Hierarchical Conditional Energy Modeling for Medical Vision–Language Pretraining

Chengsheng Mao , Yuan Luo
Department of Preventive Medicine, Northwestern University Feinberg School of Medicine
{chengsheng.mao, yuan.luo}@northwestern.edu

Abstract

Contrastive vision–language pretraining models such as CLIP align images and text in a shared embedding space but do not explicitly model or evaluate the hierarchical semantics common in medical image interpretation. We propose HCE-CLIP (Hierarchical Conditional Energy CLIP), a vision–language pretraining framework that formulates medical image–text alignment as a hierarchical label-conditional energy modeling problem. HCE-CLIP encodes an image series using transformer-based aggregation and aligns it with free-text reports and structured label state prompts across multiple semantic levels. At each level, conditional energy functions favor clinically consistent label states while suppressing contradictory alternatives, supporting uncertainty-aware inference. To assess semantic coherence, we introduce a hierarchical contradiction metric that quantifies logical inconsistencies between fine-grained disease predictions and higher-level clinical summaries. Experiments on MIMIC-CXR and other public benchmarks show that HCE-CLIP outperforms existing medical vision–language pretraining methods in seen-label, zero-shot and linear-probe settings, while producing substantially fewer hierarchical contradictions.

1 Introduction

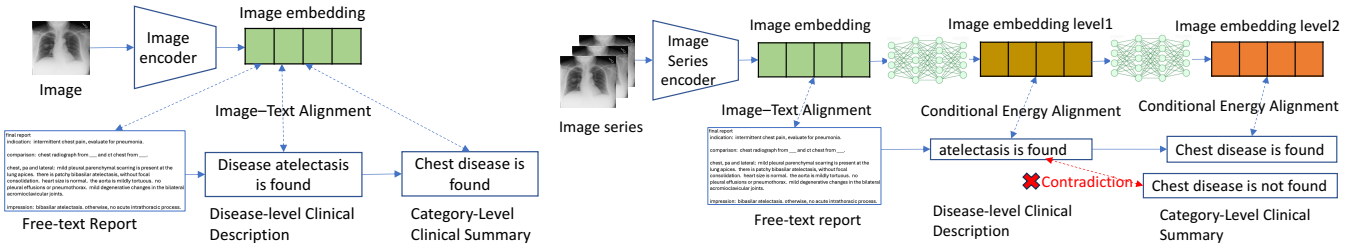
Contrastive language-image pre-training (CLIP) [Radford *et al.*, 2021] has shown remarkable success in learning transferable representations by aligning images and text within a shared embedding space. In the medical domain, recent adaptations of CLIP demonstrate promising performance for tasks such as disease classification and report retrieval [Wang *et al.*, 2022; Zhang *et al.*, 2023; Wu *et al.*, 2023; Huang *et al.*, 2021a]. However, medical image interpretation differs from general vision–language settings in two important ways that remain insufficiently addressed. First, medical semantics are inherently hierarchical. Clinical findings are often described at multiple levels of abstraction, ranging from fine-grained disease entities (e.g., atelectasis) to higher-level diagnostic summaries (e.g., abnormal chest disease). These levels are not independent; rather, they obey structured semantic relationships that conventional CLIP-style alignment does not explicitly model. Second, medical reports typically describe an entire imaging study rather than individual images. A single radiology report may summarize findings across multiple views or slices, and specific abnormalities may only be visible in a subset of images. Treating each image–text pair independently therefore introduces semantic mismatch and weakens the alignment signal.

Most existing medical vision–language models typically align a single image embedding with multiple textual descriptions corresponding to different semantic levels (Fig. 1a), implicitly assuming that one representation can simultaneously capture all levels of abstraction. This assumption can lead to under-representation of higher-level clinical semantics and inconsistent predictions across levels (e.g., predicting a disease as present while simultaneously predicting "No Finding"). Such consistency is particularly important for clinical reliability and interpretability.

To address these challenges, we propose HCE-CLIP (Hierarchical Conditional Energy CLIP), a hierarchical conditional energy-based framework for medical vision–language pretraining. HCE-CLIP represents an imaging study as a unified image series embedding using a transformer-based aggregation mechanism [Vaswani *et al.*, 2017], and models the compatibility between image series, texts, and structured clinical label states using temperature-scaled energy functions across multiple semantic levels (Fig. 1b). This formulation explicitly favors clinically consistent label states while suppressing contradictory alternatives, providing a principled mechanism for enforcing hierarchical semantic coherence.

To quantitatively assess semantic coherence, we further introduce a hierarchical contradiction-based evaluation metric that measures the extent to which model predictions violate predefined clinical hierarchy constraints. This metric directly evaluates whether fine-grained disease predictions are consistent with higher-level diagnostic summaries.

