

Fine-Grained Visual Entity Linking through Promptable Segmentation: Applications in Medical Imaging

by

Kathryn Carbone

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Kathryn Carbone 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Image analysis in domains that produce large amounts of complex visual data, like medicine, is challenging due to time and labour-constraints on domain experts. Visual entity linking (VEL) is a preliminary image processing task which links regions of interest (RoIs) to known entities in structured knowledge bases (KBs), thereby using knowledge to scaffold image understanding. We study a targeted VEL problem in which a specific user-highlighted RoI within the image is used to query a textual KB for information about the RoI, which can support downstream tasks such as similar case retrieval and question answering. For example, a doctor reviewing an MRI scan may wish to obtain images with similar presentations of a medically relevant RoI, such as a brain tumor, for comparison. By linking this RoI to its corresponding KB document, search of an imaging database with VEL-guided automatically-generated tags can be performed in a knowledge-aware manner based on exact or semantically similar entity tag matching.

Cross-modal embedding models like CLIP present straightforward solutions through the dual encoding of KB entries and either whole images or cropped RoIs, which can then be matched by a vector similarity search between these respective learned representations. However, using the whole image as the query may retrieve KB entries related to other aspects of the image besides the RoI; at the same time, using the RoI alone as the query ignores context, which is critical for recognizing and linking complex entities such as those found in medical images. To address these shortcomings, this thesis proposes VELCRO – visual entity linking with contrastive RoI alignment – which adapts an image segmentation model to VEL using contrastive learning by aligning the contextual embeddings produced by its decoder with the KB. This strategy preserves the information contained in the surrounding image while focusing KB alignment specifically on the RoI. To accomplish this, VELCRO performs segmentation and contrastive alignment in one end-to-end model via a novel loss function that combines the two objectives. Experimental results on medical VEL show that VELCRO achieves an overall linking accuracy of 95.2% compared to 83.9% for baseline approaches.

Acknowledgements

I would first like to thank my supervisors, Dr. Lukasz Golab and Dr. Robin Cohen, for their guidance, insightful feedback, and patience throughout my master’s program. I am deeply grateful for their support of my curiosity and willingness to let me explore diverse topics within a field which I am passionate about: healthcare AI. Their openness to my ideas, combined with their expertise, has been instrumental in my growth as a researcher and independent thinker. I would also like to extend my sincere appreciation for my committee members, Dr. Paulo Alencar and Dr. Jesse Hoey, for contributing their time and expertise to the evaluation of my work.

Furthermore, I would like to acknowledge the financial support provided by the NSERC Responsible AI program, as well as for creating opportunities for me to engage with researchers from a wide range of disciplines. These interactions have broadened my perspective and helped me develop as a more thoughtful and conscientious researcher.

I am also grateful for my friends and colleagues, especially those in the AI for Social Good Group, for their encouragement, camaraderie, and thoughtful advice throughout this journey. I would especially like to thank Liam Hebert, who first collaborated with me on the earliest version of this project in CS 846 and whose insights, enthusiasm, and continued involvement have helped shape its evolution into my thesis.

I owe my deepest gratitude to my family, whose unwavering love, encouragement, and belief in me have been the foundation of all my pursuits. Thank you to my parents and brother for always being in my corner, for sharing in my struggles and celebrating my triumphs. Your support means the world. Thank you to my extended family members currently and previously involved in healthcare, whose dedication has inspired my deep admiration for all those working in medicine. Finally, thank you Luke for your boundless patience, understanding, and genuine engagement with my endless musings and reflections on anything and everything, but especially my work.

Dedication

To my parents, for instilling in me a lifelong love of learning.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
Dedication	v
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Motivation	2
1.2 Contributions	2
1.3 Thesis Outline	5
2 Background	6
2.1 Entity Linking	6
2.1.1 Knowledge Bases and Dense Named Entities	7
2.1.2 Visual Entity Linking	8
2.1.3 Visual Mention Disambiguation	9

2.2	Vision-Language Pre-Training	10
2.3	Image Segmentation	11
2.4	Medical Imaging	13
3	Related Work	15
3.1	Text-Guided Mention Disambiguation	15
3.2	V2TEL	16
3.3	Open-Vocabulary Image Segmentation	17
3.4	User-Driven Image Analysis with Visual Prompts	18
4	Proposed Solution	19
4.1	Preliminaries	19
4.2	System Design	20
4.2.1	Segmentation-Supported Spatial Reasoning	22
4.2.2	BERT for Entity Representation Learning	23
4.3	Multi-Task Fusion	25
4.3.1	Combined Entity Linking Loss	26
5	Experiments	29
5.1	Data	29
5.1.1	Medical Image and RoI Data	31
5.1.2	Medical Knowledge Base	32
5.2	Metrics	33
5.3	Baselines	35
5.3.1	Representation Learning Baselines	35
5.3.2	Semantic Object Identification Baselines	36
5.3.3	Generative Entity Linking Baselines	37
5.4	Experimental Environment	38
5.4.1	Hardware and Software	38
5.4.2	Pretrained Model Weights and Hyperparameters	38

6	Results	40
6.1	VEL Performance	40
6.2	Ablations	42
6.2.1	KB Ablation	42
6.2.2	Model Size Ablation	43
6.3	Qualitative Study	44
7	Conclusion	47
	References	50

List of Figures

1.1	Three approaches to VEL using a medical example of a lesion in a lung MRI: Matching the full image to the KB, which returns “lung” and misses the lesion (a), matching only the highlighted RoI to the KB, which ignores context and returns an entity relevant to a different organ (“brain infarction”) (b), and, as used in our solution (VELCRO), combining the RoI and the surrounding image context for more accurate entity linking (c). . .	3
2.1	Entity linking pipeline: Text-based entity linking is performed in two stages: entity recognition and entity linking.	6
2.2	Two-Tower Encoders are composed of two submodels which separately process visual and textual inputs to produce embeddings within a shared latent space for similarity search.	11
2.3	Examples of mention ambiguity in medical images: (a) The same entity (lung) may present differently depending on the imaging modality. (b) A clinically significant entity like a respiratory infection (highlighted in green) may overlap with other body structures (lung, outlined in grey). (c) A relevant entity (lung nodule) may be very small compared to the overall image.	13
4.1	Overview of the VELCRO architecture: We develop our model using a two-tower design paradigm to support entity alignment during training and large-scale inference. Given an image, a bounding box identifying the general location of a visual mention, and an associated entity description, the model learns to refine its geometric and semantic understanding of the visual mention via segmentation mask generation and visual mention-entity alignment.	21

4.2	Example visualization of prompt-image cross-attention in the SAM mask decoder: cross-attention in SAM’s mask decoder shares information derived from image, prompt, and mask embeddings across all components, allowing mask embeddings to accumulate both local spatial and broader image contextual information.	24
4.3	Parallel data flows support multi-task learning through joint training on segmentation and entity alignment tasks via the repurposing of extant model components.	25
5.1	Dataset Statistics	30
5.2	Overview of evaluated model formulations: (a) text-prompt and whole image fusion approaches, as performed by LLMs (b) whole image embedding alignment using VLP models like CLIP (c) region-only linking using traditional closed-vocabulary object detection or segmentation and classification (d) VLP-guided RoI-only linking using visual prompts to crop RoIs for noise removal (e) VELCRO, which uses visual prompts and the whole image to produce contextual RoIs for alignment.	35
6.1	Qualitative Case Studies: Ground-truth segmentation masks and named entities (left), followed by the top 1 retrieved entities and detected entity locations for all evaluated region-aware models (VELCRO, ZegFormer, YOLO-obj, and YOLO-seg). The first row visualizes case (a), while the second row corresponds to case (b).	45

List of Tables

5.1	Examples of UMLS entities used in this work	32
6.1	VEL performance comparison between VELCRO and non-LLM baselines across various visual prompting methods.	41
6.2	Comparison between entity description approaches for VELCRO and both LLMs.	42
6.3	VELCRO performance at varying model sizes.	43

Chapter 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) extends this to settings where visual mentions of entities are linked to their descriptions in a KB [78], which may consist of text [83], images, or some combination thereof [74, 25, 92, 38]. VEL is useful in fields that generate high volumes of imaging data, where image analysis is both time-consuming and labour-intensive on domain experts. In ecology, for example, researchers can use VEL to gain insights into local biodiversity and organism interactions by identifying flora and fauna captured on camera trap videos [59]. Likewise, VEL applications in medicine 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 [34].

User interaction is key to effective linking when performing targeted knowledge acquisition, but the formulation of text-based queries can become burdensome when scenes are complex [90], as is often the case in knowledge-dense domains like medicine. This thesis thus studies a VEL problem in which the user provides a query image and a prompt to guide VEL. Specifically, users in this formulation draw bounding boxes around a region of interest (RoI) within the image to indicate their information intent. The goal is to link the RoI to the corresponding entity in an accompanying textual KB [90]. This “point-and-query” variant of VEL offers a more direct, user-friendly alternative to text-based queries, e.g. “What is the artifact in the top-left corner of the image?”

1.1 Motivation

Suppose that a doctor inspects a lung MRI, finds a suspicious mass in its contents, draws a bounding box around it, and requests information about this RoI from a medical KB. A popular VEL solution [74, 38, 6] is to use a cross-modal embedding model such as CLIP [61]. CLIP produces embeddings that are aligned across different 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 (for a deeper analysis of this model, we direct the reader to Chapter 2). A VEL system can encode the KB entries using CLIP, then encode the query image using CLIP, and finally perform a vector similarity search to retrieve the most similar KB entries. If we use the whole image as the query, this corresponds to the image-level design in Fig. 1.1 (a) [25, 92, 38, 13, 74]. If we only use the bounding box containing the RoI as the query, this corresponds to the mention-level design in Fig. 1.1 (b) [78]. However, as shown in the figure, 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”. Instead, our solution approach, illustrated in Fig. 1.1 (c), is to use information from both the RoI and the surrounding image for entity linking, leading to the correctly linked KB entity of “lesion of the lung”.

1.2 Contributions

This thesis proposes a novel implementation of the design in Fig. 1.1 (c) as our primary technical contribution. The idea is to compute *contextual* cross-modal embeddings that represent the RoI itself and the relevant surrounding information. These contextual embeddings can then be aligned with the KB embeddings to enable entity linking via vector similarity search, as before.

Drawing on the related field of open-vocabulary semantic segmentation, where local image regions are assigned semantic class labels [68], we approach the generation of contextual embeddings in a segmentation-first approach. Image segmentation models, such as the popular Segment Anything Model (SAM) [37], already produce contextual visual embeddings as part of the process of segmenting an image into separate objects, although they lack dense semantic awareness. Additionally, segmentation models produce segmentation masks that precisely delineate objects in an image (or within a bounding box, if provided). A segmentation model can therefore encode the query image and the RoI in a contextual way, and at the same time enable *query refinement* for our VEL problem by drawing a precise boundary around the RoI to replace a bounding box.

