Visual Medical Entity Linking with VELCRO

Kathryn Carbone KCARBONE@UWATERLOO.CA
Liam Hebert LIAM.HEBERT@UWATERLOO.CA
Robin Cohen RCOHEN@UWATERLOO.CA
Lukasz Golab LGOLAB@UWATERLOO.CA
University of Waterloo, Canada

Abstract

We study a visual entity linking (VEL) problem in which a user selects a region of interest (RoI) in an image (e.g., a brain tumour) and queries a textual knowledge base (KB) for information about the RoI. To solve this problem using cross-modal embeddings such as CLIP, we can encode the KB entries, then either encode the whole image or just the cropped RoI, and run a similarity search between the query and the KB embeddings. However, using the entire image as the query may retrieve KB entries related to other aspects of the image beyond the RoI, whereas using the RoI alone as the query ignores context, which is critical for recognizing and linking complex entities in medical images. To address these shortcomings, we propose VELCRO – visual entity linking with contrastive RoI alignment – which adapts an image segmentation model to VEL by aligning the contextual embeddings produced by its decoder with the KB using contrastive learning. This strategy preserves the information contained in the surrounding image while focusing KB alignment on the RoI. Experiments on medical VEL show that VELCRO achieves 95.3% linking accuracy compared to 83.9% or lower for baselines.

Keywords: Visual entity linking, image segmentation, contrastive learning

Data and Code Availability The imaging data, curated by Ma et al. (2024b), is publicly available at https://medsam-datasetlist.github.io/. The code and textual data are available at https://github.com/carbonkat/VELCRO.

Institutional Review Board (IRB) No IRB approval needed: secondary data analysis.

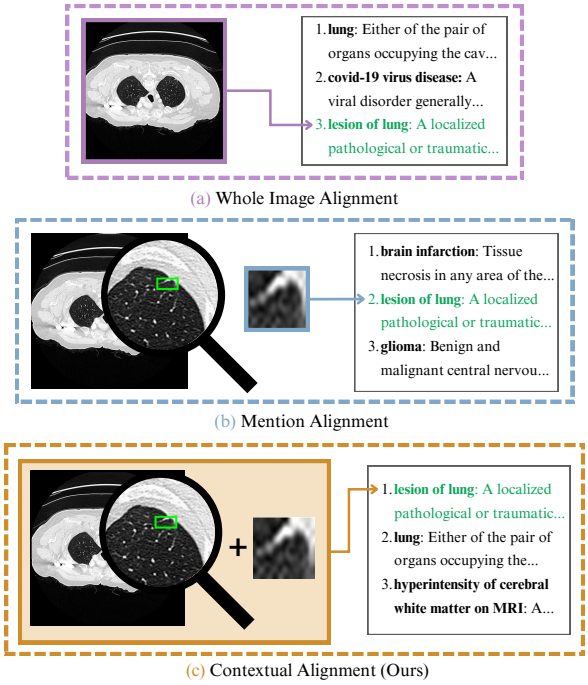

Figure 1: **Visual Entity Linking (VEL) approaches** using an CT of a lung lesion: matching the full image to the KB, which returns "lung" and misses the lesion (a), matching the highlighted RoI to the KB, which ignores context and returns an entity from the wrong organ ("brain infarction") (b), and, as used in our solution, combining the RoI and the surrounding image context for linking (c).

1. Introduction

Entity linking (EL) is the task of identifying entities mentioned in a document and linking them to the corresponding records in a knowledge base (KB). Visual entity linking (VEL) links *visual* mentions of enti-

© 2025 K. Carbone, L. Hebert, R. Cohen & L. Golab.

ties to their descriptions in a KB (Tilak et al., 2017). VEL is useful in fields such as medicine that generate high volumes of imaging data, where image analysis is both time-consuming and labour-intensive for domain experts. It can support physicians and medical students by identifying artifacts of interest in medical images. This information can then be used for tasks such as similar-case retrieval, allowing clinicians to learn and make decisions by observing the outcomes of past cases with similar visual presentations of an artifact of interest (K et al., 2025).

We study a VEL problem in which the user provides a query image and a prompt in the form of a bounding box around a region of interest (RoI) within that image. Suppose a doctor inspects a lung MRI and identifies a suspicious mass. The doctor can utilize VEL by drawing a bounding box around it and requesting relevant information from a medical KB. A popular VEL solution is to use a cross-modal embedding model such as CLIP (Radford et al., 2021). CLIP aims to produce embeddings that are aligned across modalities; e.g., the embedding of the sentence "a photo of a cat" would be similar to the embedding of an image of a cat. A VEL system can encode the KB entries and the query image using CLIP, and perform a vector similarity search to retrieve the most similar KB entries. Using the whole image as the query corresponds to the image-level design in Fig. 1 (a). It is also possible to use the bounding box containing the RoI as the query, depicted as the mention-level design in Fig. 1 (b). However, the former simply returns "lung" as the linked entity since the lung overshadows the small RoI, while the latter links to an incorrect body part due to the missing context and outputs "brain infarction". We hypothesize that a better approach would be to use information from both the RoI and the surrounding image for entity linking, as illustrated in Fig. 1 (c), leading to the correctly linked KB entity of "lesion of the lung".

We propose a novel implementation of the design in Fig. 1 (c). The idea is to compute *contextual* visual embeddings that capture both the RoI and the relevant surrounding information. These embeddings can then be aligned with the textual KB embeddings to enable entity linking via vector similarity search.

Our solution, **V**isual **E**ntity **L**inking with **C**ontrastive **R**egion-**o**f-Interest Alignment (VELCRO), is based on the observation that image segmentation models produce contextual RoI representations when delineating objects in images. Therefore, aligning these representations with the KB can enable VEL. Specifically, we add cross-modal alignment to the popular Segment Anything Model (SAM) (Kirillov et al., 2023). To represent images and RoIs within them, VELCRO repurposes SAM's mask embeddings, which accumulate information about the broader image and a specific region of interest for downstream segmentation mask generation (precise location). For alignment, our contrastive learning uses {image + RoI, KB entity} pairs for training, with positive examples containing the correct KB entity corresponding to the RoI and negative examples containing an incorrect KB entity. While VELCRO is built on existing segmentation and cross-modal representation model architectures, we combine these components in a new way via a novel loss function that integrates segmentation performance with contrastive alignment, training the end-to-end model to perform both tasks.