We pretrain HCE-CLIP on the MIMIC-CXR dataset [Johnson *et al.*, 2019] and evaluate it on multiple public benchmarks, including MIMIC-CXR, CheXpert [Irvin *et al.*, 2019], and ChestX-ray14 [Wang *et al.*, 2017]. Experimental results demonstrate that HCE-CLIP improves classfication performance on both seen and unseen disease categories, enhances representation transferability, and substantially reduces hi-



(a) conventional CLIP (b) HCE-CLIP

Figure 1: Comparison between conventional CLIP and HCE-CLIP for hierarchical medical semantics. (a) Conventional CLIP aligns a single image embedding with multiple textual descriptions at different semantic levels. (b) HCE-CLIP performs hierarchical alignment using level-specific image representations and enforces semantic consistency through conditional energy modeling over label states.

erarchical contradictions compared to existing medical vision–language models. These findings suggest that hierarchical conditional energy modeling provides an effective and general framework for medical vision–language representation learning. Our codes are available at https://github.com/mocherson/HCE-CLIP

In summary, our contributions are threefold. (1) We propose HCE-CLIP, a hierarchical conditional energy-based framework for medical vision-language pretraining that explicitly models the semantic compatibility between image series, radiology reports, and structured clinical label states across multiple levels of abstraction. (2) We introduce a Hierarchical Contradiction Rate (HCR) metric to quantitatively evaluate semantic consistency across label hierarchies, enabling systematic analysis of logically coherent clinical predictions. (3) Extensive experiments demonstrate that HCE-CLIP achieves superior classification performance on seen and unseen diseases, improved representation transferability, and substantially reduced hierarchical contradictions compared to existing medical vision-language models.

2 Related Work

2.1 Vision-Language Representation

Vision–language representation learning has been widely studied using both dual-stream and single-stream architectures for multimodal fusion [Jia *et al.*, 2021; Li *et al.*, 2021; Chen *et al.*, 2020]. Early work focused on aligning visual and textual modalities through contrastive objectives, enabling models to learn transferable representations from large-scale image–text data. Subsequent studies explored incorporating external knowledge [Cui *et al.*, 2021; Li *et al.*, 2020; Yu *et al.*, 2021] and refining pretraining objectives such as masked language and visual modeling [Huang *et al.*, 2021b; Wang *et al.*, 2023].

More recently, CLIP-style contrastive learning has become a dominant paradigm for scalable vision–language pretraining [Radford *et al.*, 2021; Li *et al.*, 2022b; Mu *et al.*, 2022; Yao *et al.*, 2022; Chen *et al.*, 2023]. Some works have investigated hierarchical or structured representations within contrastive frameworks. For example, HiCLIP [Geng *et al.*, 2023] introduces hierarchy-aware attention mechanisms, and other approaches improve robustness or generalization by modeling semantic relationships across modalities [Ge *et al.*, 2023]. Instruction-tuned multimodal models like LLaVA-Med [Li *et al.*, 2023] and Dr-LLaVA [Sun *et al.*, 2024] extend CLIP-style architectures with biomedical-specific tuning, enabling better generalization in clinical settings. LoGra-Med [Nguyen *et al.*, 2024] further augments alignment through multi-graph modeling to handle complex long-context scenarios in medical reports. However, these methods primarily focus on architectural or attention-based designs and do not explicitly model label-state semantics or structured opposition between clinically consistent and inconsistent concepts. In contrast, our work formulates vision–language alignment as a hierarchical label-conditional energy modeling problem, where semantic consistency across abstraction levels is enforced through structured energy functions rather than attention mechanisms alone.

2.2 Medical Image Understanding

Medical image understanding is a pivotal area in healthcare, and extensive research has been devoted to advancing the capabilities of computer systems in this domain [Wang *et al.*, 2017; Mao *et al.*, 2018; Mao *et al.*, 2022; Yang *et al.*, 2022]. Medical vision–language models have been adapted to align radiology images with free-text reports for tasks such as disease classification and report retrieval. MedCLIP [Wang *et al.*, 2022] decouples image–text pairing to increase training scalability, while MedKLIP [Wu *et al.*, 2023] incorporates structured medical knowledge into the alignment process. GLoRIA [Huang *et al.*, 2021a] learns both global and local representations to improve label efficiency in medical image recognition. CheXzero [Zhang *et al.*, 2023] leverages domain-specific knowledge to guide pretraining on chest radiographs. MedMax [Bansal *et al.*, 2024] introduced mixed-modal instruction tuning to scale vision-language assistants for biomedicine.

While these approaches improve medical image–text alignment, they typically rely on aligning individual images with textual descriptions and do not explicitly account for the fact that radiology reports summarize entire imaging studies. Moreover, hierarchical clinical semantics are usually treated implicitly through prompt design or supervision, rather than being modeled as structured semantic constraints.

Our method differs by encoding an image series as a uni-

fied representation and aligning it jointly with reports and structured label prompts across multiple semantic levels. By modeling label states through conditional energy functions, HCE-CLIP explicitly enforces consistency between supported and contradictory clinical concepts, enabling more structured and interpretable alignment.

3 Method

3.1 Hierarchical Label-Conditional Energy

Energy-based models (EBMs) provide a flexible framework for modeling structured relationships by assigning a scalar energy to each configuration of variables, where lower energy indicates higher compatibility [LeCun *et al.*, 2006]. EBMs define unnormalized compatibility functions and can be trained using discriminative objectives [Xie *et al.*, 2016; Du and Mordatch, 2019]. This property makes them well-suited for modeling multimodal and structured prediction problems.

We formulate medical vision–language pretraining as a hierarchical label-conditional energy modeling problem. Given an image series x, its associated report t, a medical label l, and a clinical state $s \in \mathcal{S}$ (positive, negative, or uncertain), the model assigns an energy value to each configuration, where lower energy indicates stronger semantic compatibility.

