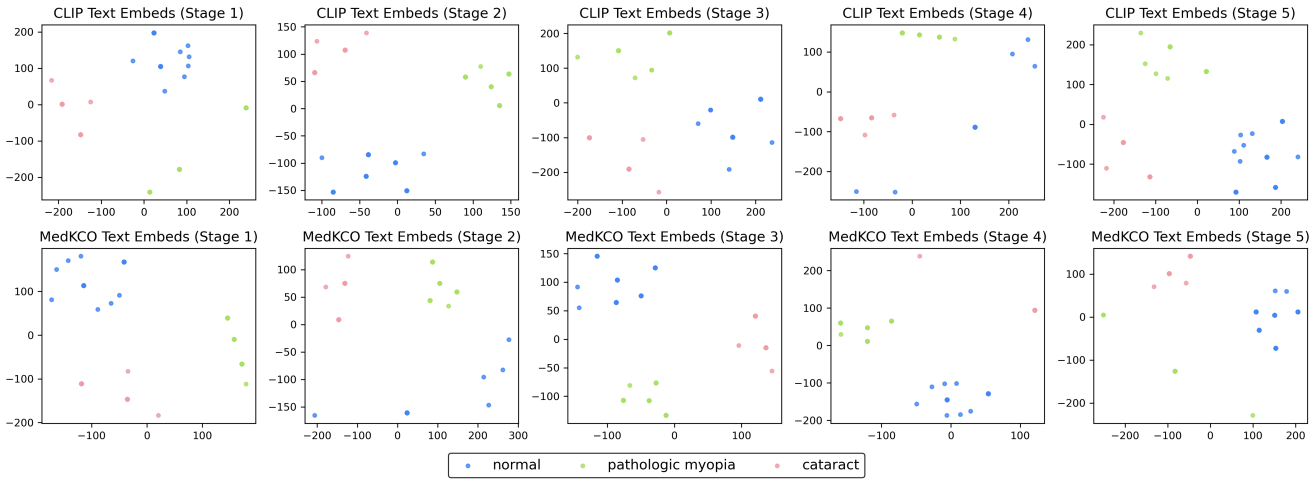

Figure 5. Visualization of text feature at different stages on ODIR200×3 dataset under the CLIP framework.

The tensor was then flattened into a feature sequence of size $B \times (HW) \times C$.

Data augmentation. For CFP and OCT modality, following the design of FLAIR [26] and KeepFIT [35] series, we resized the image to 512×512. During pretraining, we applied random image augmentations including horizontal flips, random rotations of $[-5, 5]$ degrees, zoom scaling sampled from $[0.9, 1.1]$, and color jitter. We used AdamW with a base learning rate of $1e^{-4}$ and weight decay of $1e^{-5}$. For the CXR modality, following MedCLIP [32] and CXR-CLIP [42], we resized the image to 256×256. Image augmentations were applied using horizontal flipping with 0.5 probability; color jittering with brightness and contrast ratios of $[0.8; 1.2]$; random affine transformation with degree sampled from $[-10; 10]$, maximum translation rate 0.0625 and scale factor in $[0.8; 1.1]$. We used AdamW with a base learning rate of $5e^{-5}$ and a weight decay of $1e^{-4}$.

11. Report Generation

For the report generation task, we used the vision encoder of the pretrained foundation model to obtain a sequence of visual features. Specifically, with the ResNet50 architecture used in this paper, we extracted the feature maps before the final average pooling layer and flatten their spatial dimensions. The resulting sequence of visual features was first processed by a Transformer encoder to form visual memory. Then, we concatenated the text tokens with this visual memory and fed them into the Transformer decoder. The training was conducted under an auto-regressive manner.

12. Visualization of the Text Embedding

Since the text encoders of medical VLPs are typically initialized with models pretrained on large-scale medical corpora, they inherently possess strong semantic modeling capabilities across a wide range of diseases. Consequently, textual features tend to exhibit a relatively sparse distribution in the early stage of VLP pretraining. As shown in Fig. 5, both MedKCO and CLIP demonstrated strong disease modeling abilities through their text encoders at different stages of pretraining. Meanwhile, Fig. 3 shows that MedKCO achieved a more powerful visual representation modeling performance. Together, these results indicate that the proposed knowledge–driven cognitive orchestration method effectively enhances the ability of medical VLP to model disease-related image features.