On a medical VEL use case, VELCRO achieves F1 linking accuracy (the harmonic mean of precision of recall) of 95.2%, surpassing various baselines (corresponding to Fig. 1 (a) and (b), extensions of existing solutions to related image understanding problems, as well as zero-shot and few-shot Large Language Models (LLMs)), achieving 83.9% at best and in some cases much lower. This demonstrates the practical value of adapting segmentation-based approaches to semantic tasks such as VEL.

2. VELCRO Details

2.1. Problem Overview

Let I be an image containing a visual mention of interest V. The user also provides a bounding box prompt P indicating the approximate location of V within I. Using the bounding box P and image I, the VEL system must (a) identify the exact V and (b) link it to its associated text-based entity entry T in KB, our desired knowledge base. Subtask (b) is achieved by comparing the representation of V, termed E^V, against all T_i representations (E_i^T) in KB. The T_i corresponding to the most similar E_i^T is then selected as the linked entity.

To train a VEL system, we need a dataset $D = \{(I_i, P_i, M_i^V, T_i) \cdots (I_n, P_n, M_n^V, T_n)\}$, where M_i^V is a segmentation mask identifying the exact location of V_i, broadly indicated by P_i, within I_i. Then, given (I_i, P_i), the model must learn to produce both $\hat{M}_i^V \approx M_i^V$ that correctly identify V_i **and** representations E_i^V and E_i^T from V_i and T_i that correctly align with

each other. During inference, only (I, P) are provided to the model, and $E_i^T \in KB$ are pre-computed offline to minimize real-time computation costs.

2.2. VELCRO Components

To produce entity and mention embeddings for similarity search, we design VELCRO using a popular two-tower model architecture (Radford et al., 2021) in which one submodel produces contextual embeddings of medical images and the other produces textual embeddings of the knowledge base.

Notably, the visual encoder processes both the image and a bounding box prompt indicating the user's intent. We implement it using the Segment Anything Model (SAM) due to its promptable task formulation and strong segmentation performance (Kirillov et al., 2023). SAM learns to generate object segmentations via user prompts in the form of bounding boxes, points, or "scribbles."

As shown in Fig. 2 (top left), SAM is composed of an image and prompt encoder, and a mask decoder. After the image and prompt are processed by their respective encoders, a set of special mask (`[M]`) tokens is appended to the latter. Since VELCRO only uses one prompt-type (bounding boxes), SAM will generate three masks (one per added `[M]` token). This is done to address potential prompt ambiguity by generating multiple candidate segmentation masks. Following MedSAM (Ma et al., 2024a), we extract the token with the highest estimated IoU for further processing.

After encoding, the mask token is used by the mask decoder to construct a segmentation mask. During this process, the mask token embedding is updated with information from both the image and prompt. This is crucial for the adaptation of SAM to our VEL task because it ensures that the surrounding image context is incorporated into the latent representation of the user's targeted visual mention rather than discarded.

To instantiate the text tower (Fig. 2, bottom left), we use BERT, a transformer-based language model (Devlin et al., 2019) that produces contextual representations of entire input sequences. BERT is often used in entity linking tasks such as named entity recognition (Hebert et al., 2022). As in SAM, we use the embedding of an appended [CLS] token to represent a KB entry.

2.3. Putting it All Together

We are interested in linking semantically meaningful local mentions within a larger image to their relevant entity. However, SAM does not naturally encode semantic information (no RoI label, just location), so this must be injected during training. We achieve this using two parallel dataflows, labelled in Fig. 2 as the **Promptable Mention Detection Pathway** and the **Contextual RoI Linking Pathway**.

The promptable mention detection pathway resembles typical segmentation finetuning approaches. However, whereas finetuning pipelines often freeze the image encoder, we make it learnable to enable additional domain adaptation. Key to our method is the participation of the mask token in both pathways, enabling simultaneous mask generation and classification as in non-promptable whole-image segmentation approaches (Cheng et al., 2022).

Overlap between the two pathways occurs via the extraction of the final `[M]` token produced by the SAM decoder after cross-attention, diverting this token towards the contextual RoI linking pathway. Extraction occurs at this time such that that $E_i^V = \texttt{[M]}_i$ (with $MentionDetector(I_i, P_i)$. This ensures that the mask embedding contains both contextual information about the image and the geometry of the RoI.

BERT is used to encode T_i and feed the output `[CLS]` embedding through an additional learnable projection layer to project it into the mask token's latent space. Contrastive learning is then applied such that similar mask and entity embeddings are pushed closer together, while dissimilar embeddings are pushed apart, ensuring that defining semantic features are well-separated in the latent space (Radford et al., 2021). This allows entity retrieval to occur based on the distance between the two embeddings. As a result, VELCRO is trained to generate masks that are both high-quality at a pixel level (module (a)) and semantically grounded (module (b)).

Finally, we discuss the loss function to enable entity linking using segmentation-oriented embeddings. To perform both mask generation and entity linking, we combine SAM's mask quality loss function (Kirillov et al., 2023) (denoted on the right of Fig. 2 as the difference between the predicted mask $\hat{M}_i$ and the ground truth mask provided in the training dataset M_i) with a CLIP-like contrastive entity loss function. To compute the mask loss, we first retrieve the upscaled predicted mask $\hat{M}_i$ produced by the final steps of the mask decoder. The loss is then computed as