At each hierarchy level k, HCE-CLIP defines temperature-scaled energy functions for both the image–label–state and text–label–state relationships:

$$\mathcal{E}_x^{(k)}(x,l,s) = -\frac{1}{\tau_k}\langle E_x^{(k)}, E_{l,s}^{(k)}\rangle, \tag{1}$$

$$\mathcal{E}_t^{(k)}(t,l,s) = -\frac{1}{\tau_k}\langle E_t^{(k)}, E_{l,s}^{(k)}\rangle, \tag{2}$$

where $E_x^{(k)}$ and $E_t^{(k)}$ denote the image-series and report embeddings, $E_{l,s}^{(k)}$ denotes the corresponding label-state embedding, and τ_k is a learnable temperature parameter that controls the sharpness of the energy distribution at level k. $\langle , \rangle$ is the dot product of two vectors. At each hierarchy level k, the label l is interpreted according to the corresponding semantic granularity (e.g., specific disease-level or coarse category-level), with distinct prompt sets and embeddings used to represent the label states at that level.

The conditional distributions over label states are defined as:

$$p_x^{(k)}(s \mid x,l) = \frac{\exp(-\mathcal{E}_x^{(k)}(x,l,s))}{\sum_{s'\in\mathcal{S}}\exp(-\mathcal{E}_x^{(k)}(x,l,s'))}, \tag{3}$$

$$p_t^{(k)}(s \mid t,l) = \frac{\exp(-\mathcal{E}_t^{(k)}(t,l,s))}{\sum_{s'\in\mathcal{S}}\exp(-\mathcal{E}_t^{(k)}(t,l,s'))}. \tag{4}$$

Learning proceeds by minimizing the negative log-likelihood of the ground-truth state s^* for both modalities:

$$\mathcal{L}_{\text{HCE}}^{(k)}(x,t,l,s^*) = -\log p_x^{(k)}(s^* \mid x,l) - \log p_t^{(k)}(s^* \mid t,l), \tag{5}$$

which enforces semantic consistency between the image, the report, and the corresponding clinical label state.

This objective encourages the model to assign lower energy to clinically consistent label states while increasing the energy of contradictory or uncertain alternatives. Such conditional energy formulations have been shown to be effective for structured classification and discriminative modeling [Grathwohl *et al.*, 2021; Xie *et al.*, 2021].

In addition to label-conditional energy modeling, HCE-CLIP preserves the contrastive alignment between image series x and free-text reports t through a temperature-scaled cross-modal energy formulation:

$$\mathcal{E}^{(0)}(x,t) = -\frac{1}{\tau_0}\langle E_x, E_t\rangle, \tag{6}$$

where τ_0 is a learnable temperature parameter.

The symmetric CLIP objective minimizes both the image-to-text and text-to-image conditional negative log-likelihoods:

$$\mathcal{L}_{\text{CLIP}}(x,t) = -\log p(t \mid x) - \log p(x \mid t), \tag{7}$$

where

$$p(t \mid x) = \frac{\exp(-\mathcal{E}^{(0)}(x,t))}{\sum_{t'}\exp(-\mathcal{E}^{(0)}(x,t'))}, \tag{8}$$

$$p(x \mid t) = \frac{\exp(-\mathcal{E}^{(0)}(x,t))}{\sum_{x'}\exp(-\mathcal{E}^{(0)}(x',t))}. \tag{9}$$

The full training loss combines cross-modal alignment and hierarchical label-conditional energy learning:

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}}(x,t) + \sum_{k=1}^{K}\sum_{l}\mathcal{L}_{\text{HCE}}^{(k)}(x,t,l,s) \tag{10}$$

which jointly optimizes image–report alignment and hierarchical label-conditional energy separation across both visual and textual modalities.

3.2 HCE-CLIP Architecture

We now describe how the proposed hierarchical label-conditional energy model is instantiated within a practical vision-language pretraining architecture. This section introduces the HCE-CLIP model design, including the image series encoder, text encoder, and hierarchical representation modules used to realize the energy functions defined above.

Figure 2 illustrates the overall architecture and training pipeline of HCE-CLIP. Given an image series x, its associated report t, and a set of medical labels $\{l\}$ with corresponding clinical states at each semantic level, the model learns to align visual, textual, and structured label information within a unified energy-based framework.

The image series is first encoded into a unified representation using a transformer-based aggregation module, while the report is encoded by a text encoder. Both representations are then projected into a shared embedding space and further transformed into hierarchical representations that capture progressively abstract medical semantics. In parallel, each label is converted into multiple natural-language prompts corresponding to different clinical states (e.g., positive, negative, uncertain) and encoded into label-state embeddings.

HCE-CLIP defines hierarchical label-conditional energy functions between image series, reports, and label-state embeddings, encouraging clinically consistent configurations to have low energy while suppressing contradictory alternatives.