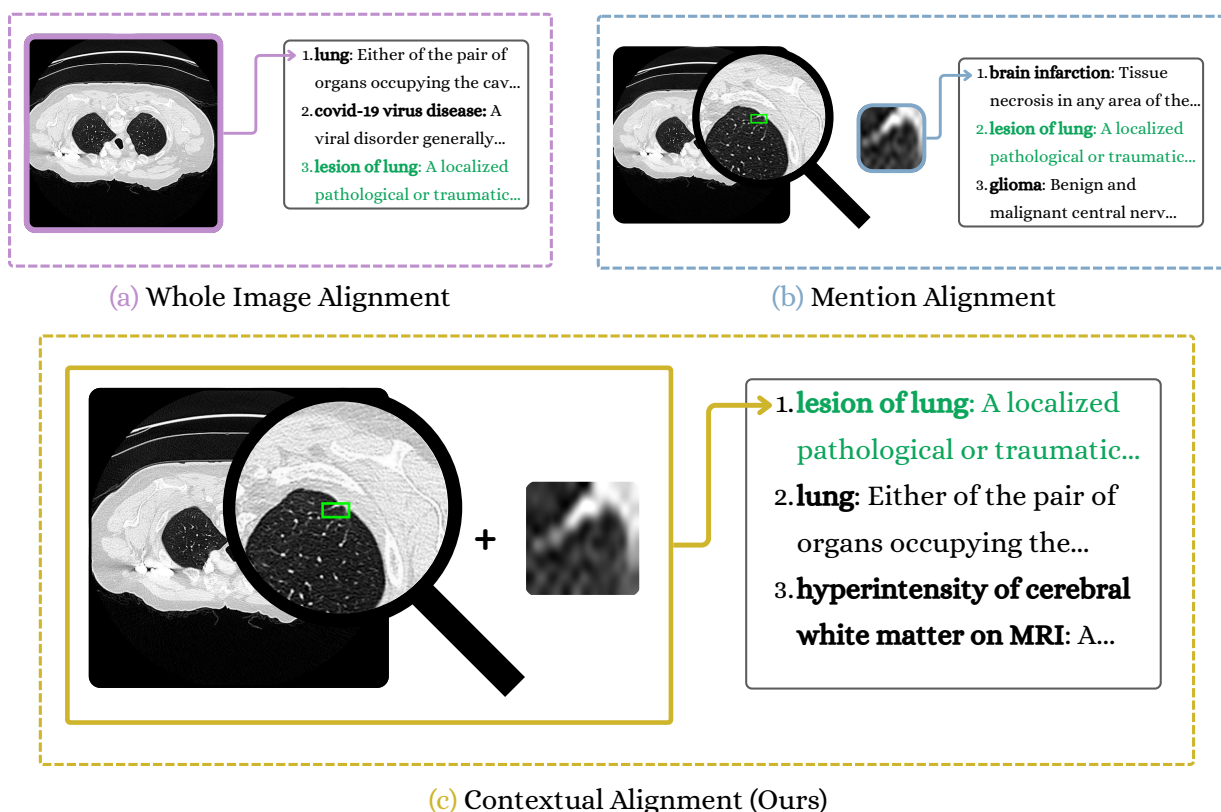


Figure 1.1: **Three approaches to VEL using a medical example of a lesion in a lung MRI:** Matching the full image to the KB, which returns “lung” and misses the lesion (a), matching only the highlighted RoI to the KB, which ignores context and returns an entity relevant to a different organ (“brain infarction”) (b), and, as used in our solution (VELCRO), combining the RoI and the surrounding image context for more accurate entity linking (c).

While numerous CLIP-based semantic segmentation models exist and may be adapted for this task [68, 86], such models either require supporting text—often in the form of image captions—to guide entity identification [12, 49, 82] or perform complete, automated segmentation without user input [68, 86, 18]. Rather than infuse these models with query refinement modules, we instead extend SAM, which natively supports querying, with cross-modal alignment to enable similarity search between contextual RoI and KB embeddings.

Our solution, **Visual Entity Linking with Contrastive Region of Interest Alignment (VELCRO)**, introduces a (CLIP-inspired) contrastive alignment scheme (a type of semantic similarity-oriented representation learning, to be further discussed in Chapter 2) into a segmentation model (SAM). 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 a randomly sampled incorrect KB entity. The training goal is to push together the image and KB embeddings of positive examples and pull apart those of negative examples, leading to an aligned cross-modal representation. To represent images and RoIs contained within them, VELCRO repurposes SAM’s mask embeddings, which accumulate information about the broader image and a specific region for downstream segmentation mask generation (further discussed in Chapter 4), to enable alignment with KB entities. By using the same embedding for both query refinement and entity linking, VELCRO learns to produce entity representations that are simultaneously morphologically and semantically aware. Notably, while VELCRO is a straightforward design based on existing segmentation and cross-modal representation model architectures, we put these components together in a new way via a novel deep learning prediction loss function which combines segmentation performance with contrastive alignment. This allows VELCRO to be trained to accept, refine, and perform entity linking on queried mentions end-to-end, thus unifying all stages of the VEL pipeline.

We experimentally compare VELCRO to the following baselines:

- An image-level CLIP solution (Fig. 1.1 (a)).
- A mention-level CLIP solution (Fig. 1.1 (b)).
- A CLIP-based semantic segmentation model which relies on text-captions to guide segment classification [18] adapted for VEL (Fig. 1.1 (c)).
- A popular CNN-based object detection and segmentation model which has seen previous use in VEL [20] (Fig. 1.1 (c)).

- Interactive entity linking using a large multi-modal model in a zero-shot fashion, by including the KB in the prompt (we tested the open-source LLaVA-Next [45] and the commercial GPT-4o-mini [57]).

Using F1 to measure overall linking accuracy (the harmonic mean of precision and recall) on a medical VEL use case with diverse, clinician-annotated target entities, VELCRO achieves 95.2%, surpassing image and mention-level solutions with performance of 81% and 80.7%, respectively, demonstrating the utility of high-quality, context-aware RoI embeddings. Furthermore, the LLM zero-shot performance was very poor, below 20%. Most notably, VELCRO, which is built on a segmentation backbone, outperforms an alternative contextual RoI embedding approach built using CLIP as a semantic alignment head (83.9%) and two SOTA variants of a classical CNN-based closed-vocabulary region localization model (63.5% and 65.9%). This demonstrates the practical value of adapting segmentation-based approaches to semantic tasks such as VEL.

1.3 Thesis Outline

Chapter 2 introduces relevant background knowledge regarding VEL in both the open and medical domains. Related work is discussed in Chapter 3, and the technical details of the proposed solution are presented in Chapter 4. Chapter 5 describes the dataset used in this work, including the dataset collection and augmentation process, our experimental setup, relevant training methods, evaluation metrics, and models for comparison. Chapter 6 concerns quantitative and qualitative performance analyses, and ablation studies. Finally, Chapter 7 describes insights drawn from the approach and discusses potential directions for future work.

Chapter 2

Background

This work draws on methods from several computing disciplines. Namely, we look at the fusion of database management methods like entity linking and AI-based fine-grained image understanding methods like semantic segmentation to further AI-supported visual knowledge management and information-seeking tasks. Section 2.1 introduces the task of entity linking, knowledge bases, and linking in visual modalities, followed by a discussion of vision-language models in Section 2.2. Section 2.3 discusses the task of semantic segmentation, which is pertinent to our approach, and Section 2.4 concludes with an overview of medical imaging characteristics.

2.1 Entity Linking

Entity linking (EL) is a natural language processing task with roots in information extraction and database management. A subtask of knowledge extraction, it aims to link

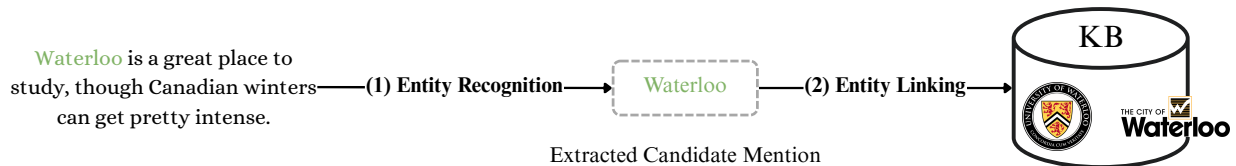


Figure 2.1: **Entity linking pipeline:** Text-based entity linking is performed in two stages: entity recognition and entity linking.

words of interest, or *entity mentions*, within text documents to their structured counterparts, called *named entities*, within an external knowledge base, or KB [66]. While algorithms may differ, linking is generally performed via similarity comparisons between the entity mentions and named entities.

EL is often discussed in close relation to another natural language processing task, Named Entity Recognition (NER), which concerns the identification of mentions within supplied documents prior to KB linking [66]. Some ambiguity surrounding the relationship between NER and EL exists. While NER has traditionally been considered a pre-processing step separate from EL [66], recent works have acknowledged the influence of correct NER or EL performance and shifted to the design of systems which jointly consider both tasks [43]. For the purposes of this thesis, we focus on the latter view of the EL task and refer to this fused approach as *end-to-end EL*.

EL methods (end-to-end and standalone) must contend with several challenges. First, entity mentions within unstructured documents may be referred to in variable ways, as when an organization like “University of Waterloo” is shorted to “UW.” Second, a single entity mention may refer to two separate named entities within a KB. For example, the term “Waterloo” may refer to the Canadian university or city, London district, or the song by the band ABBA. The non-uniform presentation of entity mentions thus requires effective EL systems to robustly adapt to ambiguity during the linking process [66].

Central to overcoming this challenge is the evaluation of the context surrounding an entity mention, which can provide valuable guidance towards its disambiguation. While early EL works performed disambiguation and coreference resolution using methods like topical coherence and feature engineering [66, 44, 99], the rise of deep learning (DL) architectures capable of handling unstructured data in more unsupervised manners through techniques like attention, in which DL models learn to focus on specific, task-relevant parts of input sequences [80], has led to the development of more robust EL systems which require less human oversight. Rather than directly compare words, AI-driven EL systems which produce embedded representations of entity mentions based on their context can better capture their true semantic meaning [5]. As a result, contextual linguistic models like ELMo and BERT have become popular choices for mention embedding representation computation [24, 69, 3].

2.1.1 Knowledge Bases and Dense Named Entities

Knowledge bases are structured repositories of information. They store information regarding specific entities and the relationships between them, supporting the semantic

structuring of data for tasks like document retrieval and analysis [66]. The Wikidata knowledge base, for example, stores structured information about Wikipedia entries, with every database entry being composed of a unique ID, the name of the entity, aliases of the entity, and additional characteristics of the entity, such as in the case of “University of Waterloo,” its location. Some entries may also contain images visually depicting the described entity [84].