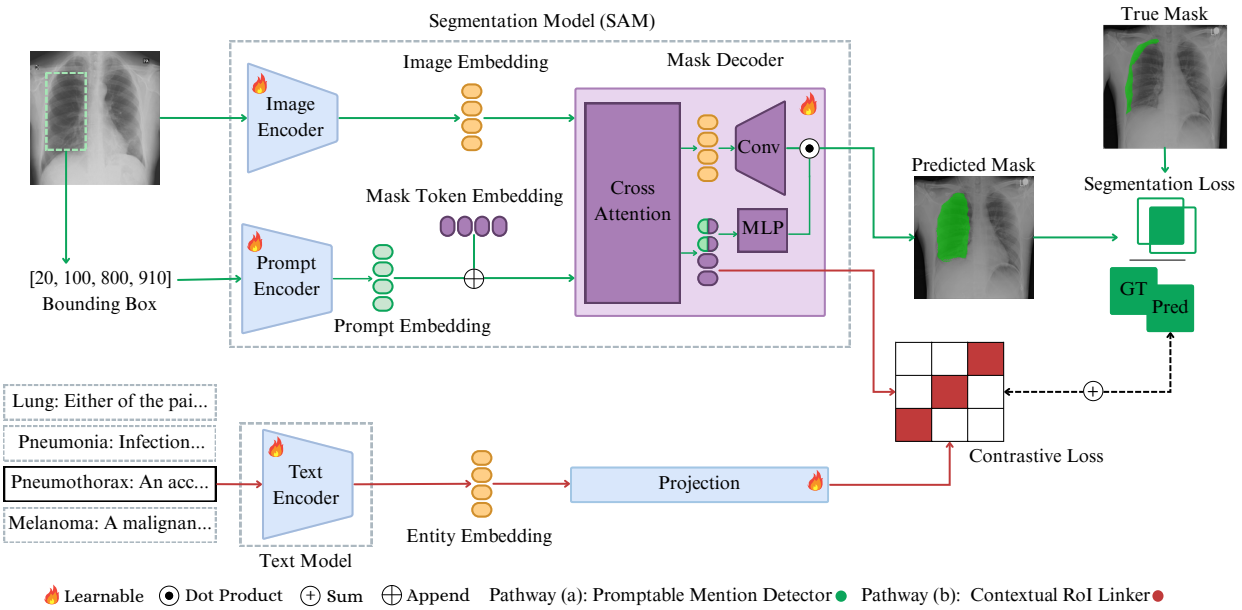

Figure 2: VELCRO architecture. **Parallel data flows** support multi-task learning through the joint optimization of segmentation and entity alignment tasks.

a linear combination of focal loss (Lin et al., 2017) and DICE loss (Milletari et al., 2016), such that, for a single datapoint with ground truth mask M_i and predicted mask $\hat{M}_i$ with N pixels, and

$$p_t = \begin{cases} \sigma(\hat{M}_i) & \text{if } M_i = 1 \\ 1 - \sigma(\hat{M}_i) & \text{if } M_i = 0 \end{cases} \tag{1}$$

where p_t is the probability that a pixel in $\hat{M}_i$ equals a pixel in M_i for all N pixels, and DICE and focal loss are as defined below.

$$L_i^{dice}(\hat{M}_i, M_i) = 1 - \frac{2\sum_{i=1}^{N} M_i \hat{M}_i}{\sum_{i=1}^{N} M_i + \sum_{i=1}^{N} \hat{M}_i} \tag{2}$$

$$L_i^{focal}(p_t) = -(1-p_t)^{\gamma} log(p_t) \tag{3}$$

In the focal loss, γ controls the degree to which the model concentrates its learning on smaller foreground areas, which often have fewer pixels than background areas. In practice, we set $\gamma = 2$ following Lin et al. (2017).

This gives us the final segmentation loss,

$$L_i^{seg} = \lambda_f L_i^{focal} + \lambda_d L_i^{dice} \tag{4}$$

where $\lambda_f = 20.0$ and $\lambda_d = 1.0$, as in the original SAM paper (Kirillov et al., 2023).

We use cosine similarity to compute the distance between entity and mask embeddings E_j^t and E_i^m.

$$S(E_i^V, E_j^T) = \frac{E_i^V \cdot E_j^T}{||E_i^V||_2 ||E_j^T||_2} \tag{5}$$

This cosine similarity score is then used in InfoNCE loss (van den Oord et al., 2019) as our entity linking loss. This loss includes a learnable temperature parameter τ, which we initialize to 0.5, to emphasize differences between embeddings following Chen et al. (2020). Thus, the loss function for a single mask embedding is defined as

$$L_i^{entity} = log \frac{exp(S(E_i^V, E_j^T)/\tau)}{\sum_{j=1}^{N} exp(S(E_i^V, E_j^T)/\tau)} \tag{6}$$

Finally, we sum the segmentation and entity losses and compute the average over batches of size B to create our final aggregate loss L_{total}:

$$L^{total} = \frac{1}{B} \sum_{i=1}^{B} (L_i^{entity} + L_i^{seg}) \tag{7}$$

3. Related Work

Numerous formulations of VEL have been studied. In multi-modal entity linking (MEL), VEL inputs con-

tain both images and text (Moon et al., 2018; Bielefeld et al., 2024; Hu et al., 2023; Xiao et al., 2025; Sui et al., 2024; Dost et al., 2020; Song et al., 2024; You et al., 2023; Gan et al., 2021; Xu et al., 2025). For example, a user may supply an image and a textual prompt such as "What is the grey blob in the top-left corner of the image?" However, MEL models may struggle when the text is missing or vaguely worded (Sun et al., 2022; Xu et al., 2025). On the other hand, in visual-to-textual entity linking (V2TEL), images and, optionally, visual prompts in the form of bounding boxes or points, are provided as input (Sun et al., 2022; Tilak et al., 2017). As seen in Fig. 1 (a) and (b), existing V2TEL methods either rely on whole-image alignment (Lerner et al., 2024) or discard possibly relevant context outside a bounding box (Sun et al., 2022; Tilak et al., 2017). VELCRO is a context-aware solution to V2TEL, capturing context by repurposing and aligning the representations produced by an image segmentation model (Fig. 1 (c)).

Related image understanding problems include object detection and semantic image segmentation. Given an image, the goal is to find bounding boxes (detection) or precise locations (segmentation) of objects within it and predict class labels for each found object[1]. In closed-vocabulary variants of the above problems, the labels come from a fixed list of known classes. In open-vocabulary variants, the input also consists of a textual prompt describing the objects in the image, and the resulting object labels are typically drawn from the nouns in this prompt (Xu et al., 2022). These models can be adapted to our VEL problem: select the predicted bounding box or segmentation mask nearest to the user-provided box, take its predicted label, and match the label to the closest KB entity. Of course, these models would have to be trained to predict labels corresponding to the KB entries in closed-vocabulary settings, or their textual prompt would have to include all the possible KB entries in open-vocabulary cases.