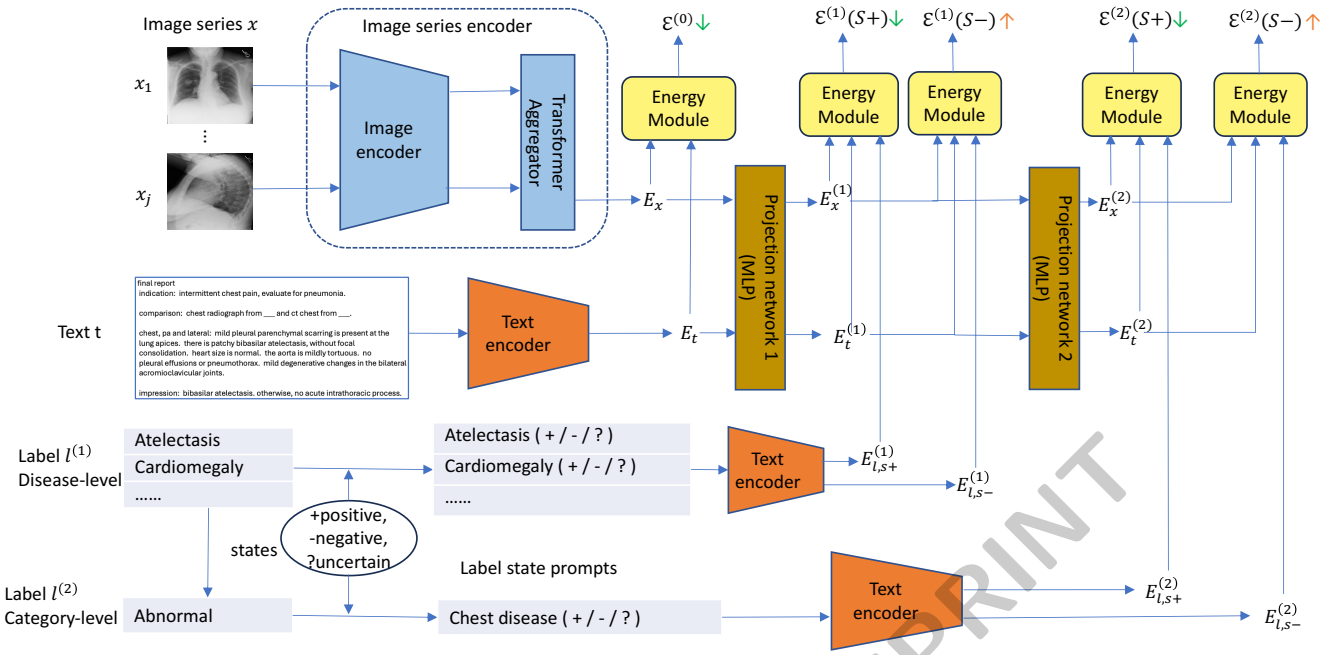


Figure 2: Overview of the HCE-CLIP framework under a hierarchical label-conditional energy modeling formulation. An image series is encoded into a unified representation using a transformer-based aggregator, while the associated report text and structured label-state prompts are encoded by a text encoder. Hierarchical projections produce multi-level representations. At each level, HCE-CLIP defines temperature-scaled energy functions between image series, reports, and label–state embeddings. Training minimizes the conditional energy of clinically consistent label states while increasing the energy of contradictory alternatives.

Image Series Encoding. Radiology reports describe an entire imaging study rather than individual images. We therefore encode an image series $x = (x_1, \cdots, x_j)$ as a unified representation. In our framework, the image series encoder consists of an image encoder and a multi-layer transformer. The image encoder processes each image in the series to produce an individual image embedding. The transformer then aggregates these individual embeddings into a unified image series embedding. A projection head is applied to transform the raw embedding to a specified dimension. The process is denoted as

$$E_{i_j} = \text{Enc}_i(x_j); \tag{11}$$

$$E'_x = \text{Transformer}([E_{i_1}, \cdots, E_{i_j}]); \tag{12}$$

$$E_x = f_s(E'_x) \tag{13}$$

where Enc_i is the image encoder; x_j is the jth image in the input image series; E'_x is the the raw output embedding of the image series through the transformer; f_s is the projection head that maps the output embeddings E'_x to a specified dimension E_x.

Text Encoding. The text encoder is to encode a text into a raw text embedding that is also projected to a specified dimension by the projection head, denoted as

$$E'_t = \text{Enc}_t(t); \qquad E_t = f_t(E'_t) \tag{14}$$

where Enc_t is the text encoder; t is the input text; f_t is the projection head that maps the raw output embeddings E'_t to a specified dimension E_t. E_t should have the same dimension with E_x for contrastive learning. The text encoder is also used to encode label-state prompts.

Hierarchical Representations. Medical semantics are inherently hierarchical, with fine-grained disease descriptions forming the basis for more abstract clinical summaries. To capture this structure, HCE-CLIP constructs multi-level representations for both image series and reports using cascaded transformation networks. Starting from the base embeddings E_x and E_t obtained from the image-series encoder and text encoder, respectively, we apply a sequence of multilayer perceptrons (MLPs) to progressively extract higher-level semantic features:

$$E_x^{(1)} = \text{MLP}_1(E_x), \qquad E_t^{(1)} = \text{MLP}_1(E_t), \tag{15}$$

$$E_x^{(2)} = \text{MLP}_2(E_x^{(1)}), \qquad E_t^{(2)} = \text{MLP}_2(E_t^{(1)}). \tag{16}$$

Each transformation produces representations corresponding to a different level of semantic abstraction. Lower levels retain fine-grained visual and textual details, while higher levels encode more abstract clinical concepts. These hierarchical embeddings are subsequently aligned with label-state prompt embeddings using the proposed label-conditional energy formulation, enabling structured semantic consistency across abstraction levels. This cascaded representation design allows higher-level clinical concepts to regularize lower-level representations, promoting coherent and interpretable multimodal alignment.