The incorporation of this wealth of information during named entity representation generation can be crucial for mention disambiguation during the linking process [25]. For example, a named entity embedding for the University of Waterloo may additionally encode information about its location (Canada). Then, given the sentence “Waterloo is a great place to study, though Canadian winters can get pretty intense.” context-aware models like BERT can use the sentence’s overall meaning to produce a “Waterloo” mention embedding more closely aligned with “Canadian” and “study,” thus allowing it to be correctly linked to the “University of Waterloo” named embedding with similar encoded information, as visualized in Fig. 2.1.

Wikidata [84], YAGO [72], and similar large-scale open-domain databases store information about many disparate entity types, like people, locations, and animals, among others. However, the information within these general-purpose knowledge bases is often insufficient to model more complex, specialized entities and entity relations in knowledge-intensive fields like medicine. Within the medical field, specialized knowledge bases like UMLS and SNOMED CT are widely used to aggregate and model biomedical entities to support research and clinical decision-making [24]. Notably, these prominent KBs are text-based, with UMLS in particular containing dense descriptions of a vast amount of medical concepts [8].

2.1.2 Visual Entity Linking

Visual entity linking (VEL) extends the task of EL into the visual domain. Much like EL, ambiguity exists surrounding the extent of this task, though this ambiguity further extends into the types of modalities comprising VEL inputs and KB constituents. Some works, for example, allow KB entities to be represented by both visual and textual content which contain complimentary information about the described entity [31, 25, 32, 97, 98, 78], while others focus exclusively on text [70]. Many of these VEL systems process pairs of text and visual inputs for joint linking, but some perform linking via only visual inputs [78, 70], making the problem space vast and diverse. Problem spaces with mixed-modality inputs or KB representations are typically referred to as multi-modal entity linking (MEL)

problems. This is in contrast to visual input to textual KB EL, which is *cross-modal* in that a single modality is being mapped to another rather than multiple modalities. Due to a lack of formalization in the literature regarding the relationship between MEL and purely visual input-based VEL, we consider both problem spaces to fall under the larger umbrella of VEL and use the term V2TEL [74] to refer specifically to VEL systems with exclusively visual input documents. While this thesis primarily focuses on the V2TEL problem formulation, we use methods from all formulations of VEL to inform our work.

2.1.3 Visual Mention Disambiguation

VEL is often challenged by mention ambiguity stemming from the visual medium. The same entity within multiple images may display variance according to environmental differences like lighting and camera angle, and the context within which the image was taken [25]. Multiple visual entities may also be present in an image scene, but not all entities may be relevant for the retrieval of that mention’s KB entry [70]. As a result, striking the balance between using additional context to support visual mention disambiguation and context removal to reduce visual noise is a significant research area in VEL [70].

MEL problem formulations often overcome this through the pairing of images with text inputs which explicitly describe target entities, providing additional context towards the interpretation of visual inputs. Indeed, early formulations of MEL considered visual inputs to be supporting documents towards largely text-guided EL [54] and thus focused the challenge of disambiguation on the text modality. While more recent solutions have studied the deeper incorporation of images into the VEL process [25], text remains a significant factor during linking [74].

The content of the text provided to these systems varies depending on the task and domain, with text captions from social media [54] and longer documents from movie reviews having seen use [25]. Because VEL systems are often used to support user-driven tasks like visual question answering (VQA), text inputs may sometimes take the form of questions or “linking queries” indicating the target mentions for linking. We consider these applications to fall under the MEL categorization as well, provided that the text is indeed used during the linking process and the system outputs KB entities. Following Hu et al. [31], we distinguish this from traditional VQA systems which instead produce unstructured responses.

V2TEL systems remove reliance on the text input modality during linking, with the aim of producing robust systems which cannot be misguided by mismatches between other modality inputs. However, this further challenges the disambiguation of visual mentions,

as a lack of supporting text also removes clues towards linking intent which are often expressed within the this modality [31, 74]. This is particularly difficult when performing user-guided tasks like information retrieval [81] and visual question answering over noisy images [90], as the system must also interpret the user’s implicit intent.

2.2 Vision-Language Pre-Training

Central to many DL-based linking systems is the creation of good entity mention and named entity representations for comparison. As introduced in Section 2.1, DL approaches to this task frequently rely on the creation of document-level, mention-level, or entity-level embeddings which capture salient information about their respective contents. When looking towards non-uni-modal problem formulations, however, the generation of good embeddings depends not just on the ability to capture semantically meaningful content within each modality, but a system’s ability to ensure that that content is aligned across modalities.

Early and small-scale approaches perform this via the fusion of uni-modal embeddings into joint multi-modal representations, which is often achieved by summing or concatenating both embeddings together for processing by task-specific model heads [26]. *Vision-language pretraining* (VLP), on the other hand, takes advantage of large-scale image and language datasets to learn robust, aligned representations of images and text for applications to a variety of downstream tasks [48]. The *two-tower encoder* architecture is a popular high-level framework used to facilitate this [26]. As visualized in Fig. 2.2, models which follow this design paradigm use modality-specific encoders to capture per-modality semantic content, then use cosine similarity to compute the degree of alignment between these representations. When working with large-scale data spaces, as when performing VEL between a single image and a large amount of candidate KB entities, this modular design allows for efficient similarity search via decoupled representation generation [26]. Because modality representations are independent of each other, target embeddings for input comparison may be precomputed offline, requiring only the input embeddings to be generated on the fly. Two-tower models are consequently popular choices for retrieval and similarity search-based tasks [26], and often used in VEL systems [74, 38, 6, 31].

Among these models, CLIP [61] has become a foundational backbone for many multi-modal semantic alignment tasks. CLIP’s vision and text towers are composed of a transformer vision encoder, ViT [19], and a transformer text encoder, respectively. The individual embeddings produced by these models are then fed through multi-layer perceptrons (MLPs) to project them into the same embedding space, allowing for direct comparison

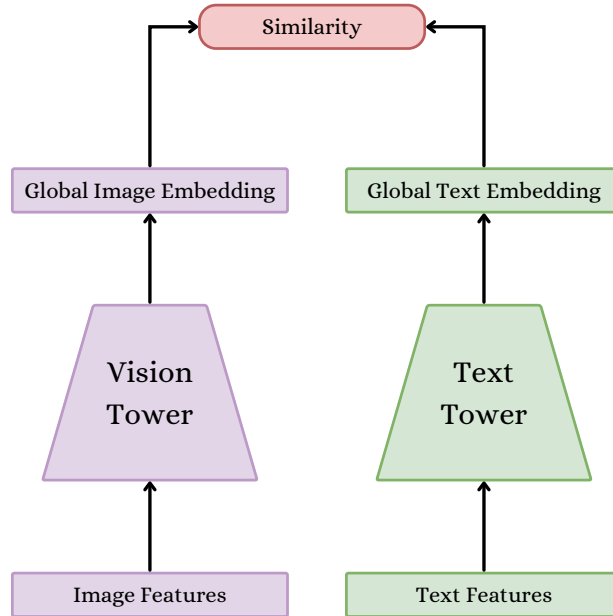


Figure 2.2: **Two-Tower Encoders** are composed of two submodels which separately process visual and textual inputs to produce embeddings within a shared latent space for similarity search.

between them. CLIP’s large-scale pretraining on image and text caption pairs has allowed it to be easily transferred to various semantic understanding tasks.

2.3 Image Segmentation

Within the field of computer vision, image processing tasks are often considered either global or local, depending on the level of visual granularity at which they operate. Global tasks like image classification make predictions based on learned representations of the entire image, while local tasks like object detection make predictions based on specific regions within the larger image. The latter categorization requires vision models to develop finer-grained understandings of visual content, making them capable of distinguishing between multiple semantically distinct visual mentions and, often, able to identify their precise locations within visual scenes [26]. Object detection and image segmentation are the two primary tasks of interest to these local-level vision models. Object detection models predict bounding boxes around objects of interest, while image segmentation models extend

this towards the prediction of pixel-level segmentation masks defining the exact contours around detected objects [91]. Both detection and segmentation models additionally assign semantic labels to detected components, making them regional classification models [26].

These models are well-suited to processing complex scenes. In scenarios requiring acute understanding of image regions, as in VEL, region-processing methods like object detection and segmentation can be used to resolve the challenges of visual noise and visual mention appearance variation which frequently disrupt linking performance [70]. In particular, image segmentation models are capable of learning more precise morphological information about detected objects, often making them superior choices when detailed spatial information is paramount, as in many medical image processing tasks [91, 35].

Image segmentation can be decomposed into three subtasks: semantic segmentation, instance segmentation, and panoptic segmentation. As the name implies, semantic segmentation operates at the class level, proposing one mask per semantic class detected within the image, even if multiple instances of that class exist within the image. Instance segmentation extends this by proposing one mask per distinct entity, allowing for more fine-grained separation of the boundaries around different manifestations of the same class. Finally, panoptic segmentation combines both approaches to produce class-level and instance-level masks [53]. In this work, we are primarily concerned with instance segmentation, but simultaneously consider semantic segmentation as an informing task.

The field of DL segmentation model designs is dominated by encoder-decoder architectures, which are used to perform sequence-sequence transformation tasks like sentence translation [53]. In segmentation, this means the translation of image inputs into pixel-level masks. Encoder-decoder segmentation models split image processing into two stages. The first stage, performed by the encoder, transforms the original image into a compressed latent embedding. This embedding is then passed into the decoder, which upscales the embedding into a pixel-level mask. Traditional encoder-decoder models use convolution layers to down-sample the image, then apply reverse convolutions, called deconvolutions, to upsample it. In semantic segmentation, the final layer of the decoder is used to perform pixel-wise class labeling on the resulting segmented regions [53]. However, such methods are limited to the semantic segmentation domain since pixel label assignment is constrained to the class level [15]. Newer instance and semantic segmentation methods address this through the decoupling of mask generation and classification, often through the use of specialized latent embeddings called query vectors which can be both passed into decoders to produce per-object segmentation masks and simultaneously classified [14].

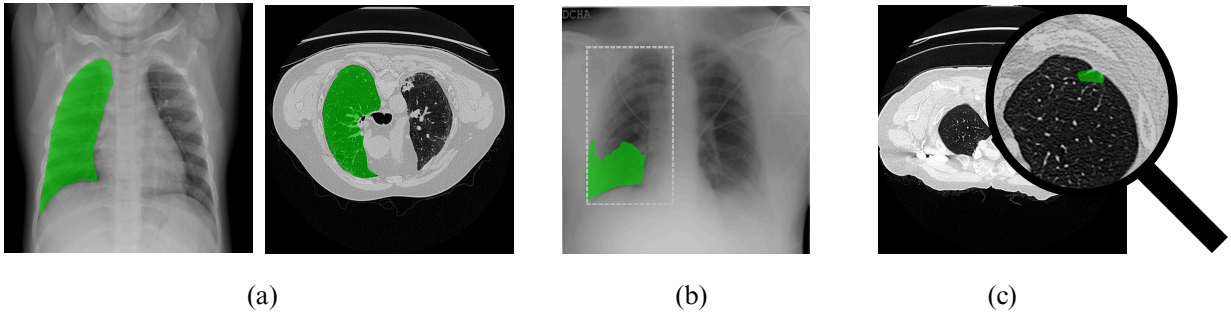


Figure 2.3: **Examples of mention ambiguity in medical images:** (a) The same entity (lung) may present differently depending on the imaging modality. (b) A clinically significant entity like a respiratory infection (highlighted in green) may overlap with other body structures (lung, outlined in grey). (c) A relevant entity (lung nodule) may be very small compared to the overall image.