We will experimentally compare VELCRO with YOLO (Redmon et al., 2016), a popular closed-vocabulary model whose variants can perform either detection or segmentation, and ZegFormer (Ding et al., 2022), a recent open-vocabulary semantic segmenter with a conceptually similar architecture to VELCRO. Specifically, ZegFormer uses a segmentation model, MaskFormer (Cheng et al., 2021), to generate mask embeddings from mask tokens, and aligns these masks with text label embeddings generated by a CLIP text encoder. However, while we use the same mask token for alignment and class-agnostic mask generation, they instead project mask tokens into two separate spaces, producing two single-task embeddings rather than a single dual-task embedding. Furthermore, they freeze CLIP's text encoder and only finetune the model's segmentation backbone, preventing the model from learning additional text-based semantic content. During inference, CLIP-generated image embeddings are also used to assist with labelling, while we exclusively use mask embeddings for training and inference.

1. Note that SAM, used in VELCRO, is not a semantic segmentation model; it only produces object locations (segmentation masks) but does not predict their labels.

4. Experiments

We first describe the data curation (Sec. 4.1) and baselines (Sec. 4.2). We then compare VELCRO to various baselines (Sec. 4.3), detail two ablation studies (Sec. 4.4) and present a qualitative study (Sec. 4.5). See Appendix A for dataset preparation and experimental setup details.

4.1. Data

Recall from Section 2.2 that VELCRO requires training data of the form $D = \{(I_1, P_1, M_i^V, T_1), \cdots (I_N, P_N, M_N^V, T_N)\}$, with images I_i, RoIs with their bounding boxes P_i, fine-grained mention locations M_i^V (to train the segmentation component), and the corresponding KB entities T_i (to train the entity linking component). However, since no full dataset of the desired form was available, we instead obtained (I_i, P_i, M_i^V) triplets from a medical image instance segmentation dataset. The segmentation targets described in this dataset were then used to guide T_i retrieval from a medical KB.

We obtained the (I_i, P_i, M_i^V) triplets from an open-source medical segmentation dataset (Ma et al., 2024b)[2]. This dataset combines multiple individual medical segmentation datasets across a variety of imaging and video modalities and domains. We excluded datasets focused on video data as they are

2. This dataset is frequently updated. Some of the entity target names and datapoints used in our work may be different from the most up-to-date version available at https://www.codabench.org/competitions/5263/. The lists of datasets and entities used in our work are released alongside our code.

outside the scope of this work. Our final dataset consisted of 29 sub-datasets from the parent dataset. Since the original test data has not been publicly released, we created train/test/validation splits by splitting at the patient case level. This was done to avoid data leakage. These sets comprise 435,208, 52,701, and 49,920 examples from 41,481, 5,187, and 5,186 patients, respectively. We ensured that all entities are present in the resulting training and testing sets.

Although the segmentation dataset contains some entity information in the form of name-based segmentation targets, it does not include entity descriptions. Thus, we use the Unified Medical Language System (UMLS) Metathesaurus to obtain the corresponding KB entries T_i (Bodenreider, 2004). UMLS comprises over 2 million data entries for 900,000 medical concepts, selected from 60 biomedical vocabularies, of which 20 terms match the segmentation targets in our image dataset. We performed instance label-UMLS entity resolution by querying target names from the constituent sub-datasets in the dataset from Ma et al. (2024b). In cases where an exact match could not be found, we manually analyzed the results for semantic similarity (e.g. the target "lung lesion" resolves to the UMLS entity "lesion of the lung"). Targets which produced no exact or sufficiently similar match were excluded from the resulting dataset.

Each final T_i consists of a text label and description, which are concatenated with the tokens `[TITLE]` and `[BODY]` indicating the start of each component (the entity name and its corresponding description) prior to text encoding. For example, the term "lesion of the lung" is represented as "`[TITLE] lesion of the lung [BODY] a localized pathological or traumatic structural change, damage, deformity, or discontinuity of the lung.`" Further details on dataset curation and preprocessing are available in Appendix A.1.

4.2. Baselines

The first two baselines use CLIP and correspond to Fig. 1 (a) and (b), respectively: *CLIP-Full*, where the input is the whole image, and *CLIP-Crop*, where the input is restricted to the RoI within the bounding box.

The next three baselines adapt related image understanding tasks to our VEL problem (recall Sec. 3): the latest version of the closed-vocabulary model YOLO in semantic object detection mode (YOLOv12-obj) fine-tuned to produce labels corresponding to our KB entities, the same model in semantic image segmentation mode (YOLOv12-seg), and ZegFormer, an open-vocabulary semantic segmentation model that can perform VEL if its input consists of an image and a prompt with all the possible KB entries that could be linked. Since YOLO and Zegformer both produce multiple segmentation predictions per image during inference, we select the mask with the highest IoU to evaluate optimistic upper-bound performance.

The last three baselines correspond to multimodal generative models performing VEL in zero-shot mode: LLaVA-Next (Liu et al., 2024), a leading open-source model, GPT-4o mini, a member of the GPT-4 class of models from OpenAI (OpenAI et al., 2024), and MedGemma, a medically-oriented LLM (Sellergren et al., 2025). MedGemma was also evaluated in a few-shot setting after initial zero-shot results analysis.

To align model inputs with the visual bounding-box prompts used by VELCRO, we overlay bounding boxes onto our dataset images. KB entities (names and descriptions) are also provided to the LLM, which is instructed to select the KB entity that best describes the visual mention within the bounding box. We performed minor prompt engineering: For instance, in LLaVA-Next, embedding the image directly into the prompt (e.g. "Classify the object in the red bounding box <image> ...", where <image> represents the provided image) allowed the model to more accurately respond to the prompt instructions with KB entities than when providing the image before the text prompt or after the KB context. Additional details regarding LLM prompting can be found in Appendix A.3.

4.3. Comparison to Baselines

Table 1 shows the performance of VELCRO and the nine baselines on four metrics: F1, recall, precision, and intersection over union (IoU) for systems that perform segmentation. We label the methods according to the nature of the provided prompt (full image vs. bounding box). Bounding boxes are used as input in CLIP-Crop, VELCRO, and for all LLMs.

VELCRO outperforms all baselines in entity linking metrics, with an F1 of 95.2%. CLIP-Crop performs slightly worse than CLIP-Full, suggesting that performance degrades only slightly when the image's

Table 1: VEL performance of VELCRO and baselines across various visual prompting methods.

