Concept-Guided Noisy Negative Suppression for Zero-Shot Classification and Grounding of Chest X-Ray Findings

Chenyu Lian[1,2], Hong-Yu Zhou[3] (✉), Chun-Ka Wong[4], and Jing Qin[1,2(✉)]

[1] The Center for Smart Health, School of Nursing, the Hong Kong Polytechnic University, Hong Kong, China
chenyu.lian@connect.polyu.hk, harry.qin@polyu.edu.hk
[2] Research Institute for Smart Ageing, the Hong Kong Polytechnic University, Hong Kong, China
[3] School of Biomedical Engineering, Tsinghua University, Beijing, China
hongyu.zhou.ai@gmail.com
[4] Queen Mary Hospital, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
wongeck@hku.hk

Abstract. Vision-language alignment using chest X-rays and radiology reports has emerged as an advanced paradigm for zero-shot classification and grounding of chest X-ray findings. However, standard contrastive learning typically treats radiographs and reports from different patients simply as negative pairs. This assumption introduces ***noisy negatives***, as different patients frequently exhibit similar findings. Such noisy negatives cause semantic ambiguity and degrade performance in zero-shot understanding tasks. To address this challenge, we propose `CoNNS`, a concept-guided noisy-negative suppression framework. To support the negative suppression mechanism, unlike previous methods that use raw reports or templatized texts, we construct a hierarchical concept ontology using large language models. The ontology structures 41 key clinical concepts by explicitly modeling presence, attributes (location and characteristics), and texts (evidential segment and presence statement). Leveraging this ontology, we implement a cross-patient pair relabeling strategy comprising three steps: (1) *Fine-Grained Breakdown* to categorize pairs based on finding presence; (2) *Noisy Negative Filtering* to resolve semantic conflicts by removing false negatives; and (3) *Hard Negative Mining* to identify subtle attribute discrepancies using a lightweight language model. Finally, we propose a Concept-Aware NCE loss to align visual features with text while suppressing the identified noisy negatives. Extensive experiments across multi-granularity zero-shot grounding tasks and five zero-shot classification datasets validate that `CoNNS` outperforms existing state-of-the-art models. The code will be publicly available and is currently at https://github.com/DopamineLcy/conns.

Keywords: Noisy negatives · Zero-shot learning · Chest X-rays

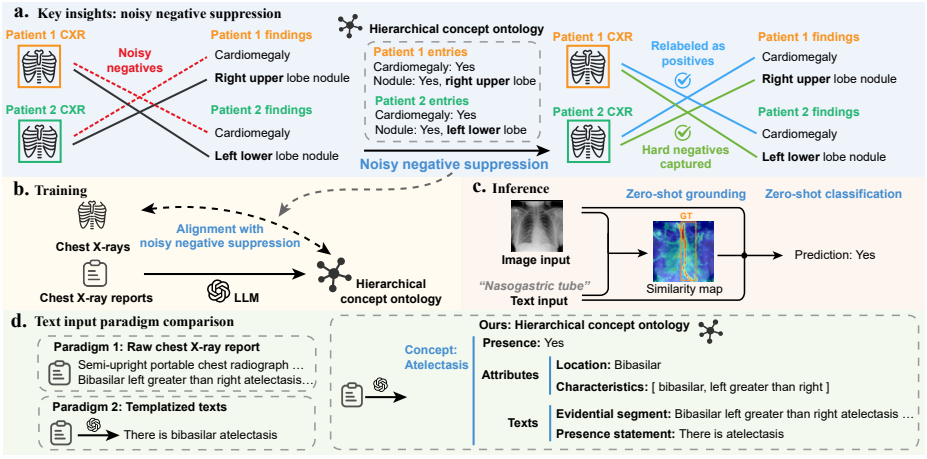

Fig. 1. **a.** Noisy negatives and the proposed suppression strategy. **b.** Training and **c.** inference flows. **d.** Comparison of two mainstream paradigms of text inputs with ours.