Label-State Prompt Encoding. Medical labels often admit multiple semantic states, reflecting the presence, absence, or uncertainty of a clinical finding. In HCE-CLIP, we model this structure explicitly by associating each med-

ical label l with a set of states corresponding to $S = \{positive, negative, uncertain\}$.

In our experiments on the MIMIC-CXR dataset, we consider 14 predefined labels, including 12 disease categories, *support devices*, and *no finding*. The label *no finding*, which summarizes the absence of pathology, is treated as a higher-level (category-level) semantic concept, while the remaining labels are treated as disease-level concepts.

For each label l and state $s \in S$, we construct a natural-language prompt to represent the label in that state. For example, for the label *Atelectasis*, the prompts to represent positive, negative, and uncertain clinical states are:

- *Disease Atelectasis is found.*
- *Disease Atelectasis is not found.*
- *Not sure if Disease Atelectasis is found.*

Each label-state prompt is encoded using the text encoder to obtain a corresponding label-state embedding:

$$P_{l,s} = \text{Prompt}(l, s), \quad E'_{l,s} = \text{Enc}_t(P_{l,s}), \quad E_{l,s} = f_l(E'_{l,s}) \tag{17}$$

where $\text{Prompt}(\cdot)$ generates the label-state prompts, Enc_t denotes the text encoder, and f_l is a projection head that maps the embeddings to the appropriate dimensionality.

To ensure compatibility with the hierarchical energy formulation, the embedding dimension of disease-level label-state representations $E^{(1)}_{l,s}$ matches that of the corresponding image and text representations $E^{(1)}_x$ and $E^{(1)}_t$, while the category-level label-state representations $E^{(2)}_{l,s}$ match the dimensionality of $E^{(2)}_x$ and $E^{(2)}_t$. This design enables structured alignment between multimodal representations and hierarchically organized clinical semantics.

Energy Modules. HCE-CLIP employs a set of energy modules to compute scalar compatibility scores between image representations, text embeddings, and label-state prompts at different semantic levels. Each energy module implements an energy function defined in Eq. 1, Eq. 2, or Eq. 6, depending on the input modalities. During training, energies corresponding to the ground-truth label state are minimized, while energies that contradict the ground-truth label state are maximized. This formulation enables HCE-CLIP to explicitly enforce semantic compatibility and hierarchical consistency across disease-level and category-level concepts.

3.3 Training Objective

With the above architecture, HCE-CLIP produces aligned representations for image series ($E^{(k)}_x$), report texts ($E^{(k)}_t$), and structured label prompts ($E^{(k)}_{l,s}$) across multiple semantic levels. Based on these representations, the corresponding energy functions are defined under the hierarchical label-conditional energy modeling framework described in Section 3.1. The training objective of HCE-CLIP jointly optimizes cross-modal image–text alignment and hierarchical label-conditional energy learning.

Image–Text Alignment. At the base level, HCE-CLIP preserves the original CLIP objective by minimizing the symmetric image-to-text and text-to-image conditional negative log-likelihoods defined by the cross-modal energy $\mathcal{E}^{(0)}(x, t)$ (Eq. 7). This encourages globally consistent alignment between image series and their associated free-text reports.

Hierarchical Label-Conditional Learning. At each hierarchy level $k \geq 1$, the model is trained to assign lower energy to the ground-truth clinical label state and higher energy to contradictory or uncertain alternatives using the label-conditional energy functions $\mathcal{E}^{(k)}_x(x, l, s)$ and $\mathcal{E}^{(k)}_t(t, l, s)$ (Eqs. 1–5). This enforces semantic consistency among the image series, the report, and the structured label prompts across multiple levels of clinical abstraction.

Joint Optimization. The overall training loss combines the cross-modal alignment objective and the hierarchical label-conditional energy losses across all labels and hierarchy levels (Eq. 10). By optimizing these objectives jointly, HCE-CLIP learns representations that are simultaneously aligned across modalities and constrained by structured clinical semantics.

3.4 Inference

At inference time, HCE-CLIP predicts the clinical state of a medical label by minimizing the hierarchical label-conditional energy. For image classification, given an image series x and a target label l, the model evaluates the compatibility between the input image series and each possible clinical state $s \in \mathcal{S}$ by computing the image–label–state energy $\mathcal{E}_x(x, l, s)$ using Eq. 1. The final prediction is obtained by selecting the state with the minimum energy at the hierarchy level corresponding to l:

$$\hat{s} = \arg\min_{s \in \mathcal{S}} \mathcal{E}_x(x, l, s) \tag{18}$$

Text-based classification can be performed analogously by minimizing the text–label–state energy $\mathcal{E}_t(t, l, s)$.

This energy-based inference formulation provides a unified mechanism for zero-shot classification, uncertainty-aware prediction, and hierarchical semantic consistency. New medical concepts can be introduced at test time through textual label descriptions without retraining the image encoder.