Finetuned	Prompt	Model	F1	Precision	Recall	IoU
✓	Full Image	CLIP-Full	0.810	0.814	0.839	-
✓	Bounding Box	CLIP-Crop	0.807	0.795	0.829	-
✓	Full Image	YOLOv12-obj	0.635	0.665	0.724	0.539
✓	Full Image	YOLOv12-seg	0.659	0.702	0.768	0.554
✓	Full Image	ZegFormer	0.839	0.858	0.887	0.780
✓	Bounding Box	**VELCRO**	**0.952**	**0.934**	**0.981**	**0.806**
✗	Bounding Box	LLaVA-Next	0.113	0.130	0.173	-
✗	Bounding Box	GPT-4o mini	0.179	0.270	0.313	-
✗	Bounding Box	MedGemma (Zero-shot)	0.199	0.282	0.279	-
✗	Bounding Box	MedGemma (Few-shot)	0.275	0.391	0.393	-

surrounding context is removed. The results obtained by ZegFormer and VELCRO further indicate that an alignment scheme that leverages RoI information without removing whole-image awareness is more conducive to the task of VEL than either strategy separately. However, VELCRO outperforms ZegFormer in F1 by 12.6%. This suggests that applying open-domain semantic models, such as CLIP, as classification heads in specialized domains like medicine requires additional text-encoder training, as demonstrated by VELCRO.

VELCRO outperformed both YOLO models by a significant margin. In particular, YOLO models failed to generate valid masks on 20.1% of the testing dataset in the segmentation setting and 34.4% in object detection and were especially challenged by small entities such as cells. This behaviour can be adjusted by tuning hyperparameters, such as the confidence threshold for detected objects, to increase the number of predicted masks. This, however, also increases false-positive predictions.

Finally, both open-domain LLMs exhibited poor zero-shot performance. Although GPT-4o mini achieved better results with an F1 of 17.9%, this still fell considerably short of all non-LLM baselines. In the zero-shot setting, MedGemma achieved slightly higher F1 performance than its non-medical peers (+11.2/% F1 over GPT-4o mini), though this was still significantly lower than all finetuned models. The all-around poor task performance observed with off-the-shelf LLMs thus indicates a lack of adequate latent reasoning capabilities suitable for VEL. In a few-shot setting with two examples, however, MedGemma achieved a 38.2% relative F1 increase over its zero-shot counterpart and a 53.6% increase over GPT-4o mini. This increase suggests that, short of model finetuning, few-shot prompting is a promising path toward achieving comparable VEL performance.

4.4. Ablations

4.4.1. KB Ablation

To study the effect of dense KB entities on linking in VELCRO and in the LLMs, which displayed poor task performance, linking was further evaluated with shortened entity descriptions of the form `''An image of a {entity}''` (Table 2). Our results indicate that overall linking performance was similar between the two approaches, with full-caption VELCRO yielding slightly higher performance over its ablated counterpart. However, VELCRO's negligible decrease in segmentation performance further suggests that variances in entity description lengths were minimally influential. LLaVA-Next demonstrated improved overall linking performance under these simplified conditions, with a substantial recall increase from 17.3% to 31.2%, suggesting it likely suffered from overcontextualization during full-KB inference. GPT-4o mini was less sensitive to KB variances and displayed consistent performance across settings, but still only achieved a maximum F1 of 17.9%. Zero-shot MedGemma with ablated captions achieved higher precision and lower recall than its full-caption counterpart, indicating a reduced incidence of false positives at the cost of fewer overall positive predictions. Finally, few-shot, short-caption MedGemma demonstrated a 24.7% decrease in overall relative performance. Based on our experiments, LLMs were found to be more sensitive to KB caption density than VELCRO.

Table 2: Comparison between entity description approaches for VELCRO and all LLMs.

Entity Description	Model	F1	Precision	Recall	IoU
Full	**VELCRO**	**0.952**	**0.934**	**0.981**	**0.806**
	LLaVA-Next	0.113	0.130	0.173	-
	GPT-4o mini	0.179	0.270	0.319	-
	MedGemma (Zero-shot)	0.199	0.282	0.279	-
	MedGemma (Few-shot)	0.275	0.391	0.393	-
Short	VELCRO	0.948	0.926	0.981	0.798
	LLaVA-Next	0.196	0.238	0.312	-
	GPT-4o mini	0.175	0.253	0.371	-
	MedGemma (Zero-shot)	0.182	0.366	0.255	-
	MedGemma (Few-shot)	0.207	0.287	0.310	-

4.4.2. Model Size Ablation

VELCRO's consistent performance across entity description settings prompted further evaluation of the influence of different submodel component sizes. We hypothesized that when VELCRO's vision model size increased, segmentation performance would improve. Likewise, increases in text model size would improve linking. We compared our base model VELCRO$_B$ with 203M parameters to VELCRO$_T$, a variant that uses a larger BERT$_L$ text model with 340M parameters instead of BERT$_B$ with 110M, and VELCRO$_V$, a variant with a larger SAM$_L$ vision model with 312M parameters.

The results in Table 3 indicate that larger model sizes did not constantly improve performance. Nearly all models achieved comparable performance in both linking and segmentation, suggesting that VELCRO$_B$ is sufficient for the task on this dataset. We further investigated this by freezing different parts of the model architecture to explore the impact of individual component finetuning on performance. In VELCRO$_{B/V}$, SAM's image encoder, ViT (Dosovitskiy et al., 2021) was frozen, resulting in a smaller model with 113M trainable parameters. In VELCRO$_{B/T}$, BERT was frozen such that only the text projection layer was learnable, which produced a smaller model with 119M parameters. We observed that freezing BERT did not significantly affect performance, as evidenced by VELCRO$_{B/T}$'s F1 of 93.7% in comparison to VELCRO$_B$'s score of 95.2%. In contrast, freezing SAM's image encoder caused a 2% performance drop in linking F1 and a 13.6% drop in segmentation IoU. This implies that learning effective visual representations is the primary driver behind region localization and, consequently, linking.

4.5. Qualitative Study

To explore performance differences among segmentation-based methods, we conduct a qualitative analysis across several cases. Fig. 3 shows two examples. In case (a), top row, VELCRO identifies the correct entity (the optic cup structure within the eye), while both ZegFormer and YOLO-obj fail. Notably, the predicted entity for each of these models, the optic disk, is a larger structure surrounding the optic cup, suggesting that imprecise RoI localization likely contributed to this failure. While YOLO-seg predicts the correct entity, its segmentation mask is also poor, indicating that this model may have failed to learn the distinction between the smaller optic cup and larger optic disk. Only VELCRO performs both high-quality object localization *and* correct entity linking, demonstrating harmony between RoI refinement and semantic alignment.