1 Introduction

Vision-language alignment using chest X-rays (CXR) and radiology reports significantly advances zero-shot classification and grounding of radiological findings without the need for costly manual annotations (e.g., class labels or bounding boxes) [11,12,16,19,27]. Despite these advancements, standard image-text contrastive learning in radiology suffers from a critical, often overlooked limitation: ***noisy negatives***. The conventional CLIP-style training objective naively treats image-text pairs from different patients as negative samples [19]. However, in radiology datasets, distinct patients frequently exhibit identical findings [22]. As shown in Fig. 1a, Patient 1 and Patient 2 both present "cardiomegaly". Forcing these semantically equivalent pairs apart as negatives introduces contradictory supervision signals, causing semantic ambiguity that hinders the learning of discriminative features. Alongside the noisy negative issue, there exists a dilemma in utilizing text supervision. While previous work has observed that using template text (e.g., "There is `[finding]`") instead of raw reports improves prompt consistency between training and inference stages [11], this reduction in linguistic diversity impairs the understanding of complex, naturalistic clinical texts.

To bridge these gaps, we propose `CoNNS`, a concept-guided framework to suppress noisy negative pairs while simultaneously preserving clinical report diversity and prompt consistency. Fig. 1d compares the proposed `CoNNS` with two mainstream text utilization paradigms. Paradigm 1 utilizes raw report texts [2,28]; while this preserves natural diversity, it suffers from poor prompt consistency between training (long, dense reports) and inference (short, atomic queries). Paradigm 2 employs fixed-style templates [11,16], which ensures prompt consistency but sacrifices the linguistic diversity required to handle complex cases. Instead of relying on raw text or rigid templates, we construct a ***hierarchical***

concept ontology derived from radiology reports. The ontology decomposes reports into structured knowledge, explicitly modeling presence, attributes (location and characteristics), and texts (evidential segments and presence statements). The extracted evidential segments preserve linguistic diversity, while the standardized presence statements ensure prompt consistency.

Building on this concept ontology, we introduce a ***cross-patient pair relabeling*** strategy incorporating three steps: (1) *Fine-Grained Breakdown* to categorize pairs based on finding presence; (2) *Noisy Negative Filtering* to resolve semantic conflicts by removing false negatives; and (3) *Hard Negative Mining* to identify subtle attribute discrepancies using a lightweight language model. Finally, to achieve vision-language alignment using these relabeled pairs, we propose a ***Concept-Aware NCE Loss***, which aligns visual features with all semantically consistent reports in the batch while masking out noisy negatives. **Main contributions. 1.** We address the critical challenge of *noisy negatives* in radiology vision-language alignment, proposing a new learning paradigm that shifts from patient-based to concept-based negative pair identification. **2.** The proposed *hierarchical concept ontology* simultaneously preserves clinical text diversity and ensures prompt consistency. **3.** We propose a *Concept-Aware NCE loss* tailored for the concept-based framework, suppressing noisy negatives during vision-language alignment. **4.** Extensive experiments across multi-granularity (word, phrase, and sentence levels) grounding tasks and five zero-shot classification datasets validate that `CoNNS` outperforms previous state-of-the-art methods.

2 Methodology

Overview. As illustrated in Fig. 2, our framework comprises three primary stages: 1. ***construction of hierarchical concept ontology*** for 41 key clinical concepts (Sec. 2.1); 2. ***cross-patient pair relabeling***, which generates a relation matrix for image-text pairs based on the hierarchical concept ontology (Sec. 2.2); and 3. ***vision-language alignment with noisy negative suppression***, using the proposed Concept-Aware NCE (CA-NCE) loss (Sec. 2.3).