2.4 Medical Imaging

Medicine in particular is a domain which generates large amounts of text and imaging data, which can be used to assist physicians when practicing case-based medicine, identifying candidates for clinical studies and performing research [34]. Medical images can be particularly illuminating and are frequently used both for patient-facing diagnostic and physician-facing knowledge acquisition tasks [71].

These images are obtained through a variety of technologies. Most prevalent are X-ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), ultrasound, and Positron Emission Tomography (PET) modalities, which each offer unique views of the human body and, similarly, present unique challenges, making automated analysis a non-trivial task [33].

Much like classic open-domain visual ambiguity challenges, entities within medical images also display characteristics like differing entity appearances in different contexts, as when an organ like a lung appears differently depending on the imaging modality used (Fig. 2.3 (a)). Variability between human anatomies, imaging operator methodologies, and differences in disease presentation due to factors like disease severity also contribute to entity variability [33]. Medical images also frequently display overlapping body structures, especially in common modalities like X-rays wherein clinically significant findings within soft-tissue structures, like respiratory infections, overlap with the structures they are found within, as is in the case of lungs (Fig 2.3 (b)). Artifacts of interest may also

be small in size compared to the overall image and displayed blurred boundaries, making direct application of global image analysis methods challenging [75]. Thus, to fully unlock the utility of visual data, methods for imaging analysis must be robust and well-grounded in domain knowledge [40].

Chapter 3

Related Work

3.1 Text-Guided Mention Disambiguation

Works which are closest in spirit to traditional EL approaches use text as the primary document source, but additionally provide visual images as supplementary context to support entity linking. Two methods are used: global image incorporation and local region discovery.

(i) *Global Image Incorporation*: The multi-modal named entity disambiguation task proposed by Moon et. al. in 2018 is an early work which aimed to support text-based NER on social media posts through the fusion of global visual features extracted from a CNN with document-level and finer-grained mention-level text features [54]. CLIP2CLIP [6] performs a similar process using the powerful VLP model CLIP to create global embeddings of image and text captions, which are then fused into a joint embedding for direct similarity comparison with similarly fused image-text KB entity representations. In the medical domain, CLIP has been used to produce fused embeddings from medical images and corresponding descriptions of symptoms for linking with medical diseases [22]. The fusion of coarse-grained visual and textual inputs has also been widely used in the adaptation of generative models towards MEL, particularly in user-guided contexts where textual inputs take the form of unstructured questions or prompts [31, 85, 67, 11]. Unlike the similarity-based systems proposed in prior works, these methods all focus on the use of vision-aware large language models (LLMs), which interpret user-provided visual and textual inputs and generate human-like text, to generate the entity names themselves rather than embeddings. However, images continue to be provided at the coarse-grained level. Alternative approaches use CLIP to individually encode each modality from the inputs

and KB, then compute the pair-wise similarity across all global uni-modal embeddings to obtain a final similarity score [38, 32]. The MEL system proposed by Gan et al., on the other hand, does not create fused representations for KB linking, but rather computes the pairwise alignment across all input and KB modalities to obtain KB entities which satisfy inter-modality agreement [25].

(ii) *Local Region Discovery*: As discussed in Chapter 2, MEL effectiveness has been noted to suffer in the presence of noisy visual inputs, which may contain irrelevant entities [73]. To circumvent this, the incorporation of finer-grained visual features related to specific entities has been explored as a method for more precise mention representation. Intermediate feature maps which capture finer-grained visual features have been extracted from input images and fused with earlier, more global feature maps to produce finer-grained visual representations for MEL [73]. While this can be performed by extracting all late-stage feature maps, several works have taken a more precise approach via the incorporation of object detection models into the EL pipeline [20, 98, 93, 70, 97]. In these systems, object detectors like YOLO [20] and Mask R-CNN [93] are first used to identify the location and class of detected visual mentions. Each mention’s corresponding features can then be extracted and used for similarity linking on their own [20, 98] or concatenated with even finer-grained task-specific features like facial features for more refined embeddings [70]. Alternatively, the class label assigned to each detected object can be compared to or used to replace ambiguous text mentions to improve text-guided linking [93].

3.2 V2TEL

In knowledge-seeking tasks, correct interpretation of user intent is central to correct entity linking [31]. However, without high-quality text inputs to encode this intent, the linking performance of MEL systems degrades, making purely visual linking an important and non-trivial task [74]. An early approach to this challenge circumvented the need for input text through internal caption generation, which was used to identify candidate entity names [78]. Visual noise was reduced via object detection for visual mention detection, followed by image cropping around each proposed box for further visual disentanglement from extraneous visual content. More recently, the Visual Named Entity Linking task was proposed to highlight methods which exclusively rely on visual inputs during the linking process [74]. As with their MEL counterparts, multi-modal models like CLIP have been used to generate image and KB embeddings for similarity comparison [74, 38].

With such image-only problem formulations, however, the challenge of user intent encoding is further pronounced. Without text to explicitly identify entities of interest, works

have primarily approached the challenge through the use of bounding boxes and image cropping to remove unnecessary entities [78, 74]. However, such approaches remove potentially valuable context from the remainder of the scene. Recently, the task of Pixel-Level VEL has been proposed to address the task of VEL within noisy scenes and model user intent through the use of visual prompts in the form of segmentation masks [90]. Like text captions, these visual prompts identify exactly the entity of interest for linking, albeit in a visual medium. The proposed baseline model for this approach uses a preliminary segmentation model to extract visual features corresponding to the segmentation mask, then fuses these features with the global image features for candidate entity generation via an LLM.

3.3 Open-Vocabulary Image Segmentation

Object detectors have been previously used to support visual mention disambiguation in VEL, but these models have traditionally operated on closed vocabularies, meaning that they are restricted to predicting a set number of predefined entities and may struggle to generalize towards novel objects [26]. Furthermore, object detectors fail to capture fine-grained morphological structures and may fail on complex scenes with overlapping entities [91]. The open-vocabulary segmentation task aims to alleviate both challenges by producing fine-grained segmentation masks with detected object labels drawn from an unbounded set of possible classes derived from text prompts at inference time [49]. To facilitate this, these open-vocabulary works shift away from traditional classification methods towards alignment-based approaches, allowing them to better adapt to unseen classes.

VLMs with a wealth of pretrained and generalized semantic knowledge are frequently adapted for this task, transforming segmentation mask label prediction from a classification task to a retrieval task [26]. Most relevant to our work on VEL is the role of language supervision in the segmentation pipeline. Open-vocabulary works typically use either image level [12, 55, 49, 86, 47, 82] or class level [16, 18, 28, 89, 88, 87] labels. It has been observed that more specific text prompts can yield better segmentation results [12], but the formulation of such prompts places additional burdens on users [90]. Furthermore, each of these works is challenged by their dependence on text surface forms during class alignment and do not address scenarios in which user-provided text is vague or captions are poorly formed, making their direct adaptation to VEL tasks challenging due to a lack of entity mention disambiguation. PixelClip stands out as a method which uses learnable class labels for semantic region classification rather than direct alignment with static text

captions, reducing this issue [68], but does not directly inject entity contextual background knowledge to deepen the model’s semantic understanding of learned classes.

3.4 User-Driven Image Analysis with Visual Prompts

Unlike text-referring visual region understanding tasks, visual prompting mechanisms have emerged as a more user-friendly solution to visual reasoning tasks [9]. VipLLaVA [9] and GPT-4RoI [96] are recent LLM-oriented works which aim to infuse visual-aware language models with regional understanding through either the direct overlay of visual prompts like bounding boxes on images, as in VipLLaVA, or the injection of spatial prompts like numerical bounding box coordinates into the text prompt, as in GPT-4RoI.

In segmentation literature, SAM is a popular encoder-decoder solution which employs a variety of visual prompts, such as bounding boxes, points, and scribbles, to guide image segmentation in an interactive, user-driven manner [37]. MedSAM extends this to the medical domain, using primarily bounding box prompts to produce a medical segmentation foundation model [50]. Both works, however, focus solely on class-agnostic segmentation and possess no semantic understanding of segmented objects, making their direct application to semantically-demanding tasks like VEL impractical. Faulkenberry et al. proposed a visual prompt-guided incremental finetuning method for medical semantic segmentation models which allows clinicians to select artifacts of interest within images for their model to further train on [23]. However, this ultimately produces a set of submodels which each model focuses on detecting a single semantic class, challenging its scalability to the large entity sets often present in VEL.

Within VEL literature, user-driven visual prompting mechanisms are sparse, despite the abundance of focused text-based prompting methods available. Frozen segmentation models like SAM and SEEM have recently been used to produce pixel-level masks which can be provided as standalone prompts or supplementary to textual prompts for VEL [90], but the proposed approach does not evaluate the performance of end-to-end training with learnable segmentation and linking methods, which may challenge its generalization to domains with particularly noisy visual characteristics like medical images. There thus exists a significant research gap in the task of end-to-end visual prompt-based VEL, which this work aims to address.

Chapter 4

Proposed Solution

Recall the VEL problem setting described in Chapter 1 and the proposed solution design from Fig. 1.1 (c). The user’s query is an image with a visual mention of an entity¹ highlighted by way of a bounding box. We assume the existence of an accompanying textual KB, and the system output is a KB entity corresponding to the above visual mention. Our solution approach is to (1) compute contextual embeddings that capture information about the RoI as well as its surroundings, (2) which must then be aligned with embeddings of the textual KB entries for entity linking. For (1), we extract mask embeddings from a segmentation model, which also performs a kind of *query refinement* by predicting the precise location of the RoI within its user-supplied bounding box. For (2), we fine-tune a language model to generate KB embeddings and we apply a CLIP-inspired contrastive learning step to align these two embeddings for entity linking. Notably, segmentation and alignment happen simultaneously in a single end-to-end architecture thanks to a novel loss function that combines these two tasks (details in Sec. 4.3.1). Fig. 4.3 summarizes the design of VELCRO, with the details discussed in the remainder of this chapter.

4.1 Preliminaries

This section defines the task of user-guided VEL and relevant notation. Let I be an image containing a visual mention of interest V which the user would like linked. It is implicitly understood that the process of linking may be a preliminary for another downstream task like IR or VQA, but since the proposed solution is only concerned with the act of

¹We use the terms ‘visual mention’ and ‘RoI’ interchangeably.

entity linking we focus our discussion solely on the relevant task. To direct the system to link the correct entity, the user also provides a location prompt P indicating the general location of the candidate V within the larger image. Using the prompt P and image I , the system must (a) correctly identify the exact V of interest and (b) link it to its associated text-based entity entry T in KB , an external knowledge base. Although (a) is often called “visual mention detection” in the associated literature [74], we instead call this task “query refinement” to reflect the act of using a provided P assist in image parsing. 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 final linked entity.