Case (b), bottom row, visualizes VELCRO's superior linking performance even in scenarios with comparable detected RoIs. Despite all models producing similar segmentation masks, only VELCRO correctly linked the generated RoI, demonstrating a stronger learned separation between RoI features corresponding to the entities "glioma" and "brain infarction". We note, however, that the ability of the baselines to perform linkage to another entity within the same organ, a brain infarction, suggests an understanding of entity *location*, if not semantics.

5. Conclusion and Future Work

We presented VELCRO, a novel model for VEL that utilizes user-supplied bounding boxes as queries. VELCRO's ability to perform both segmentation and

Table 3: VELCRO performance at varying model sizes.

Model	Components	Parameter#	F1	Precision	Recall	IoU
$\text{VELCRO}_{B/V}$	$\text{SAM}_B/ViT + \text{BERT}_B$	113M	0.844	0.812	0.945	0.691
$\text{VELCRO}_{B/T}$	SAM_B + Text Proj.	119M	0.937	0.914	0.985	0.793
VELCRO_B	$\text{SAM}_B + \text{BERT}_B$	203M	**0.952**	**0.934**	**0.981**	**0.806**
VELCRO_T	$\text{SAM}_B + \text{BERT}_L$	429M	0.950	0.933	0.978	0.783
VELCRO_V	$\text{SAM}_L + \text{BERT}_B$	422M	0.923	0.932	0.944	0.801

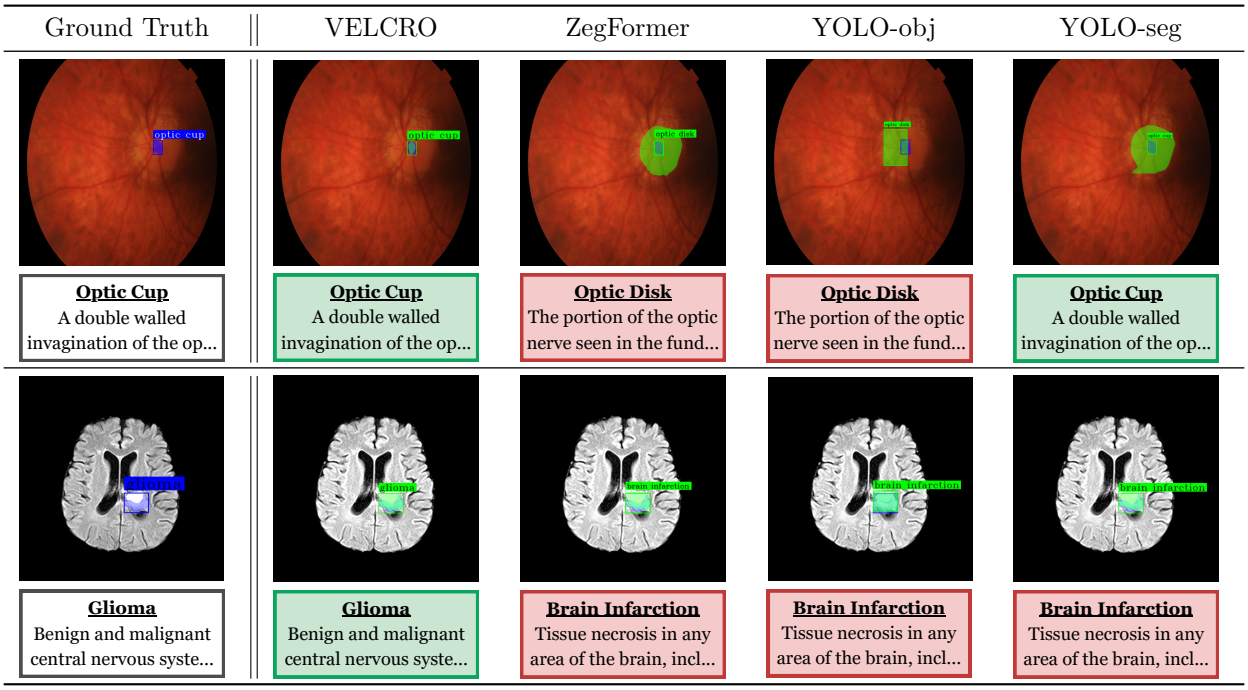

Figure 3: **Case Studies:** Ground-truth segmentation masks and named entities (left), and the top 1 entity and detected entity location for all RoI-aware models (VELCRO, ZegFormer, YOLO-obj, and YOLO-seg).

entity linking eliminates the need for dual-model pipelines by collapsing the two primary requirements of VEL—mention detection and linking—into a single, unified model capable of generating high-quality, semantically-oriented segmentation masks. Our experimental results indicate that exploring segmentation model extensions to semantically oriented tasks such as VEL is a promising direction.

Our experiments used a subset of the UMLS Metathesaurus. Expanding our image and entity datasets would allow us to evaluate the scalability of our approach and explore VELCRO's stability under more challenging conditions. This would be especially valuable for analyzing the influence of dense KB descriptions on linking, as our results suggested a heavier reliance on visual data than textual semantic knowledge. Specifically, incorporating additional finer-grained, closely-related entities and conducting top-k metric evaluations would be beneficial for characterizing task performance in more realistic clinical settings. Furthermore, the image encoder backbones (CLIP, YOLO, ZegFormer, and SAM) used in our experiments are pretrained on different image resolutions, with SAM requiring the highest resolution and consequently learning from more detailed features. Studying the effect of image resolution on linking performance is thus another avenue for future work, specifically using lower-resolution imaging data to increase computational efficiency. Finally, our experiments focus on image-to-text entity linking, but other variations, such as image-to-image-plus-text entity linking, are also applicable to medical imaging.

Acknowledgments

Hebert gratefully thanks financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Vanier Graduate Scholarship as well as from the IEEE Nick Cercone Graduate Scholarship. The authors also thank the institutional support provided by the University of Waterloo.

References

Philipp Bielefeld, Jasmin Geppert, Necdet Güven, Melna John, Adrian Ziupka, Lucie-Aimée Kaffee, Russa Biswas, and Gerard De Melo. Wiki-VEL: Visual entity linking for structured data on wikimedia commons. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 186–194, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.alvr-1.16. URL https://aclanthology.org/2024.alvr-1.16/.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue): D267–D270, January 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh061.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, July 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/950a4152c2b4aa3ad78bdd6b366cc179-Paper.pdf.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. On visual-textual-knowledge entity linking. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 190–193, 2020. doi: 10.1109/ICSC.2020.00039.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 993–1001, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475400. URL https://doi-org.proxy.lib.uwaterloo.ca/10.1145/3474085.3475400.