2.1 Construction of Hierarchical Concept Ontology

To facilitate the proposed noisy negative suppression, we construct a hierarchical concept ontology derived from radiology reports in MIMIC-CXR [10]. Utilizing locally deployed Llama-3.3-70B-Instruct [8], we process the reports to extract information for 41 key clinical concepts, which are selected based on previous datasets and discussions with radiologists. As illustrated in Fig. 2 (Stage 1), the ontology decomposes each finding into ***three layers***: (1) 41 key clinical concepts of chest X-ray findings, such as "atelectasis". (2) Components including `Presence`, `Attributes`, and `Texts`, where `Presence` guides the fine-grained breakdown in Stage 2.1 and noisy negative filtering in Stage 2.2; `Attributes` are used for hard negative mining in Stage 2.3; while `Texts` serve as text inputs in Stage 3. (3) Granular details, where `Attributes` include `Location` and

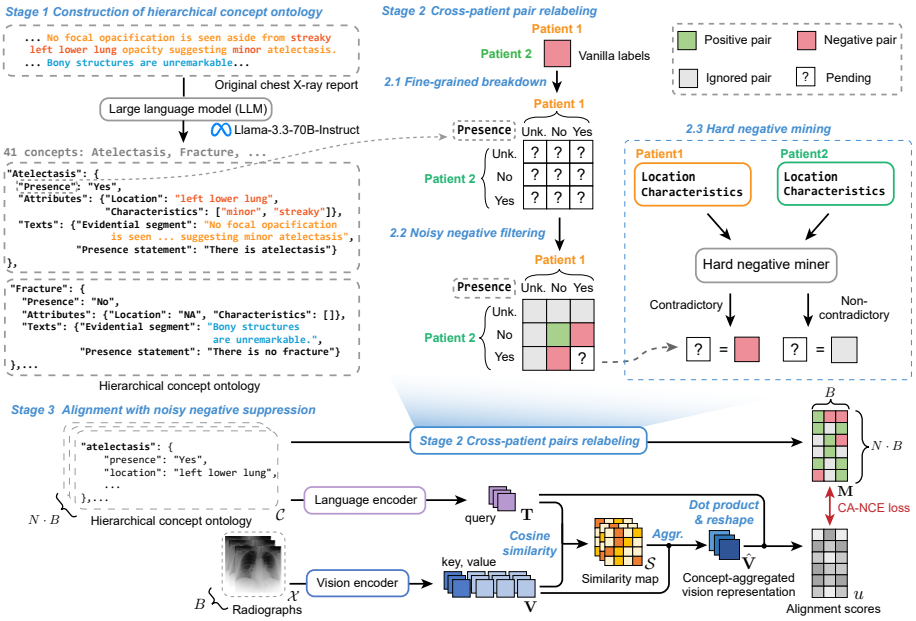

Fig. 2. The proposed `CoNNS` consists of three stages. (1) We construct a hierarchical concept ontology via LLM. (2) Cross-patient pairs are relabeled into a relation matrix based on the ontology (Unk. = Unknown). (3) We perform vision-language alignment with noisy negative suppression using the proposed Concept-Aware NCE loss.

`Characteristics`, while `Texts` comprise `Evidential segment` and `Presence statement` (each sampled with a 50% probability during training).

2.2 Cross-Patient Pair Relabeling

Following standard practice, we regard pairs consisting of images and text from the same patient as positive samples. Regarding image-text pairs from different patients, standard CLIP-style training assumes them to be negative samples. However, this assumption often fails in radiology, where descriptions like "No pneumothorax" are frequently shared by different patients, resulting in ***noisy negatives***. To mitigate this, we propose a relabeling strategy to construct a ***relation matrix*** $\mathbf{M} \in \{1, 0, \varnothing\}^{(N \cdot B) \times B}$, where N and B refer to the number of sampled texts per image (default 8) and the batch size (default 192), respectively. There are $N \cdot B$ texts in a batch. Entries in $\mathbf{M}$ represent positives (1), negatives (0), and ambiguous pairs ($\varnothing$). The construction of $\mathbf{M}$ proceeds in three steps:
1. Fine-grained breakdown. We first construct a presence-based 3×3 grid by pairing three possible statuses: "Yes" (present), "No" (absent), or "Unk." (Unknown) (Fig. 2, Stage 2.1), which provides the foundation for subsequent logic.
2. Noisy negative filtering. Based on the breakdown, we categorize the pairs into four initial types and populate the relation matrix $\mathbf{M}$ (Fig. 2, Stage 2.2):