3.5 Hierarchical Consistency Metric

Standard classification metrics (AUC/ACC/F1) do not measure whether predictions are logically consistent across the label hierarchy. For example, predicting 'Atelectasis = positive' while predicting 'No Finding = positive' constitutes a hierarchical contradiction. To quantify semantic inconsistency, we define the *Hierarchical Contradiction Rate* (HCR), the fraction of violated hierarchical constraints.

In our hierarchy, level-2 corresponds to *No Finding / Normal* (denoted as NF), which summarizes the absence of any abnormalities. Therefore, a logically consistent prediction should not simultaneously indicate that a specific disease is present and that the study is normal.

Let $\mathcal{L}_1$ denote the set of level-1 disease labels (excluding NF). For a test sample x_i, let $\hat{s}_{i,l} \in \mathcal{S}$ be the predicted clinical

	MIMIC-CXR test			CheXpert		
	AUC	**ACC**	**F1**	AUC	ACC	F1
HCE-H2-E-F	**0.8735**	0.8505	**0.5897**	**0.8921**	**0.8654**	**0.6973**
HCE-H2-NE-F	0.8543	0.8367	0.5587	0.8526	0.7723	0.6145
HCE-H1-E-F	0.8547	0.8489	0.5714	0.8601	0.8365	0.6413
HCE-H1-NE-F	0.8549	0.8367	0.5641	0.8578	0.7892	0.6297
HCE-H2-E-S	0.8617	**0.8559**	0.5631	0.8686	0.8535	0.6713
MedKLIP	0.5951	0.4456	0.2783	0.7993	0.7261	0.5474
MedCLIP	0.6063	0.3321	0.2673	0.5863	0.3271	0.3300

Table 1: Prompt-based classification results for seen labels on MIMIC-CXR test and CheXpert datasets

	seen diseases			unseen diseases		
	AUC	ACC	F1	AUC	ACC	F1
HCE-H2-E-F	**0.7975**	0.8428	**0.3345**	**0.6048**	**0.6231**	**0.1657**
HCE-H2-NE-F	0.7818	0.8401	0.3191	0.5588	0.5484	0.1575
HCE-H1-E-F	0.7843	0.8373	0.3275	0.5676	0.4938	0.1575
HCE-H1-NE-F	0.7815	**0.8546**	0.3252	0.5763	0.5583	0.1581
HCE-H2-E-S	0.7804	0.8416	0.3233	0.5860	0.5670	0.1581
MedKLIP	0.6922	0.7121	0.2268	0.5551	0.4880	0.1547
MedCLIP	0.4037	0.1532	0.1510	0.4591	0.3159	0.1230

Table 2: Prompt-based classification results for seen and unseen labels on ChestXray14 dataset

state of label l, where $\mathcal{S} = \{\text{positive}, \text{negative}, \text{uncertain}\}$. We define a contradiction for disease label l as:

$$\text{viol}_{i,l} = \mathbb{I}\big[\hat{s}_{i,l} = \text{positive} \wedge \hat{s}_{i,\text{NF}} = \text{positive}\big] \tag{19}$$

We report the HCR as the fraction of disease–normal contradictions across labels and samples:

$$\text{HCR} = \frac{1}{N|\mathcal{L}_1|} \sum_{i=1}^{N} \sum_{l \in \mathcal{L}_1} \text{viol}_{i,l} \tag{20}$$

where N is the number of test samples. Lower HCR indicates better hierarchical semantic consistency. Following common practice for uncertainty-aware evaluation, we do not count $\hat{s}_{i,\text{NF}} = \text{uncertain}$ as a contradiction.

4 Experiments

4.1 Datasets

Pretraining Dataset. We pretrain HCE-CLIP on MIMIC-CXR [Johnson *et al.*, 2019], a large public dataset of 377,110 chest radiographs, corresponding to 227,835 imaging studies paired with free-text radiology reports. MIMIC-CXR predefined 14 labels, including 12 disease, support devices, and no findings. We follow the official patient-level split and train on 222,758 studies, validating on 1,808 studies to select the best model.

Evaluation Dataset. (1) **MIMIC-CXR test set** is a split from MIMIC-CXR dataset. It contains 5,159 images from 3,269 studies with the same disease label set as the training set. This evaluates the performance on seen labels. (2) **CheXpert** [Irvin *et al.*, 2019], the test set contains 500 radiologist-annotated test studies with the same label set as MIMIC-CXR, enabling external validation on seen diseases. (3) **ChestX-ray14** [Wang *et al.*, 2017] contains 112,120 frontal-view X-rays with 14 disease labels. Seven labels overlap with MIMIC-CXR (seen), and seven are unseen, enabling zero-shot evaluation on novel disease categories. Each image is treated as a single-image study.

4.2 Implementation Details

We benchmarked on the Mask R-CNN [He *et al.*, 2017] with SwinTransformer [Liu *et al.*, 2021] followed by a feature pyramid networks (FPN) [Lin *et al.*, 2017] as the image encoder. We used the BERT model [Devlin *et al.*, 2019] as the backbone of text encoder. The image encoder and text encoder are initialized with the GLIP-T [Li *et al.*, 2022a] model. The original text embedding dimension output from text encoder is 768. The original image series embedding dimension output from image series encoder is 1280 and is then reduced to 768 by a linear projection head. The level 1 (disease-level) and level 2 (category-level) embeddings are 128 and 64, respectively. The transformer to aggregate the image embeddings consists 2 layers of Multihead Attention. All the learnable temperature τ_k is initialized to 0.07. The training batch size is 32 and the max epoch is 30. we use AdamW optimizer [Loshchilov and Hutter, 2019] with both initial weight decay and initial learning rate equal to 0.0001 during pre-training.