To train a model to perform VEL, a dataset $D = \{(I_i, P_i, M_i^V, T_i) \cdots (I_n, P_n, M_n^V, T_n)\}$ sets must be provided, where M_i^V is a segmentation mask identifying the exact location of V_i within I_i . Then, given (I_i, P_i) , the model must both learn to produce $\hat{M}_i^V \approx M_i^V$ which correctly identify V_i **and** produce representations E_i^V and E_i^T from V_i and T_i which correctly align with each other. During inference, only (I, P) are provided to the model, and $E_i^T \in KB$ are pre-generated offline.

4.2 System Design

To produce mention and entity embeddings for similarity search, we design our system using a two-tower model architecture as introduced in Chapter 2. Recall that this design consists of two uni-modal sub-models which transform features from their respective modality—images or text—into lower-dimensional embeddings summarizing semantic information, which may then be projected into a shared latent space for cross-modal similarity comparisons. This architecture maintains separation between the image and text modalities during processing, which allows image and text embeddings to be generated independently during inference. As a result, static KB entity embeddings can be computed once offline and stored in an index during inference, thus minimizing similarity search latency by requiring only the vision tower to create embeddings in real time [26]. In addition to KB size scalability, the two-tower model architecture also supports a robust post-deployment life cycle by allowing each sub-model to be further finetuned independently. For example, if new imaging data is acquired after deployment, the text tower can be frozen such that the model only needs to focus on learning to align image embeddings with frozen entity embeddings. This prevents the entire model from needing to be continually retrained, saving time and computational resources [26]. Consequently, selecting a two-tower model as the basis of our approach ensures that our model is able to quickly respond to user

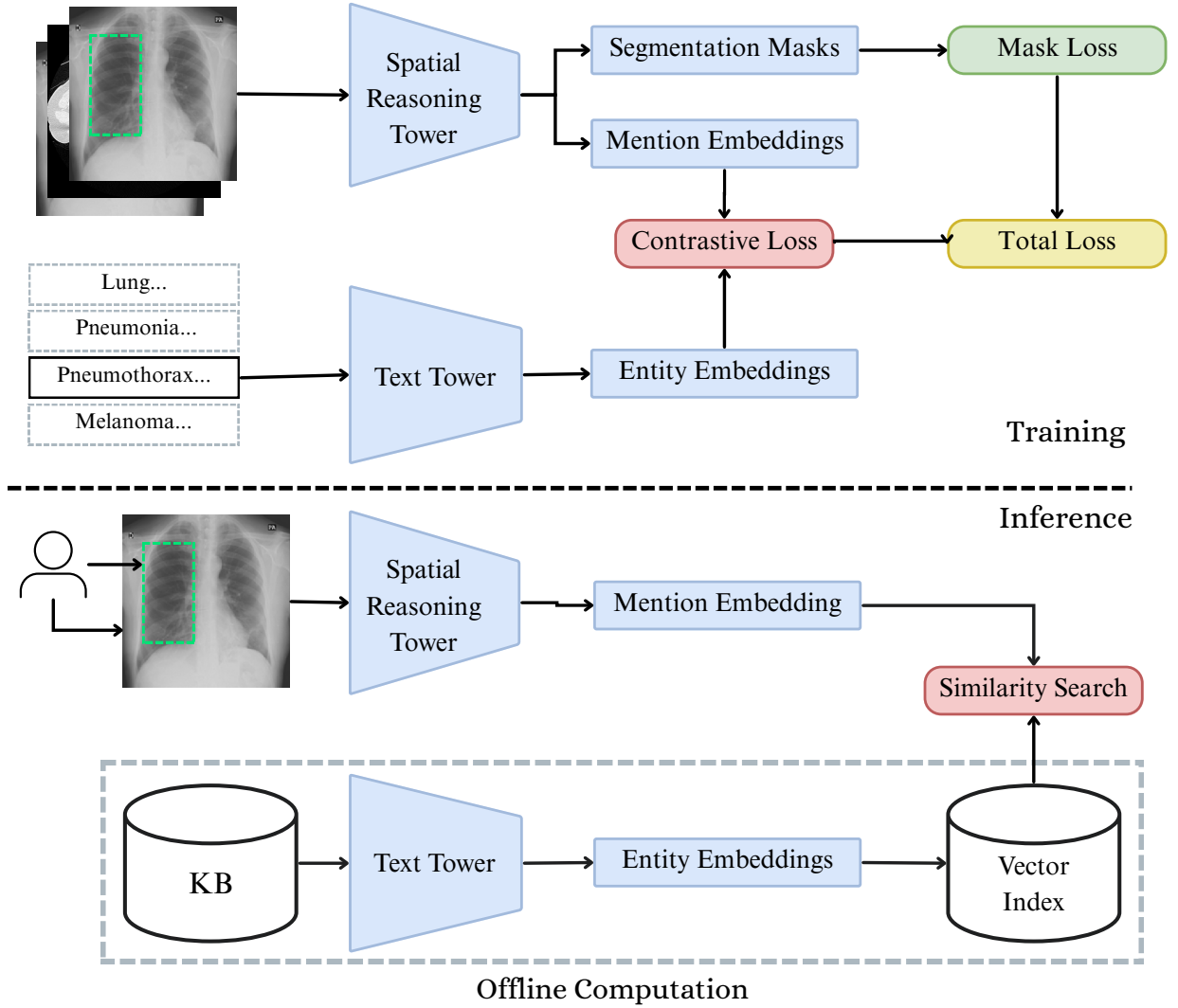


Figure 4.1: **Overview of the VELCRO architecture:** We develop our model using a two-tower design paradigm to support entity alignment during training and large-scale inference. Given an image, a bounding box identifying the general location of a visual mention, and an associated entity description, the model learns to refine its geometric and semantic understanding of the visual mention via segmentation mask generation and visual mention-entity alignment.

requests, scalable to large knowledge bases, and capable of integrating new information after deployment, allowing it to keep pace with advancements in medical knowledge.

Because the overall model must be able to incorporate user intent when analyzing images, our vision sub-model must also accept a user-provided visual prompt alongside the provided image, which is used to guide the selection and emphasis of targeted mention-specific features during the embedding process. Our proposed solution thus adapts the two-tower framework for the task of prompt-guided, interactive VEL by using a promptable, segmentation-based spatial reasoning sub-model (Section 4.2.1), rather than a vision model which only produces global image embeddings, and a text encoding sub-model (Section 4.2.2) to process features from their respective input modalities.

4.2.1 Segmentation-Supported Spatial Reasoning

In our model, the spatial reasoning tower is responsible for processing both a user’s queried image and an additional bounding-box prompt indicating the user’s intent. This sub-model must thus utilize the information contained within each input to produce a singular mention embedding. The query refinement process is essential for the generation of a visual representation which accurately reflects the user’s intent.

Given the complex nature of medical images and medical artifacts, we select a segmentation model to form the backbone of the spatial reasoning tower. This allows us to precisely control the information emphasized in the final embedding via precise RoI extraction. Specifically, we select Segment Anything due to its promptable task formulation and powerful segmentation performance [37]. As previously discussed in Chapter 2, recent segmentation model approaches perform decoupled mask generation and classification using specialized learnable query vectors. However, the number of query vectors per image is typically static, meaning that segmentation models can only detect a fixed number of objects [15]. First proposed in 2023, Segment Anything (SAM) shifts away from this static formulation to perform dynamic, promptable segmentation and has since been applied to numerous open-domain and medical image segmentation tasks [50]. Rather than assume a fixed number of possible objects per image, SAM instead learns to generate segmentations on the fly via diverse user prompts in the form of bounding boxes, points, or “scribbles.” By restricting segmentation only to user-provided prompts, SAM reduces mask prediction overhead and produces segmentation masks which better reflect user intent. Since our task is concerned specifically with user-guided VEL, SAM’s prompt-guided approach is an effective method to achieve targeted entity linking while limiting mask generation only to a select set of visual artifacts of interest, avoiding unnecessary computations. This makes it a strong choice for the instantiation of our spatial reasoning tower.

Mask Decoding: Central to the segmentation process is the mask decoder, which takes the image embedding representing the entire image and the prompt embedding representing the general location of the object of interest, and uses the information contained in each to predict which pixels in the overall image correspond to the object identified by the prompt. A special *mask token* [M] is introduced during this process to accumulate the information necessary for mask prediction, which includes contextual geometric information about the global image and specific spatial information corresponding to the prompt embedding. The mask token is concatenated with the prompt embedding prior to decoding to form a set of “tokens” for processing.

Information within the image and prompt embeddings is transferred via a sequence of stacked self and cross attention blocks in the mask decoder. In self attention, individual tokens within a sequence update their understanding of information from other tokens within the sequence, producing contextual embeddings which better encapsulate information from the entire sequence. Self attention relies on three components: keys (K , with a dimension of d_k), queries (Q), and values (V), where queries represent the current information encoded by a sequence, keys represent, at a high-level, potentially relevant content within the sequence, and values represent the actual information within the sequence [80]. Each token is projected through an MLP to obtain these values, which are then used in the attention equation

$$Attn(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V$$

As visualized in Fig. 4.2, cross attention is performed when two input sequences are passed into the attention block, with attention scores for all tokens in one sequence being computed via information contained within the second sequence. The SAM decoder implements two types of cross attention—image-prompt and prompt-image—to allow for bidirectional information flow. As a result, image embeddings are updated with spatial information contained within the prompt and mask embeddings, while the prompt and mask embeddings are updated with information from the image. This cross-embedding information sharing is crucial for the adaptation of SAM to our V2TEL task.