• Positive pairs ($\mathbf{M}_{i,j} = 1$): "No-No" pairs, a consistent absence of the concept.
• Negative pairs ($\mathbf{M}_{i,j} = 0$): "Yes-No" pairs, where the presence is contradictory.
• Ignored pairs ($\mathbf{M}_{i,j} = \varnothing$): Pairs containing an "Unknown" status are considered semantically ambiguous and are therefore ignored when computing loss.
• Pending pairs: "Yes-Yes" pairs, where both patients share the finding but may differ in characteristics (e.g., small vs. large) and location (e.g., left vs. right). These are deferred to the next stage to avoid introducing noise.

3. Hard negative mining. To facilitate fine-grained discrimination, we tackle the pending ambiguous pairs using `Location` and `Characteristics` (Fig. 2, Stage 2.3). We compare the attributes of Patient 1 against Patient 2 using nli-deberta-v3-small [20]. If the attributes are explicitly contradictory (e.g., Left vs. Right), the pair is relabeled as a **hard negative** ($\mathbf{M}_{i,j} = 0$). Otherwise, if attributes are similar or insufficient to distinguish, we treat them as potential noisy negatives and mask them out ($\mathbf{M}_{i,j} = \varnothing$) to avoid semantic ambiguity.

2.3 Vision-Language Alignment with Noisy Negative Suppression

Building upon the negative pair relabeling in Sec. 2.2, we introduce the **Concept-Aware NCE (CA-NCE)** loss. Guided by the relation matrix $\mathbf{M}$, this loss function aligns concept-aggregated visual representations with corresponding texts.

Counterfactual prompting. To enhance the diversity of negative image-text pairs, we introduce counterfactual texts for each target concept by sampling descriptions from patients with the opposite `Presence` status. In these cases, fine-grained attributes (i.e., `location` and `characteristics`) are set to null to prevent the introduction of spurious correlations during hard negative mining.

Vision and text encoding. Let $\mathcal{X} = \{x_j\}_{j=1}^{B}$ be a batch of chest X-rays and $\mathcal{C} = \{c_i\}_{i=1}^{N \cdot B}$ be the concept-centric input texts. We employ a dual-encoder architecture with a frozen Rad-DINO-Maira-2 [1,18] followed by two trainable ViT layers [7] as the vision encoder (input size 518×518), and BiomedVLP-CXR-BERT [3] as the text encoder (length of 128 tokens). Inputs are processed to obtain the visual and textual representations: $\mathbf{V} = f_{vis}(\mathcal{X})$ and $\mathbf{T} = f_{txt}(\mathcal{C})$. Here, $\mathbf{V} \in \mathbb{R}^{B \times L \times D}$ denotes the batch of image patch embeddings, where $\mathbf{v}_{j,l} \in \mathbb{R}^D$ represents the raw feature vector of the l-th patch in the j-th image ($j \in \{1, \ldots, B\}, l \in \{1, \ldots, L\}$). Correspondingly, $\mathbf{T} \in \mathbb{R}^{(N \cdot B) \times D}$ denotes the text embeddings, where $\mathbf{t}_i \in \mathbb{R}^D$ corresponds to the i-th text ($i \in \{1, \ldots, (N \cdot B)\}$).