We evaluate several variants of HCE-CLIP, including models with one or two hierarchical levels (H1/H2), with or without energy-based modeling (E/NE), and using either full image series or single images as input (F/S). We name the models like HCE-H2-E-F. Baselines include MedCLIP [Wang *et al.*, 2022] and MedKLIP [Wu *et al.*, 2023], evaluated using the same prompt format and inference procedure for fair comparison.

4.3 Classification Results

Tables 1 and 2 report classification performance using prompt-based, classifier-free inference. MIMIC-CXR Test and CheXpert use the same label set as training (seen labels), while ChestX-ray14 additionally contains unseen disease categories for zero-shot evaluation.

For seen-label classification, HCE-H2-E-F achieves the best overall performance, with the highest AUC and F1 scores across all three datasets. All HCE-CLIP variants substantially outperform MedCLIP and MedKLIP, demonstrating the effectiveness of hierarchical energy-based alignment for prompt-based classification.

For unseen diseases on ChestX-ray14, HCE-H2-E-F consistently outperforms all baselines, achieving the highest AUC, accuracy, and F1. In contrast, MedCLIP and MedKLIP exhibit weaker generalization, highlighting the advantage of hierarchical energy modeling for zero-shot medical classification.

4.4 Linear-Probe Evaluation

Table 3 reports linear-probe evaluation results on MIMIC-CXR and ChestX-ray14. In this setting, the image encoder is frozen and a linear classifier is trained on top of the learned representations, allowing us to assess feature transferability.

HCE-CLIP with two hierarchical levels achieves the best performance on both datasets, with the highest AUC, accuracy, and F1 scores, outperforming both MedKLIP and Med-

	MIMIC-CXR			ChestXray14		
	AUC	ACC	F1	AUC	ACC	F1
HCE-CLIP	**0.8711**	**0.8454**	**0.5871**	**0.8285**	**0.8916**	**0.3767**
MedKLIP	0.8044	0.7950	0.4657	0.7639	0.8588	0.3037
MedCLIP	0.8158	0.8039	0.4840	0.7726	0.8579	0.3075

Table 3: Linear-probe evaluation results

	MIMIC-CXR test	CheXpert	ChestXray14
HCE-H2-E-F	**0.0000**	0.0020	**0.0002**
HCE-H2-NE-F	0.0015	0.0020	0.0020
HCE-H1-E-F	0.0190	0.0580	0.0059
HCE-H1-NE-F	0.1233	0.1000	0.0212
HCE-H2-E-S	**0.0000**	**0.0000**	0.0003
MedKLIP	0.7420	0.6100	0.2937
MedCLIP	0.0801	0.0900	0.3063

Table 4: Hierarchical Contradiction Rate on benchmark datasets.

CLIP by a clear margin. These results indicate that HCE-CLIP learns more discriminative and transferable visual representations than existing medical vision–language pretraining methods, further validating the effectiveness of hierarchical conditional energy modeling.

4.5 Hierarchical Consistency Evaluation

Table 4 summarizes the HCR (Eq. 20) results across MIMIC-CXR Test, CheXpert, and ChestX-ray14. HCE-H2-E-F and HCE-H2-E-S achieve near-zero contradiction rates on all datasets, demonstrating that two-level hierarchical energy modeling enforces strong semantic consistency. In contrast, removing hierarchical energy constraints (HCE-H2-NE-F) leads to higher contradiction rates, confirming that architectural hierarchy alone is insufficient without explicit energy-based supervision.

The single-level variants show substantially worse consistency. HCE-H1-E-F exhibits noticeably higher HCR, while HCE-H1-NE-F suffers from severe inconsistency across all datasets. This indicates that both hierarchical depth and energy-based modeling are necessary to maintain semantic coherence.

Baseline methods MedCLIP and MedKLIP show significantly higher contradiction rates than all HCE-CLIP variants. Overall, these results demonstrate that hierarchical conditional energy modeling is crucial for enforcing clinically meaningful semantic consistency, substantially reducing contradictory predictions compared to both non-energy variants and existing medical vision–language models.

4.6 Ablation Insights

Effect of Hierarchical Energy Modeling. We compare HCE-H2-E-F with its non-energy counterpart HCE-H2-NE-F, as well as HCE-H1-E-F with HCE-H1-NE-F. Across all datasets, models with hierarchical energy supervision achieve higher classification performance and substantially lower HCR. In particular, HCE-H2-E-F consistently outperforms HCE-H2-NE-F in AUC and F1 while reducing contradiction rates to near zero. These results confirm that explicit hierarchical energy modeling is essential for both predictive accuracy and logical coherence.

Effect of Image-Series Encoding. We compare HCE-H2-E-F with the single-image variant HCE-H2-E-S to evaluate the impact of modeling complete imaging studies. While both variants achieve low HCR, image-series encoding improves performance on most datasets, indicating that aggregating multiple views better captures clinical context than relying on individual images alone.

