# Supplementary material: Relaxing Binary Constraints in Contrastive Vision-Language Medical Representation Learning

Xiaoyang Wei, Camille Kurtz, Florence Cloppet
Université Paris Cité, LIPADE, F-75006, Paris, France
xiaoyang.wei@etu.u-paris.fr

## Abstract

*This document provides supplementary materials to the manuscript entitled "Relaxing Binary Constraints in Vision-Language Medical Contrastive Learning", submitted to the WACV 2025 scientific event. The two sections are related to the ablation study (Sec. 5).*

## 1. Knowledge extraction

To evaluate the effectiveness of each part in the knowledge extraction module, we conduct several extensive ablative experiments (illustrated in Fig. 1):

- With the same semantic matching loss, in experiment one (ID 1), instead of building image-text similarity target using the extracted knowledge, we randomly generate noise vectors and inject them into the loss function (referred to as NB-CLIP (noise));
- In experiment two (ID 2), we keep image nodes and the extracted CUI nodes in our KG paired with *has_attribute* edge between them but remove the hierarchical *is_a* edges derived from external UMLS ontology to model the subtle medical context for CUI nodes. We name this approach as NB-CLIP (UMLS without ontology);
- Finally in experiment three (ID 3) we adopt our proposed approach NB-CLIP (UMLS) which combines knowledge hidden in the dataset with external domain knowledge to model image-text relationships as a comparison.

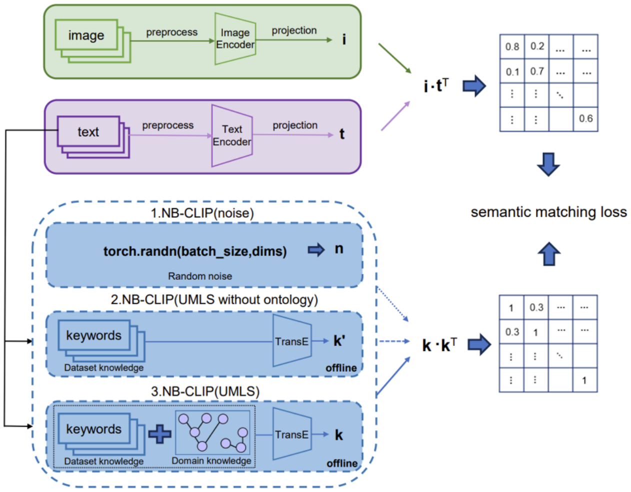

Figure 1. Ablation study of knowledge extraction strategies. We compare our method with noise-guided NB-CLIP and keywords-guided NB-CLIP (i.e. UMLS without ontology) in experiments 1 & 2 respectively.

According to the results provided in Table 1, we can draw the following observations: First (ID 1), randomly relaxing binary CLIP using semantic matching loss paired with noise deteriorates the performance of CLIP (fine-tuned on ROCO). Second (ID 2), without the hierarchical relationships for CUI nodes, we fail to position related CUIs in a reasonable context. In this case, modelling the representation of image nodes solely based on extracted CUIs does not show better supervision compared with CLIP. Third (ID 3), with incorporation of semantic knowledge from external knowledge base, the model learns finer grain correlation between image-caption pairs, and achieves the best results which confirms the necessity of knowledge fusion for dataset knowledge (DK) and external domain knowledge (EK). It also showcases the potential of relaxing binary constraints in CLIP using semantic matching loss paired with various knowledge.

## 2. Choice of the loss function

As discussed in the main document, the objective of NB-CLIP is to minimize the distance between the soft targets and the predicted image-text logits. Here we compare the effectiveness of several possible losses as shown in Table 2. NB-CLIP supervised by MSE loss and Sigmoid loss does

|  |  |  | ROCO |  |  | Custom retrieval dataset |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  |  | CUI@K |  |  | P@K |  |  |  |  |  |  |  |  |
| ID | Knowledge |  | @5 | @10 | @50 | @5 | @10 | @30 | @5 | @10 | @30 | @5 | @10 | @30 |
|  | DK | EK |  |  |  | Modality |  |  | Organ |  |  | Modality/Organ |  |  |
| 1 |  |  | 38.75 | 39.80 | 43.61 | 89.60 | 87.80 | 81.20 | 64.17 | 60.00 | 49.03 | 92.29 | 90.00 | 76.03 |
| 2 | ✓ |  | 39.72 | 40.88 | 44.79 | 90.65 | 88.47 | 84.73 | 67.98 | 66.52 | 54.76 | 92.54 | 90.37 | 75.08 |
| 3 | ✓ | ✓ | **41.82** | **43.25** | **48.61** | **92.80** | **92.00** | **89.33** | **76.67** | **72.50** | **57.67** | **95.24** | **95.25** | **81.92** |
| CLIP |  |  | 39.85 | 40.84 | 44.86 | 91.20 | 88.60 | 84.60 | 67.92 | 66.46 | 54.86 | 92.38 | 90.48 | 75.24 |

Table 1. Ablation study of knowledge extraction strategies under the image retrieval downstream task (evaluation on the ROCO dataset). DK refers to the keywords extracted from the dataset, EK refers to the ontology of the external knowledge base.

|  |  | ROCO |  |  | Custom retrieval dataset |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | CUI@K |  |  | P@K |  |  |  |  |  |  |  |  |
| Methods | Loss | @5 | @10 | @50 | @5 | @10 | @30 | @5 | @10 | @30 | @5 | @10 | @30 |
|  |  |  |  |  | Modality |  |  | Organ |  |  | Modakity/Organ |  |  |
| CLIP | InfoNCE | 39.85 | 40.84 | 44.86 | 91.20 | 88.60 | 84.60 | 67.92 | 66.46 | 54.86 | 92.38 | 90.48 | 75.24 |
| NB-CLIP | MSE | 39.76 | 40.97 | 45.74 | 92.40 | 90.80 | 88.27 | 58.75 | 55.83 | 46.25 | **96.19** | 93.33 | 84.44 |
| NB-CLIP | Sigmoid [1] | 39.84 | 41.06 | 45.73 | 90.80 | 90.80 | **89.93** | 59.92 | 55.21 | 44.17 | 95.24 | 94.76 | **86.67** |
| NB-CLIP | SM(ours) | **41.82** | **43.25** | **48.61** | **92.80** | **92.00** | 89.33 | **76.67** | **72.50** | **57.67** | 95.24 | **95.25** | 81.92 |

Table 2. Ablation study of various loss function for soft targets-based NB-CLIP under the image retrieval downstream task (evaluation on the ROCO dataset). MSE refers to Mean Squared Error loss, SM refers to the Semantic matching loss.

not show additional gain when compared with CLIP with binary NB-CLIP loss. However, according to the first row and the last row, we can observe that replacing the InfoNCE loss with a non-binary soft target-based semantic matching loss can lead to notable improvement on all metrics.

## References

[1] Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *CVPR, Procs.*, pages 14172–14182, 2024. 2