Similarity map and concept-aggregated vision representation. Let $\bar{\mathbf{t}}_i$ and $\bar{\mathbf{v}}_{j,l}$ denote the ℓ_2-normalized text and image patch embeddings, respectively. To achieve concept-aware fine-grained alignment, we first compute the similarity map $\mathcal{S}_{i,j} \in \mathbb{R}^L$, which consists of patch-wise similarity scores $s_{i,j,l}$, defined as the scaled dot-product: $s_{i,j,l} = (\bar{\mathbf{t}}_i^\top \bar{\mathbf{v}}_{j,l})/\tau_{attn}$. Subsequently, we normalize these scores via softmax to generate attention weights $w_{i,j,l}$, which are used to aggregate the visual features into a concept-specific representation $\hat{\mathbf{v}}_{i,j}$ ($\hat{\mathbf{V}}$ in Fig. 2 refers to representations $\{\hat{\mathbf{v}}_{i,j}\}$ in a batch):

$$w_{i,j,l} = \frac{\exp(s_{i,j,l})}{\sum_{k=1}^{L} \exp(s_{i,j,k})}, \quad \mathbf{z}_{i,j} = \sum_{l=1}^{L} w_{i,j,l} \bar{\mathbf{v}}_{j,l}, \quad \hat{\mathbf{v}}_{i,j} = \frac{\mathbf{z}_{i,j}}{\|\mathbf{z}_{i,j}\|_2}. \tag{1}$$

The final alignment score $u_{i,j}$ is computed as $u_{i,j} = \bar{\mathbf{t}}_i^\top \hat{\mathbf{v}}_{i,j}$.

Notably, for zero-shot grounding in inference, $\mathcal{S}_{i,j}$ is reshaped to the spatial grid size $H \times W$ (where $H \cdot W = L$) and interpolated to the original image resolution to serve as the localization heatmap.

Concept-aware NCE loss. To suppress noisy negatives identified in the relabeling stage and accommodate arbitrary numbers of positive pairs, we propose Concept-Aware NCE loss based on InfoNCE loss [15], which incorporates a scaling temperature τ_{loss}. We first derive binary masks from the relation matrix $\mathbf{M}$ constructed in Sec. 2.2. The **positive mask** $\mathbf{M}_{pos}$ indicates positive pairs, and the **valid mask** $\mathbf{M}_{valid}$ indicates pairs participating in the loss:

$$\mathbf{M}_{pos}(i,j) = \mathbb{I}(\mathbf{M}_{i,j} = 1), \quad \mathbf{M}_{valid}(i,j) = \mathbb{I}(\mathbf{M}_{i,j} \neq \varnothing). \tag{2}$$

The Text-to-Image ($\mathcal{L}_{T2I}$) and Image-to-Text ($\mathcal{L}_{I2T}$) losses are defined as:

$$\mathcal{L}_{T2I}(i) = -\log\left(\frac{\sum_{j=1}^{B} \mathbf{M}_{pos}(i,j) \cdot \exp(u_{i,j}/\tau_{loss})}{\sum_{k=1}^{B} \mathbf{M}_{valid}(i,k) \cdot \exp(u_{i,k}/\tau_{loss}) + \epsilon}\right), \tag{3a}$$

$$\mathcal{L}_{I2T}(j) = -\log\left(\frac{\sum_{k=1}^{N\cdot B} \mathbf{M}_{pos}(k,j) \cdot \exp(u_{k,j}/\tau_{loss})}{\sum_{m=1}^{N\cdot B} \mathbf{M}_{valid}(m,j) \cdot \exp(u_{m,j}/\tau_{loss}) + \epsilon}\right). \tag{3b}$$

The total objective is the average over entries containing positive targets:

$$\mathcal{L} = \frac{\sum_i \mathbb{I}(p_i > 0)\mathcal{L}_{T2I}(i)}{\sum_i \mathbb{I}(p_i > 0)} + \frac{\sum_j \mathbb{I}(q_j > 0)\mathcal{L}_{I2T}(j)}{\sum_j \mathbb{I}(q_j > 0)}. \tag{4}$$

where $p_i = \sum_j \mathbf{M}_{pos}(i,j)$ and $q_j = \sum_k \mathbf{M}_{pos}(k,j)$.