Effect of Hierarchical Depth. We compare two-level models (HCE-H2-E-F) with single-level models (HCE-H1-E-F). The two-level model consistently achieves higher AUC and F1 scores and lower HCR. In contrast, increasing hierarchical depth without energy modeling (HCE-H2-NE-F vs. HCE-H1-NE-F) does not improve classification performance and only marginally improves consistency. This indicates that hierarchical depth alone is insufficient to improve performance without explicit energy-based supervision.

Overall, these ablations validate the design choices of HCE-CLIP and highlight the importance of structured energy modeling for reliable medical vision–language understanding.

5 Discussion and Conclusion

Our results demonstrate that hierarchical conditional energy modeling provides a strong structural guidance for learning hierarchically coherent medical representations. Across MIMIC-CXR, CheXpert, and ChestX-ray14, HCE-CLIP consistently improves both classification performance and hierarchical semantic consistency.

A key finding is that hierarchical energy supervision, rather than hierarchical structure alone, drives these improvements. Ablation studies show that increasing hierarchical depth without energy constraints does not improve performance or consistency, while energy-based supervision substantially reduces hierarchical contradictions. This indicates that structured energy modeling is essential for enforcing meaningful semantic relationships between fine-grained disease findings and higher-level clinical concepts.

HCE-CLIP also achieves near-zero Hierarchical Contradiction Rate, producing logically coherent predictions across semantic levels. Such consistency is particularly important in clinical settings, where contradictory outputs (e.g., detecting a disease while predicting "No Finding") may undermine trust. In addition, while single-image models can also achieve low contradiction rates, HCE-CLIP with image-series encoding improves robustness by aggregating multi-view information into a unified representation.

Despite these strengths, our approach relies on predefined label hierarchies and prompt-based label descriptions, and our evaluation focuses on chest X-ray data. Future work may explore learning hierarchical structures automatically, extending to other imaging modalities, and integrating uncertainty-aware decision thresholds for clinical deployment.

Overall, HCE-CLIP provides a principled and effective framework for hierarchical conditional energy-based medical vision-language pretraining, enabling more accurate, consistent, and interpretable clinical predictions.

References

[Bansal *et al.*, 2024] Hritik Bansal, Daniel Israel, Siyan Zhao, Shufan Li, Tung Nguyen, and Aditya Grover. Medmax: Mixed-modal instruction tuning for training biomedical assistants. *arXiv preprint arXiv:2412.12661*, 2024.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[Chen *et al.*, 2023] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15095–15104, 2023.

[Cui *et al.*, 2021] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Du and Mordatch, 2019] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy-based models. In *Advances in Neural Information Processing Systems*, 2019.

[Ge *et al.*, 2023] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023.

[Geng *et al.*, 2023] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention. In *The Eleventh International Conference on Learning Representations*, 2023.

[Grathwohl *et al.*, 2021] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2021.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Huang *et al.*, 2021a] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[Huang *et al.*, 2021b] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[LeCun *et al.*, 2006] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu-Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

[Li *et al.*, 2020] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[Li *et al.*, 2022a] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[Li *et al.*, 2022b] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and

Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022.

[Li *et al.*, 2023] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[Mao *et al.*, 2018] Chengsheng Mao, Liang Yao, Yiheng Pan, Yuan Luo, and Zexian Zeng. Deep generative classifiers for thoracic disease diagnosis with chest x-ray images. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1209–1214. IEEE, 2018.

[Mao *et al.*, 2022] Chengsheng Mao, Liang Yao, and Yuan Luo. Imagegcn: Multi-relational image graph convolutional networks for disease identification with chest x-rays. *IEEE Transactions on Medical Imaging*, 41(8):1990–2003, 2022.

[Mu *et al.*, 2022] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

[Nguyen *et al.*, 2024] Duy MH Nguyen, Nghiem T Diep, Trung Q Nguyen, Hoang-Bao Le, Tai Nguyen, Tien Nguyen, TrungTin Nguyen, Nhat Ho, Pengtao Xie, Roger Wattenhofer, et al. Logra-med: Long context multi-graph alignment for medical vision-language model. *arXiv preprint arXiv:2410.02615*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Sun *et al.*, 2024] Shenghuan Sun, Alexander Schubert, Gregory M Goldgof, Zhiqing Sun, Thomas Hartvigsen, Atul J Butte, and Ahmed Alaa. Dr-llava: Visual instruction tuning with symbolic clinical grounding. *arXiv preprint arXiv:2405.19567*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2017] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[Wang *et al.*, 2022] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[Wang *et al.*, 2023] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[Wu *et al.*, 2023] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[Xie *et al.*, 2016] Jianwen Xie, Yang Zhu, and Ying Nian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, 2016.

[Xie *et al.*, 2021] Jianwen Xie, Yang Zhu, and Ying Nian Wu. Cooperative learning of energy-based model and latent variable model via mcmc teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[Yang *et al.*, 2022] Minqiang Yang, Yuhong Zhang, Haoning Chen, Wei Wang, Haixu Ni, Xinlong Chen, Zhuoheng Li, and Chengsheng Mao. Ax-unet: A deep learning framework for image segmentation to assist pancreatic tumor diagnosis. *Frontiers in Oncology*, 12:894970, 2022.

[Yao *et al.*, 2022] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

[Yu *et al.*, 2021] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[Zhang *et al.*, 2023] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.