4.2.2 BERT for Entity Representation Learning

BERT is a transformer-based NLP model proposed by Google in 2019 [17] capable of producing contextual representations of entire input sequences. It has seen much use in

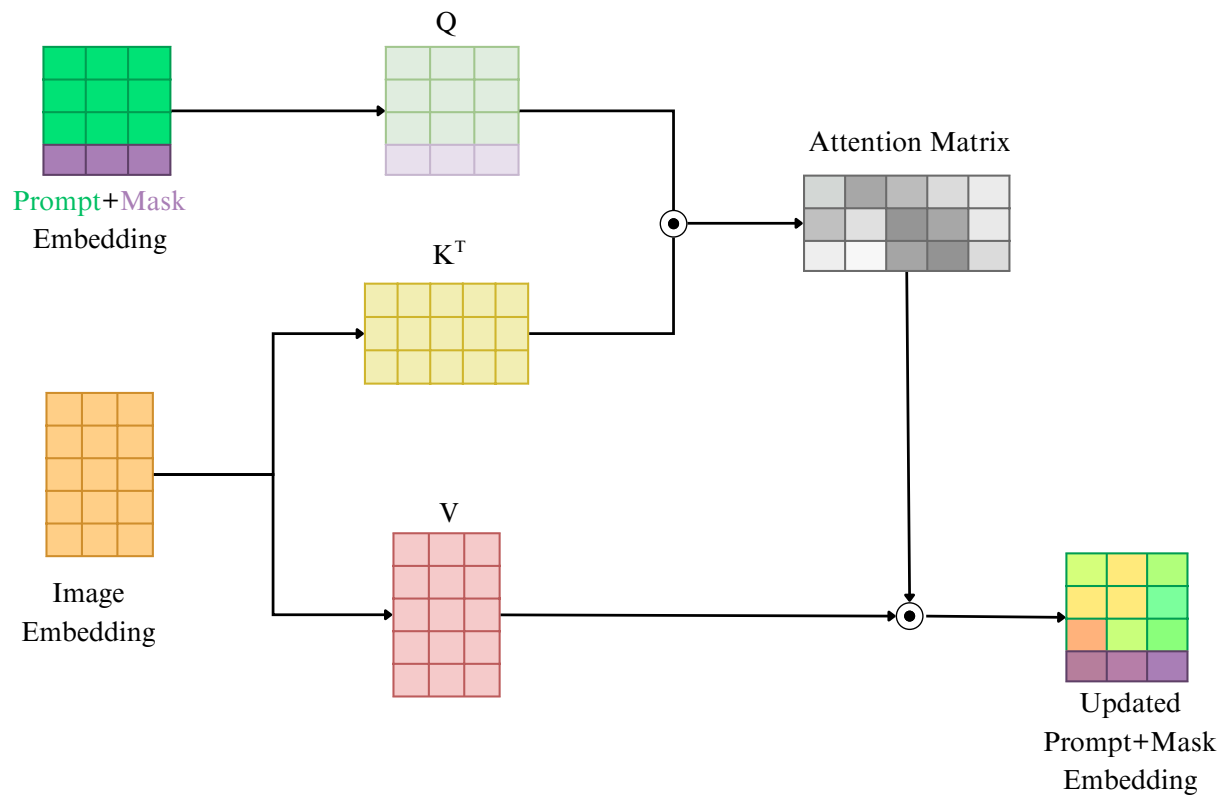


Figure 4.2: **Example visualization of prompt-image cross-attention in the SAM mask decoder:** cross-attention in SAM’s mask decoder shares information derived from image, prompt, and mask embeddings across all components, allowing mask embeddings to accumulate both local spatial and broader image contextual information.

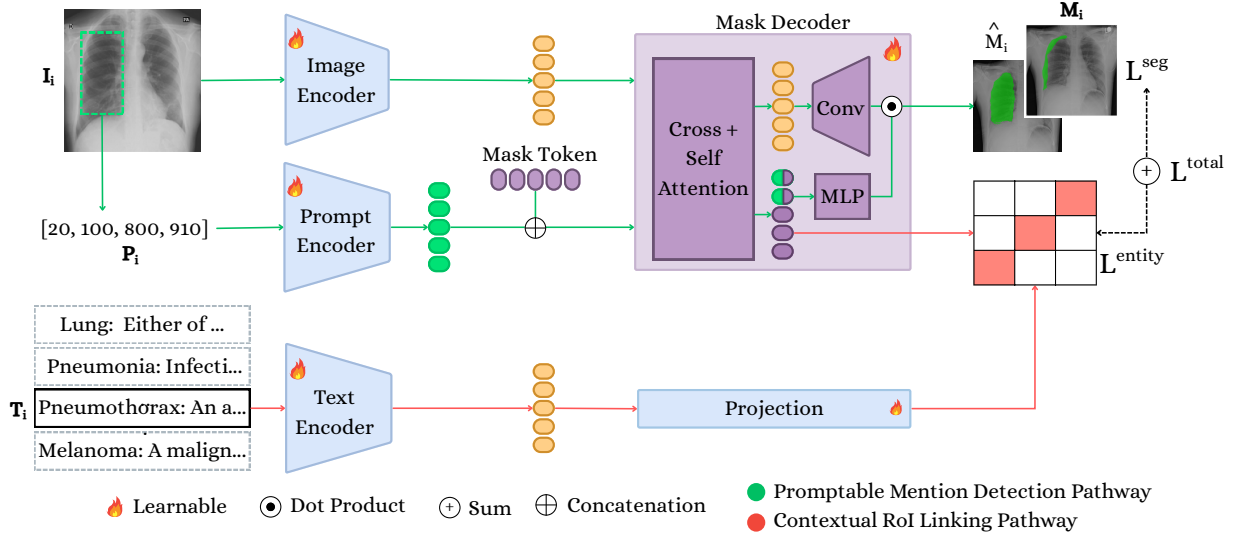


Figure 4.3: **Parallel data flows** support multi-task learning through joint training on segmentation and entity alignment tasks via the repurposing of extant model components.

entity linking-adjacent tasks like named entity recognition [30], making it an ideal candidate for the entity processing portion of our task.

The attention performed by this model is *bidirectional self-attention*, meaning that, for a given input sequence, attention for each token embedding is computed by focusing on the token embeddings to its left and right in the sequence. As in the SAM mask decoder, a special *classification token* embedding [CLS] is prepended to the input sequence passed to the model prior to attention. This embedding attends to all other token embeddings and is thus representative of all information contained within the provided sequence. In our work, we treat this classification token embedding as the final entity representation.

4.3 Multi-Task Fusion

Fig. 4.3 visualizes all components of the VELCRO model and highlights the flow of input data throughout the model, illustrating where important components, such as the mask token, are introduced and how they factor in to VELCRO’s dual-task training objective. We term these data-flows the **Promptable Mention Detection Pathway** and the **Contextual RoI Linking Pathway**. The promptable mention detection pathway illustrates how image and prompt feature information is aggregated in the mask token and used

to refine the selection of salient fine-grained object features via segmentation, while the contextual RoI linking pathway illustrates how that same information is used to produce semantically-aware contextual RoI embeddings for entity embedding alignment in a shared latent space. The dual participation of the mask token in both pathways is key for our methodology, allowing us to use the same latent mention representation for both tasks with minimal direct modifications to the token generation process.

The promptable mention detection pathway is largely representative of typical segmentation finetuning approaches. However, whereas finetuning pipelines often freeze the image encoder, we instead make this learnable to allow for additional domain adaptation. 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 $E_i^V = [M]_i$. This ensures that the mask embedding contains both contextual information about the image and the geometry of the RoI via the decoder’s cross attention.

We use BERT 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 can be pushed closer together, while dissimilar embeddings are pushed farther apart. This enforces semantic meaning through learned proximity, allowing entity retrieval to occur based on the distance between the two embeddings.

4.3.1 Combined Entity Linking Loss

We now discuss the loss functions used to make entity linking with segmentation-oriented embeddings feasible. To train VELCRO to perform both mask generation and entity linking, we combine SAM’s original mask quality loss function [37] (denoted on the right hand side of Fig. 4.3 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. Loss is then computed as a linear combination of focal loss [42] and DICE loss [52], with DICE loss quantifying overall segmentation performance by computing the intersection and overlap of the ground truth and predicted masks. The focal loss, which is a modified variant of cross-entropy (classification) loss, emphasizes the model’s ability to classify pixels as either foreground or background in difficult instances, such as when foreground objects are small relative to the overall image. This is achieved

by weighting the loss contribution of less-frequent class instances, such as foreground pixels in the aforementioned scenario, higher than the contribution of more frequent background pixels. Since medicine often features such class imbalances, this combination of DICE and focal loss ensures that the model focuses on learning to detect important structures regardless of size.

For a single datapoint with ground truth mask m_i , predicted mask $\hat{m}_i$ with N pixels, and

$$p_t = \begin{cases} \sigma(\hat{M}_i) & \text{if } M_i = 1 \\ 1 - \sigma(\hat{M}_i) & \text{if } M_i = 0 \end{cases} \quad (4.1)$$

where σ denotes the sigmoid function and p_t is the probability that a pixel in $\hat{M}_i$ equals a pixel in M_i for all N pixels, DICE and focal loss are defined as follows:

$$L_i^{dice}(\hat{M}_i, M_i) = 1 - \frac{2 \sum_{i=1}^N M_i \hat{M}_i}{\sum_{i=1}^N M_i + \sum_{i=1}^N \hat{M}_i} \quad (4.2)$$

$$L_i^{focal}(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (4.3)$$

Here, γ controls the degree to which infrequent classes dominate the loss. In practice, we set $\gamma = 2$ following Lin et al. [42].

This gives us the final segmentation loss,

$$L_i^{seg} = \lambda_f L_i^{focal} + \lambda_d L_i^{dice} \quad (4.4)$$

where $\lambda_f = 20.0$ and $\lambda_d = 1.0$, as in the original SAM paper [37].

We use cosine similarity to compute the distance between an entity embedding E_j^T and mask embedding E_i^V .

$$S(E_i^V, E_j^T) = \frac{E_i^V \cdot E_j^T}{\|E_i^V\|_2 \|E_j^T\|_2} \quad (4.5)$$

This cosine similarity score is then used in InfoNCE loss [79] as our final entity linking loss. This loss includes a learnable temperature parameter τ , which we initialize to 0.5, to emphasize differences between embeddings. For a batch of training examples consisting of a set of mask-entity pairs, InfoNCE loss treats the true entity pairing for a given mask as a positive sample, and the pairing of that mask with all other entity embeddings as negative samples. Then, the function need only minimize the distance between each positive

sample and maximize the distance of all negative samples for every mask. This encourages robust latent space separation by comparing masks against a variety of potential entity representations, rather than one carefully selected negative sample per positive sample as in other contrastive learning losses such as triplet loss [65]. Drawing negative samples from the batch itself enables training to efficiently and effectively scale to large KBs by striking a balance between single negative comparison and comparison with all possible entities in the KB, making InfoNCE an effective loss for our V2TEL task. The loss function for a single mask embedding over all K entity embeddings in a batch is defined as

$$L_i^{entity} = \log \frac{\exp(S(E_i^V, E_j^T)/\tau)}{\sum_{j=1}^K \exp(S(E_i^V, E_j^T)/\tau)} \quad (4.6)$$