3 Experiments

3.1 Dataset and Evaluation Metrics

Training dataset. We employ MIMIC-CXR [10] for vision-language alignment training. This dataset comprises 227,835 radiographic studies from 65,379 patients. We follow the official split for training and validation. Each study consists of a radiology report and one or more chest X-ray images, totaling 377,100 images. These reports serve as the source for constructing the hierarchical concept ontology, which is described in Sec. 2.1.

Evaluation datasets. We conduct a comprehensive evaluation using three zero-shot grounding datasets and five zero-shot classification datasets. For ***zero-shot grounding***, we systematically evaluate the performance of ***word-level, phrase-level, and sentence-level*** grounding tasks on ChestXDet10 [13], MS-CXR [3], and PadChest-GR [4], respectively. To measure grounding performance, we utilize the Pointing Game score [24], which calculates the hit rate where the maximum value falls within the ground-truth bounding box. For **zero-shot classification**, we employ ChestXDet10 [13], Open-I [6], ChestXray14 [21], CheXpert [9], and PadChest-GR [4]. The Area Under the Receiver Operating Characteristic curve (AUROC) is adopted as the classification metric. Specifically, the

Table 1. Comparison with SOTAs. Pointing Game scores (%) for zero-shot grounding and AUROCs (%) for zero-shot classification are reported as mean ± std over 3 runs. "-": not applicable. The best results are in **bold** and the second-best ones are underlined.

Method	Venue	Zero-shot grounding			Zero-shot classification				
		ChestXDet10	MS-CXR	PadChest-GR	ChestXDet10	Open-I	ChestXray14	CheXpert	PadChest-GR
CoNNS	Ours	**65.6**±0.2	**87.2**±0.3	**86.6**±0.2	**82.5**±0.2	**87.1**±0.1	**81.9**±0.1	92.0±0.2	**88.1**±0.2
RadZero [16]	NeurIPS'25	62.2	84.4	85.3	78.7	84.7	80.4	90.0	84.7
CARZero [11]	CVPR'24	54.3	79.4	77.1	79.6	83.8	81.1	**92.3**	86.1
BiomedCLIP [25]	NEJM AI'24	-	-	-	63.0	57.7	63.9	67.7	64.7
KAD [26]	Nat. Commun.'23	39.1	19.2	29.2	73.5	80.7	78.9	90.5	67.3
MedKLIP [23]	ICCV'23	48.1	40.7	34.8	71.3	75.9	72.6	87.9	67.9
BioViL-T [2]	CVPR'23	35.1	71.9	22.8	70.8	70.2	72.9	78.9	52.9

Table 2. Pointing Game scores (%) per category for zero-shot grounding on ChestXDet10. The best results are in **bold** and the second-best ones are underlined.

Method	Venue	Mean	ATE	CALC	CONS	EFF	EMPH	FIB	FX	MASS	NOD	PTX
CoNNS	Ours	**65.6**	**83.3**	**44.7**	74.4	80.2	76.9	**65.9**	**40.8**	63.3	**63.6**	**62.9**
RadZero [16]	NeurIPS'25	62.2	64.6	36.8	82.4	**85.7**	**87.2**	58.5	25.0	**76.7**	50.6	54.3
CARZero [11]	CVPR'24	54.3	60.4	18.4	82.4	78.2	84.6	56.1	18.4	70.0	28.6	45.7
KAD [26]	Nat. Commun.'23	39.1	64.6	13.2	69.9	61.8	64.4	24.4	19.9	26.7	31.6	14.3
MedKLIP [23]	ICCV'23	48.1	62.5	13.2	**83.7**	67.5	73.4	30.5	22.4	73.3	31.2	22.9
BioViL-T [2]	CVPR'23	35.1	43.8	0.0	63.0	50.4	84.6	39.0	2.6	50.0	0.0	17.1