Liam Hebert, Raheleh Makki, Shubhanshu Mishra, Hamidreza Saghir, Anusha Kamath, and Yuval Merhav. Robust candidate generation for entity linking on short social media texts. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 83–89, Gyeongju, Republic of Korea, October 2022.

Association for Computational Linguistics. URL https://aclanthology.org/2022.wnut-1.8/.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12065–12075, October 2023.

Karthik K, Sowmya Kamath S, Supreetha R, and Ashish Katlam. Content-based medical retrieval systems with evidence-based diagnosis for enhanced clinical decision support. *Expert Systems with Applications*, 272:126678, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2025.126678. URL https://www.sciencedirect.com/science/article/pii/S0957417425003008.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval*, pages 421–438. Springer, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024a. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z.

Jun Ma et al. Efficient medsams: Segment anything in medical images on laptop, 2024b. URL https://arxiv.org/abs/2412.16085.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1186. URL https://aclanthology.org/P18-1186/.

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

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.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Shezheng Song, Shan Zhao, ChengYu Wang, Tianwei Yan, Shasha Li, Xiaoguang Mao, and Meng Wang. A dual-way enhanced framework from text matching point of view for multimodal entity linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19008–19016, March 2024. doi: 10.1609/aaai.v38i17.29867. URL https://ojs.aaai.org/index.php/AAAI/article/view/29867.

Xuhui Sui, Ying Zhang, Yu Zhao, Kehui Song, Baohang Zhou, and Xiaojie Yuan. MELOV: Multimodal entity linking with optimized visual features in latent space. In *Findings of the Associ-*

ation for Computational Linguistics: ACL 2024, pages 816–826, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.46. URL https://aclanthology.org/2024.findings-acl.46/.

Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Visual named entity linking: A new dataset and a baseline. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 2403–2415, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.178. URL https://aclanthology.org/2022.findings-emnlp.178/.

Neha Tilak, Sunil Gandhi, and Tim Oates. Visual entity linking. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 665–672, 2017. doi: 10.1109/IJCNN.2017.7965916.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. Grounding language models for visual entity recognition. In *Computer Vision – ECCV 2024*, pages 393–411, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73247-8.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

Zhengfei Xu, Sijia Zhao, Yanchao Hao, Xiaolong Liu, Lili Li, Yuyang Yin, Bo Li, Xi Chen, and Xin Xin. Reverse region-to-entity annotation for pixel-level visual entity linking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12): 12981–12989, April 2025. doi: 10.1609/aaai.v39i12.33416. URL https://ojs.aaai.org/index.php/AAAI/article/view/33416.

Jiuxiang You, Zhenguo Yang, Qing Li, and Wenyin Liu. A retriever-reader framework with visual entity linking for knowledge-based visual question answering. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 13–18, 2023. doi: 10.1109/ICME55011.2023.00011.

Appendix A. Additional Experimental Details

A.1. Data Curation Details

As mentioned earlier, we obtained the (I_i, P_i, M_i^V) triplets from an open-source medical segmentation dataset. Since this dataset was originally created for universal instance segmentation, datapoints were of the form (I_i, M_i), where M_i is a segmentation mask potentially containing multiple entity instances. Raw M_i masks vary in format across sub-datasets, with some assigning each distinct object in an image a unique numerical label, regardless of its semantic meaning. This makes the disambiguation of unique instances of the same entity from instances of different entities a non-trivial task, which is nevertheless necessary for the accurate assignment of accompanying T_i descriptions. We thus left the incorporation of these sub-datasets as future work and instead focused on normalizing datasets containing easily separable M_is.

The remaining M_is were refined by identifying non-contiguous entities within each mask. Each M_i was expanded into a set of $\{M_i^{V_i} \cdots M_i^{V_j}\}$ submasks, each identifying the location of a single distinct entity. Bounding boxes for each entity were then extracted using the contours of its associated M_i^V to produce the final (I_i, P_i, M_i^V) datapoint. For example, a mask containing segments for both lungs would be split into two masks with one lung each. This was done to ensure that each datapoint neatly maps to a single T_i.

Some medical imaging modalities produce three-dimensional scans of the patient's body. An I_i datapoint (and its associated M_i) may thus take the form $(T \times H \times W)$, where T refers to the number of individual slices (or "time steps") which are composed to produce the full image. In these instances, 3-D scans are decomposed into $(1 \times H \times W)$ 2-D slices, each representing an individual image and associated entities of interest within the patient's body at a single time step. The collection of these 2-D slices is referred to as a "case," since it represents information about one patient's health at a specific time. In 3-D views of a patient's body, there may exist slices where no entities of interest are present in the mask. In such instances, we discard the image and mask and keep only the slices containing relevant information.

Fig. 4 visualizes dataset entity distributions after expansion. We observe extreme class imbalances, with the dataset largely being dominated by the entities "glioma,", "lung," and "cell," which we address with a combination of over and under-sampling during training. Table 4 shows some examples of UMLS entities represented in the resulting dataset.

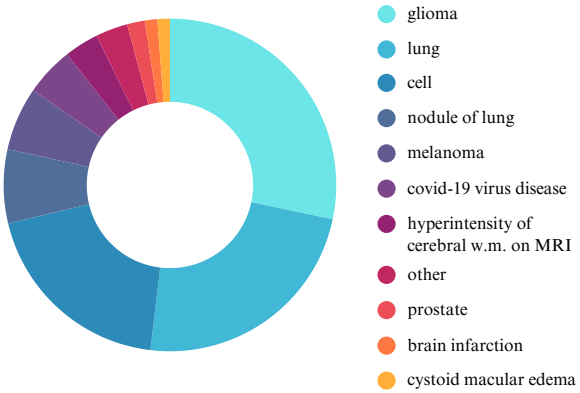

Figure 4: **Post-processed Entity Distribution**

Table 4: Examples of UMLS entities used in this work

lung	Either of the pair of organs occupying the cavity of the thorax that effect the aeration of the blood.
cervical cancer	A tumor of the uterine cervix.
nodule of lung	A benign or malignant small, oval or round growth in the lung. It can be due to infectious, inflammatory or neoplastic processes.
optic disk	The portion of the optic nerve seen in the fundus with the ophthalmoscope. It is formed by the meeting of all the retinal ganglion cell axons as they enter the optic nerve.