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} :

$$L^{total} = \frac{1}{B} \sum_{i=1}^B (L_i^{entity} + L_i^{seg}) \quad (4.7)$$

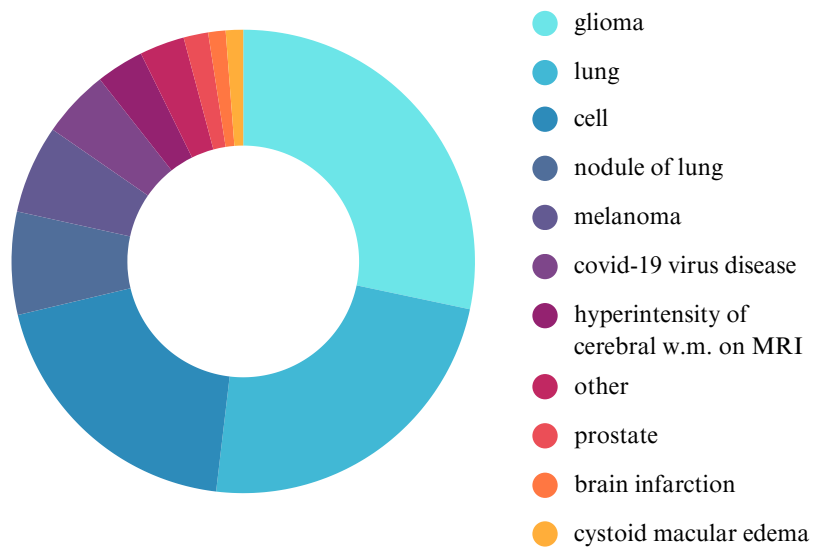
Chapter 5

Experiments

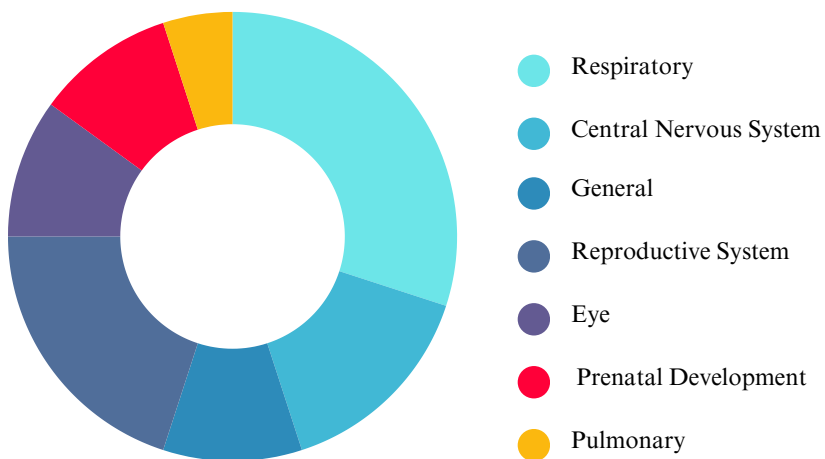
This chapter concerns details regarding experimental design and setup. Dataset curation and pre-processing steps are discussed in Section 5.1, followed by an overview of evaluation metrics in Section 5.2 and baseline models for comparison in Section 5.3. Finally, Section 5.4 describes the hardware and software used during experimentation.

5.1 Data

As we discussed in Sec. 4.2, VELCRO requires a training dataset of the form $D = \{(I_1, P_1, M_1^V, T_1), \dots (I_N, P_N, M_N^V, T_N)\}$, with images I_i , RoIs with their bounding boxes P_i and 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 dataset of the desired form was available, we instead obtained (I_i, P_i, M_i^V) triplets from an existing medical image instance segmentation dataset, then used each segmentation target, in the form of an entity name, described in this dataset to guide T_i retrieval from a medical knowledge base. In Section 5.1.1, the segmentation dataset and image pre-processing methods are described, while Section 5.1.2 details the medical knowledge base and T_i curation process.



(a) **Post-processed Entity Distribution**



(b) **Entity Supercategory Distribution**

Figure 5.1: **Dataset Statistics**

5.1.1 Medical Image and RoI Data

(I_i, P_i, M_i^V) triplets were obtained through an open-source medical segmentation dataset created for medical image segmentation [50]¹. This dataset combines various individual segmentation datasets covering a variety of medical imaging and video modalities, domains, and clinical artifacts for the purpose of universal medical image and video segmentation. We exclude video data, as this is outside the scope of this work. Each sub-dataset is associated with a list of segmentation targets in the form of an entity name.

Since this dataset was originally created for the purpose of universal instance segmentation, datapoints were originally of the form (I_i, M_i) , where M_i refers to a segmentation mask potentially containing multiple entity instances. Raw M_i masks vary in format between sub-datasets, with some assigning each distinct object within an image a unique numerical label regardless of their 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 the normalization of datasets containing easily separable M_i s.

The remaining M_i s were refined by identifying non-contiguous entities within each mask. Each M_i was expanded into a set of $\{M_i^{V_i} \dots M_i^{V_j}\}$ 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 is a set of 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.

¹The segmentation dataset used in this work is continually updated. Some of the entity target names and datapoints used in our experiments may be different from the most up-to-date version made publicly available here: <https://www.codabench.org/competitions/1847/>

Table 5.1: 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 tumour 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.

Fig. 5.1a visualizes dataset entity distributions after expansion. We observe extreme class imbalances, with the dataset largely being dominated by the entities “glioma,” “lung,” and “cell.” We address this with a combination of over and under-sampling during training. Finally, entity supercategory membership according to physical system and body structure are visualized in Fig. 5.1b. We observe that, of the 20 segmentation targets within the original dataset, 30% of those targets were related to the respiratory system, making it the most diverse supercategory represented in the dataset. Conversely, the pulmonary system was the most underrepresented system, making up just 5% of the overall dataset. Nine different imaging modalities and 20 unique medical concepts are represented in the dataset.

5.1.2 Medical Knowledge Base

Although the segmentation dataset contains some entity information in the form of name-based segmentation targets, it does not include entity descriptions. To ensure that the model obtains sufficient semantic information for knowledge-rich entity embedding generation, we thus manually retrieve associated entity descriptions from a known medical ontology to obtain the T_i s in this work.

The Unified Medical Language System (UMLS) Metathesaurus² curated by the United States National Library of Medicine [8] was selected as the target KB for T_i retrieval due to its extensive medical knowledge and popularity in medical EL [7, 69, 27] and image

²https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

analysis works [63]. UMLS comprises over two million data entries relating to 900,000 medical concepts selected from 60 biomedical vocabularies, from which we identify 20 terms matching the segmentation targets described in the image dataset [51]. Some UMLS entities may contain multiple definitions from different biomedical vocabularies. Since we are focused on imbuing the model with deep semantic understanding, we prioritize the selection of more extensive descriptions over concise ones in these cases.

Each resulting T_i consists of a text label and description, which are concatenated with tokens indicating the start of each component (title and description). For example, the term “lung lesion” would be represented as “[TITLE] lesion of the lung [BODY] a localized pathological or traumatic structural change, damage, deformity, or discontinuity of the lung.” The known segmentation target for each decomposed (I_i, P_i, M_i^V) triplet in the post-processed image dataset is then used to guide T_i matching to produce the final (I_i, P_i, M_i^V, T_i) sets for model training. Table 5.1 shows additional examples of UMLS entities represented in the resulting dataset.

5.2 Metrics

We evaluate the performance of VELCRO and the comparison baselines using four metrics: F1, recall, precision, and intersection over union (IoU). F1, recall, and precision measure entity-linking performance, while IoU evaluates segmentation quality for systems that perform segmentation. Although IoU excludes baselines that simply encode images or RoIs using CLIP, we nevertheless include it when appropriate as a valuable insight into linking performance among applicable models. Consequently, we treat IoU as a supporting metric meant to contextualize entity linking performance through the direct analysis of one aspect of linking: mention detection. F1, recall, and precision, which are common across all models, are treated as the primary metrics for model performance comparison.

Precision and recall measure the quality of all links made between visual mentions and entities. They are calculated on a per-entity basis and averaged across all entities to produce the final precision and recall scores for the entire dataset. These metrics rely on four statistics for calculation:

- True Positives (TPs): The number of correctly linked positive visual mention instances for a given entity. This quantifies the creation of true links between matching visual mention-entity pairs.

- True Negatives (TNs): The number of negative visual mention instances which, correctly, were not linked to a given entity. This represents true missing links between mismatched mention-entity pairs.
- False Positives (FPs): The number of negative visual mention instances incorrectly linked to a given entity. This quantifies the creation of false links between mismatched mention-entity pairs.
- False Negatives (FNs): The number of positive visual mention instances which failed to be linked to a given entity. This represents missed links between matching mention-entity pairs.

An entity’s precision measures the number of times visual mentions were correctly linked to that entity (TPs) out of all the times visual mentions were linked to it (TPs and FPs).

$$Precision = \frac{TP}{TP + FP} \quad (5.1)$$

An entity’s recall, on the other hand, is the amount of associated visual mentions that were correctly linked to it (TPs) out of all associated visual mentions within the dataset (TPs and FNs).

$$Recall = \frac{TP}{TP + FN} \quad (5.2)$$

Finally, the harmonic mean of precision and recall gives the F1 measure, which thus encapsulates both metrics and provides an overall understanding of linking performance.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5.3)$$

IoU is a measure of segmentation quality. For the purposes of VEL, it can be used to determine how well the model learned to separate each visual mention from its surrounding context, irrespective of its associated entity. For a given mention’s predicted ($\hat{M}^V$) and ground truth (M^V) masks, it is computed as

$$IoU = \frac{\hat{M}^V \cap M^V}{\hat{M}^V \cup M^V} \quad (5.4)$$

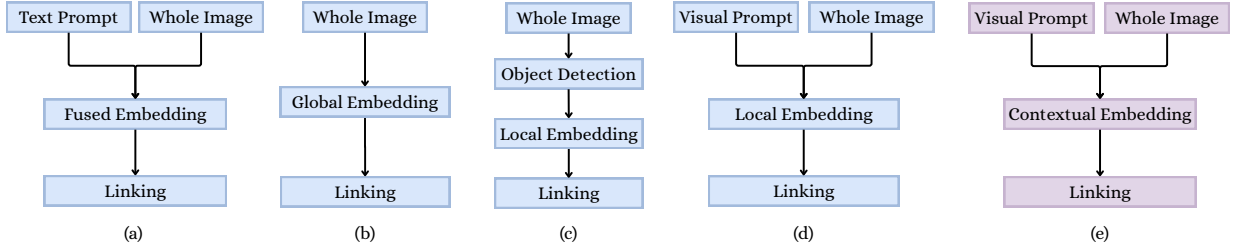


Figure 5.2: **Overview of evaluated model formulations:** (a) text-prompt and whole image fusion approaches, as performed by LLMs (b) whole image embedding alignment using VLP models like CLIP (c) region-only linking using traditional closed-vocabulary object detection or segmentation and classification (d) VLP-guided RoI-only linking using visual prompts to crop RoIs for noise removal (e) VELCRO, which uses visual prompts and the whole image to produce contextual RoIs for alignment.

5.3 Baselines

Since the proposed approach draws from both multi-modal representation learning works and object localization and identification works, VELCRO’s performance is evaluated against representative models from each. Furthermore, the increased use of LLMs, which emphasize user interactivity, for the task of VEL [90, 67, 31, 10] motivates our exploration of zero-shot LLM-based approaches.