test set of **ChestXDet10** includes 542 images with manually annotated bounding boxes for 10 diseases. **MS-CXR** contains 1,153 phrase-bounding box pairs. For a fair comparison, evaluation is conducted on the test set released by [5], comprising 167 images. **PadChest-GR** contains 915 images in the test set, highlighting box-level labels with sentence-level descriptions of 24 findings. **Open-I** comprises 7,470 chest X-ray images with annotations (refer to CARZero [11] codebase) for 18 diseases. The official test set of **ChestXray14** contains 22,433 images annotated with 14 disease labels. **CheXpert** includes 500 patients in the test set; our evaluation focuses on the 5 officially recommended diseases.

3.2 Results and Analyses

Comparison with the state-of-the-art. We compare the proposed CoNNS against state-of-the-art methods, spanning zero-shot understanding [11,16] and vision-language pretraining [2,23,25,26]. As presented in Table 1, CoNNS achieves superior performance across both zero-shot grounding and zero-shot classification tasks. Specifically, for zero-shot grounding, CoNNS significantly outperforms the runner-up RadZero [16] across all multi-granularity datasets ($p < 0.01$), achieving absolute gains of 3.4%, 2.8%, and 1.3% on ChestXDet10, MS-CXR, and PadChest-GR, respectively. A fine-grained analysis on ChestXDet10 (Table 2) further reveals that CoNNS achieves the highest Pointing Game scores in 6 out of 10 categories. Notably, in challenging categories where prior methods struggle, such as CALC and FX, we observe substantial relative improvements of 21.5% and 63.2%, respectively. In Table 2, we adopt the following abbreviations for the findings: Atelectasis (ATE), Calcification (CALC), Consolidation (CONS), Effusion (EFF), Emphysema (EMPH), Fibrosis (FIB), Fracture

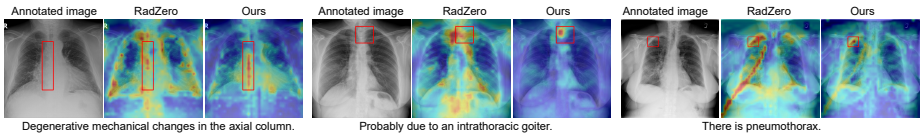

Fig. 3. Qualitative comparison of similarity maps between ours and the leading competing method [16]. Red bounding boxes indicate ground-truth annotations.

Table 3. Ablation study of main components in `CoNNS`. Pointing Game scores and AUROCs (%) are reported for zero-shot grounding and classification, respectively.

Setting	Noisy negative filtering	Hard negative mining	Counterfactual prompting	Zero-shot grounding			Zero-shot classification		
				ChestXDet10	MS-CXR	PadChest-GR	ChestXDet10	Open-I	ChestXray14
(a)	✓			61.1 (−4.5)	82.0 (−5.2)	88.2 (+1.6)	81.2 (−1.3)	84.0 (−3.1)	81.6 (−0.3)
(b)			✓	58.5 (−7.1)	85.6 (−1.6)	86.2 (−0.4)	81.7 (−0.8)	86.9 (−0.2)	82.1 (+0.2)
(c)	✓	✓		62.6 (−3.0)	87.4 (+0.2)	87.9 (+1.3)	81.4 (−1.1)	87.0 (−0.1)	81.9 (−0.0)
(d)	✓		✓	61.2 (−4.4)	82.6 (−4.6)	87.0 (+0.4)	82.0 (−0.5)	84.7 (−2.4)	81.2 (−0.7)
Ours	✓	✓	✓	65.6	87.2	86.6	82.5	87.1	81.9

(FX), Mass (MASS), Nodule (NOD), and Pneumothorax (PTX). In zero-shot classification, `CoNNS` also exhibits higher AUROCs on four out of five datasets ($p < 0.01$) while achieving performance comparable to CARZero [11] on CheXpert ($p = 0.12$). Partial results in Table 1 and 2 are cited from [11] and [16].