A.2. Experimental Environment

Experiments were primarily conducted using two RTX 6000 Ada Generation GPUs with 32 GB of RAM and 8 CPU cores. We use an 80-10-10 train-test-validation split with a seed of 42. To manage dataset imbalance in the training data, weighted random sampling was performed to undersample datapoints from overrepresented entities and oversample datapoints from underrepresented entities.

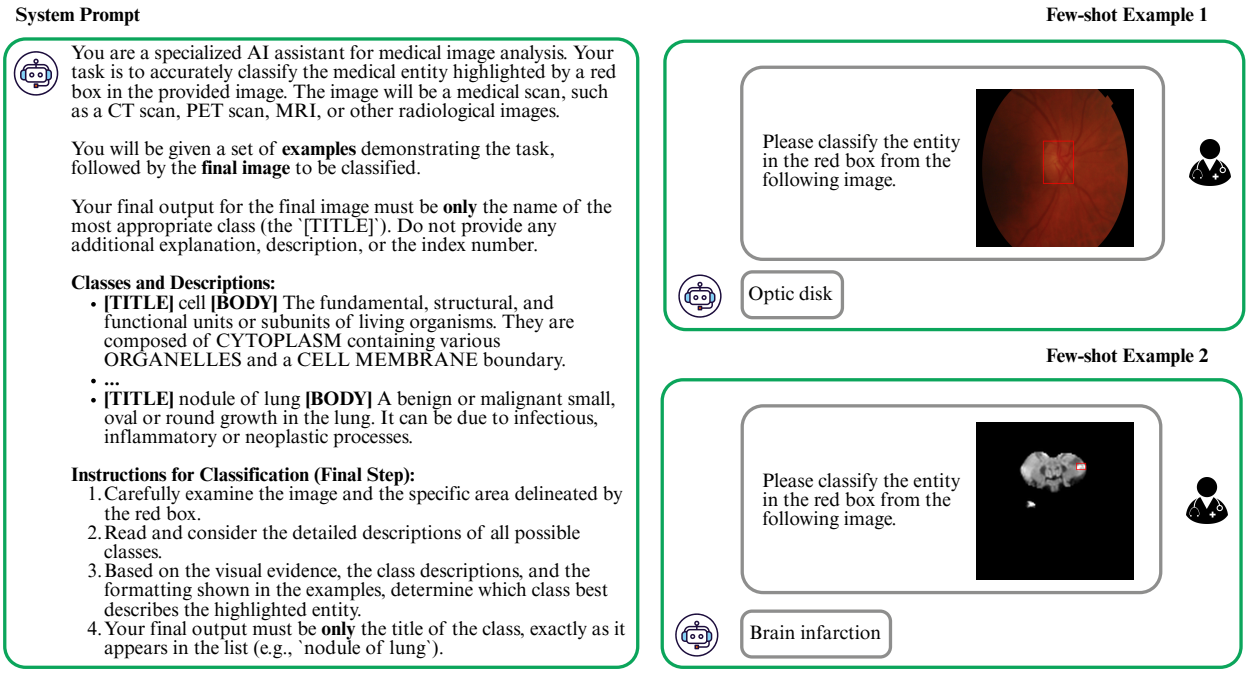

Figure 5: **Few-shot MedGemma prompt:** Best LLM results were achieved using few-shot prompting with two examples. Some class descriptions have been omitted from the figure for brevity.

VELCRO is initialized using the pretrained weights from SAM-base with 16 patches[3] and BERT-base[4]. We specifically use the uncased version of BERT-base, which ignores capitalization during sequence tokenization. All CLIP baselines were initialized with the pretrained weights from CLIP-base with 16px patches[5]. YOLOv12 was initialized using the pretrained weights from the largest YOLOv12 variant, YOLOv12-X, with 59.1M trainable parameters for object detection and 64.5M trainable parameters for segmentation[6]. To process images and text, LLaVA-Next comprises a visual encoder and a text-based LLM backbone. We perform inference using the pretrained weights of LLaVA-Next composed of a frozen CLIP-large/14-patch image encoder and Meta's LLaMA 3 LLM backbone with 8B parameters[7]. We perform inference on GPT-4o mini using OpenAI's API. MedGemma experiments were conducted using the multi-modal 27B-it variant[8].

We select all CLIP-based model hyperparameters using a grid search with options $[1e^{-6}, 1e^{-5}]$ for learning rate and a batch size of $[16, 32, 64, 128, 256]$. For VELCRO, we restrict the batch size to 8 to account for the increased computational complexity of SAM and set the learning rate to $1e^{-5}$. Both YOLOv12 models and ZegFormer were similarly trained with a learning rate of $1e^{-5}$ and a batch size of 32. To simulate a loss function similar to VELCRO and emphasize the importance of both accurate mention localization and entity identification, classification, and object detection losses were weighted equally. All finetuned models used an AdamW optimizer without learning rate decay and a learning rate on a plateau scheduler to decrease the learning rate by $1e-1$ when a monitored metric, per-epoch validation loss, stops improving.

Finally, we use in-batch negatives when performing contrastive learning for all trained embedding alignment models. This means that, for each mention and entity embedding in a training batch, all other

3. https://huggingface.co/facebook/sam-vit-base
4. https://huggingface.co/google-bert/bert-base-uncased
5. https://huggingface.co/openai/clip-vit-base-patch16
6. https://github.com/sunsmarterjie/yolov12/tree/main
7. https://huggingface.co/llava-hf/llama3-llava-next-8b-hf
8. https://huggingface.co/google/medgemma-27b-it

unassociated (or negative) embeddings of the opposite modality are considered dissimilar when manipulating embedding distances in latent space. During inference, however, the complete set of candidate entities was provided for fair evaluation.

A.3. LLM Prompting

All LLMs were evaluated using greedy sampling. Few-shot prompting experiments with 2, 5, 6, and 8 examples were performed on MedGemma. The classes represented in these examples were selected from the set of lowest-performing entities in the zero-shot setting. Few-shot prompting with two examples achieved the highest performance. The final, full-caption prompt with both few-shot examples is presented in Fig. 5.