Qualitative analysis. To provide an intuitive comparison of grounding capabilities, we visualize the similarity maps of the proposed `CoNNS` alongside the leading state-of-the-art method, RadZero [16]. As illustrated in Fig. 3, `CoNNS` generates a more concentrated focus that aligns precisely with the ground-truth bounding boxes. These qualitative results validate that the proposed noisy negative suppression strategy effectively mitigates semantic ambiguity, enabling the precise localization of complex findings.

Ablation study. We conduct an ablation study to validate the contributions of the proposed noisy negative suppression (comprising noisy negative filtering and hard negative mining) alongside the counterfactual prompting strategy. As shown in Table 3, removing the entire noisy negative suppression leads to the sharpest performance drop (Setting b). The consistent trend of performance degradation across all settings indicates the necessity of these core components.

3.3 Implementation Details

Our code is implemented using PyTorch 2.7.0 [17]. Experiments were conducted using 4 NVIDIA RTX PRO 5000 Blackwell (48 GB) GPUs. We adopt AdamW [14] as the optimizer, with a weight decay of 0.05, β_1 of 0.9, and β_2 of 0.95. A warm-up strategy is employed, where the learning rate increases linearly to 1e-4 and then decreases using a cosine decay schedule. The model is trained for 10 epochs, during which the checkpoint yielding the lowest validation loss is selected. We evaluate comparative methods using their official code and weights.

4 Conclusion

In this paper, we propose CoNNS, a concept-guided framework to suppress noisy negatives in chest X-ray vision-language alignment, enhancing zero-shot classification and grounding of findings. The proposed hierarchical concept ontology and Concept-Aware NCE loss mitigate semantic ambiguity. Comprehensive experiments validate the advancements in zero-shot tasks. This concept-centric paradigm offers a solution for mitigating noisy negatives in medical data, which can be generalized to broader modalities in future work.

Acknowledgments. This work was supported in part by a Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under Shenzhen Municipal Science and Technology Innovation Commission (project no. SGDX20230821092359002) and a project under Innovation and Technology Support Programme of Hong Kong Innovation and Technology Commission (project no. ITS/202/23). C. Lian is supported in part by the Research Institute for Smart Ageing of the Hong Kong Polytechnic University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
2. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
4. de Castro, D.C., Bustos, A., Bannur, S., Hyland, S.L., Bouzid, K., Wetscherek, M.T., Sánchez-Valverde, M.D., Jaques-Pérez, L., Pérez-Rodríguez, L., Takeda, K., et al.: Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. NEJM AI **2**(7), AIdbp2401120 (2025)
5. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G.S.K., Cheng, L.T.E., Thng, C.H., Xu, X., Liu, Y., et al.: Medical phrase grounding with region-phrase context contrastive alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 371–381. Springer (2023)
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2015)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
10. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
11. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11137–11146 (2024)
12. Lian, C., Zhou, H.Y., Liang, D., Qin, J., Wang, L.: Efficient medical vision-language alignment through adapting masked vision models. IEEE Transactions on Medical Imaging **44**(11), 4499–4510 (2025)
13. Liu, J., Lian, J., Yu, Y.: Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. arXiv preprint arXiv:2006.10550 (2020)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
15. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
16. Park, J., Yoon, B., Kim, S., Choi, K.: Radzero: Similarity-based cross-attention for explainable vision-language alignment in chest x-ray with zero-shot multi-task capability. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Perez-Garcia, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence **7**(1), 119–130 (2025)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019)
21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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. pp. 2097–2106 (2017)
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)

23. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21372–21383 (2023)
24. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)
25. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024)
26. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
27. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
28. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: The Eleventh International Conference on Learning Representations (2023)