5.3.1 Representation Learning Baselines

Multi-modal representation learning models are trained to align the semantic information contained within disparate modalities like images and text, as is performed in CLIP. We thus evaluate two variants of CLIP which employ different strategies for visual mention identification.

The first baseline, denoted CLIP-Full [61], is conceptually illustrated in Fig. 1.1 (a) and described in more technical detail in Fig. 5.2 (b). This is a standard CLIP model finetuned on our dataset. Here, the query is the full image, which is embedded and projected into a shared image-text embedding space. No RoI indicators are provided, and no masked RoI embedding generation is performed.

The second baseline, CLIP-Crop [74], is conceptually illustrated in Fig. 1.1 (b) and detailed in Fig. 5.2 (d). This is a standard (fine-tuned) CLIP model similar to CLIP-Full, but the image is cropped around the user-provided prompt prior to embedding generation.

The query seen by the model thus only contains the information of the visual mention itself and any context immediately surrounding it.

5.3.2 Semantic Object Identification Baselines

Unlike pure representation-learning models like CLIP, which perform entity linking via similarity search on learned embeddings, it is also possible to treat linking as a classification task applied to relevant mentions within a provided document. Given a whole image query, the act of query refinement can be transformed into a pure object localization task, while entity linking is translated into a semantic labelling task applied to the resulting detected candidate mentions. Unlike cross-modal alignment-based approaches, in which the KB entities are directly encoded and searched, the resulting labels in this formulation are instead used as intermediate keys for KB entity lookup. Consequently, good visual mention feature learning is entirely controlled by visual modality information. To evaluate this alternative formulation of VEL, we investigate the use of a well-known closed-vocabulary object localization model with both object detection and segmentation heads as an alternative solution.

YOLO, first introduced in 2016 [62] was originally proposed as an end-to-end convolutional neural network (CNN)-based object detection model. It detects objects by dividing an image into grid cells which both predict bounding box coordinates encapsulating the potential object at each cell’s centre and class probabilities for each predicted bounding box. Much like VELCRO, YOLO achieves multi-task performance via a composite loss function combining classification loss and detection loss. While many YOLO variants have been proposed since its introduction [64], we use YOLOv12, the most recent and highest-performing architectural variant, in our experiments [76]. We apply both YOLO for object detection and YOLO for instance segmentation, in which YOLO’s bounding box prediction head is removed and feature maps from upscaled convolutions are instead passed into segmentation head, which predicts candidate instance-level segmentation masks. Both of these methods for object localization are encapsulated in Fig. 5.2 (c). To integrate user-provided visual prompts in real-world, targeted VEL scenarios, IoU-based one-to-many matching between the generated bounding box or segmentation mask and the provided prompt is performed. In practice, this pipeline performs query refinement as a final post-processing step before linking, rather than an early processing step intended to guide mention representation generation.

We also evaluate performance on an open-vocabulary semantic segmentation model, ZegFormer, which combines MaskFormer, a fixed query-based segmentation model [15],

with a CLIP classification head to transform the task of mask classification to a vision-language alignment problem [18]. However, while we directly use the same mask token (or query vector, in MaskFormer) for semantic alignment and class-agnostic mask generation, they instead project mask tokens into two separate spaces, thus producing two separate, single-task oriented embeddings rather than one dual-task embedding. Furthermore, they also 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 use mask embeddings as our only visual signal during training and inference. This approach is represented by Fig. 1.1 (c) and Fig 5.2 (d). As with YOLO, one-to-many matching between generated segments and user prompts may be performed at inference time in promptable scenarios.

5.3.3 Generative Entity Linking Baselines

Inspired by recent generative approaches to the VEL task [90, 67, 31, 10], we explore the zero-shot application of LLMs with visual reasoning capabilities for user-driven VEL. Specifically, we compare performance between LLaVA-Next [45], a powerful open-source model, and GPT-4o-mini, a member of the GPT-4 class of models from OpenAI [57]. This problem formulation is represented by Fig. 5.2 (a). In general, these models feature vision-processing modules which transform visual inputs into a form that can be understood by the underlying text-based LLM backbone. LLaVA-Next, for example, uses a frozen CLIP image encoder to extract image features, which are then transformed into language tokens via an MLP-based projector. This allows visual inputs to be merged with tokenized text prompts for final processing via the LLM backbone [46, 45].

We approach LLM evaluations as contextual visual reasoning tasks, where the LLM must analyze the content of provided images and make decisions with the help of user-supplied context. To align 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 entity from the KB which best describes the visual mention in the bounding box. Minor prompt engineering was performed to ensure that both LLMs were able to appropriately analyze the provided images. For LLaVA-Next, for example, it was observed that 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.

5.4 Experimental Environment

5.4.1 Hardware and Software

Experiments were primarily conducted using two RTX 6000 Ada Generation GPUs with 32 GB of RAM and 8 CPU cores. VELCRO was trained and evaluated using Pytorch 2.5.1 [4] and PyTorch-Lightning 2.4.0 [21], a deep learning framework which wraps PyTorch for streamlined model deployment and efficient distributed training across multiple GPUs.

We use an 80-10-10 train-test-validation split for our dataset. We perform train, test, and validation dataset splitting at the case level to prevent data leakage during training. The train, test, and validation datasets comprise 435,208, 52,701, and 49,920 unique datapoints from 41,481, 5,187, and 5,186 distinct cases, respectively. Thus, while we are unable to obtain exactly balanced datasets through case-based splitting, we are able to achieve a close approximation. To manage dataset imbalance in the training data, as discussed in Section 5.1.1, weighted random sampling was performed to undersample datapoints from overrepresented entities and oversample datapoints from underrepresented entities. We achieved this by first assigning a weight to each datapoint according to its entity membership. The list of datapoints and their associated weights were then provided to PyTorch’s WeightedRandomSampler, which randomly selects the next batch of datapoints from the dataset to be passed to the model during training according to the provided weights. This ensures equal class distribution during training. All random operation seeds, including dataset splitting and sampling, were set to 42 for all experiments³.

5.4.2 Pretrained Model Weights and Hyperparameters

VELCRO is initialized using the pretrained weights from SAM-base with 16 patches⁴ and BERT-base⁵. 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 patches⁶, while YOLOv12 was initialized using the pretrained weights from the largest YOLOv12 variant, YOLOv12-X. This variant features 59.1M trainable parameters for object detection and 64.5M trainable parameters

³All code, including experimental scripts and model and dataset configurations, are publicly available at <https://github.com/carbonkat/VELCRO.git>

⁴<https://huggingface.co/facebook/sam-vit-base>

⁵<https://huggingface.co/google-bert/bert-base-uncased>

⁶<https://huggingface.co/openai/clip-vit-base-patch16>

for segmentation [77]⁷. To process images and text, LLaVA-Next is composed of 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 parameters⁸. Finally, we perform inference on GPT-4o-mini using OpenAI’s API. Although GPT-4o-mini’s size is not publicly available, its status as a “small language model,” according to OpenAI [56], We set the LLM temperature parameter, which controls the amount of randomness or “creativity” used when generating model outputs, to 0.0. This mitigates the influence of LLM non-determinism on output reproducibility, which may cause these models to output different responses to the same input on consecutive queries.

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 use a learning rate of $1e^{-5}$, which was found to most effectively balance training convergence with overall performance. Consequently, both YOLOv12 models and ZegFormer were similarly trained with a learning rate of $1e^{-5}$ as well. A batch size of 32 was used for all three models, following the default hyperparameters of their respective implementations. To simulate a similar loss function to VELCRO and emphasize the importance of both accurate mention localization and entity identification, classification and object detection loss were weighted equally. All finetuned models used an AdamW optimizer without learning rate decay and a learning rate on plateau scheduler to decrease the learning rate by $1e - 1$ when a monitored metric, per-epoch validation loss, stops improving.

Finally, as discussed in Sec. 4.3.1, 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 unassociated (or negative) embeddings of the opposite modality are considered dissimilar when manipulating embedding distances in latent space. If two visual mentions corresponding to the same entity were present within a training batch, duplicate entities were removed such that each entity embedding within the batch was unique. This circumvented issues in which visual mentions could be both pushed towards and pulled away from the same entity. During inference, the full set of unique candidate entities was provided for fair evaluation.

⁷<https://github.com/sunsmarterjie/yolov12/tree/main>

⁸<https://huggingface.co/llava-hf/llama3-llava-next-8b-hf>

Chapter 6

Results

This section presents our experimental results. Sec. 6.1 begins with a quantitative performance comparison of VELCRO against the baselines described in Sec. 5.3. Sec. 6.2 describes ablation studies on KB data and model size, while Sec. 6.3 concludes with two case studies qualitatively analyzing VELCRO performance.

6.1 VEL Performance

Table 6.1 shows the F1, precision, recall and IoU of VELCRO and all baselines. We label the methods according to the nature of the provided prompt (full image, bounding box) used for *pre-processing* visual inputs (meaning whether or not a prompt is seen by a given model before a semantic label is attached to a visual mention) and the entity linking formulation: CLIP-based models rely solely on learned semantic embeddings to perform linking ($\mathcal{S}$), while YOLO’s detection-based approaches rely on object detection and classification ($\mathcal{O}$). Generative models like LLaVA and GPT-4 autoregressively predict entity types ($\mathcal{G}$). Bounding boxes are used in CLIP-Crop and for both LLMs, while CLIP-Full, YOLO-obj, and YOLO-seg do not use any explicit prompting. Finally, VELCRO and ZegFormer are classified as $\mathcal{S} + \mathcal{O}$ mixed-method models. Though the open-vocabulary semantic segmentation formulation typically treats text captions as segmentation prompts, we instead provide KB entities to ZegFormer’s classification head and thus treat this as a full-image prompting model.

Table 6.1: VEL performance comparison between VELCRO and non-LLM baselines across various visual prompting methods.

Prompt	Model	Type			F1	Precision	Recall	IoU
		$\mathcal{S}$	$\mathcal{O}$	$\mathcal{G}$				
Full Image	CLIP-Full	✓	✗	✗	0.810	0.814	0.839	-
Bounding Box	CLIP-Crop	✓	✗	✗	0.807	0.795	0.829	-
Bounding Box	LLaVA-Next	✗	✗	✓	0.113	0.130	0.173	-
Bounding Box	GPT-4o-mini	✗	✗	✓	0.179	0.270	0.313	-
Full Image	YOLOv12-obj	✗	✓	✓	0.635	0.665	0.724	0.539
Full Image	YOLOv12-seg	✗	✓	✗	0.659	0.702	0.7676	0.554
Full Image	ZegFormer	✓	✓	✗	0.839	0.858	0.887	0.780
Bounding Box	VELCRO	✓	✓	✗	0.952	0.934	0.981	0.806

We first observe that VELCRO outperformed all other models in entity linking metrics, with an F1 of 95.2%. The $\mathcal{S}$ baselines CLIP-Crop and CLIP-Full, which make no use of RoI embeddings, show poor performance in comparison to ZegFormer and VELCRO, with CLIP-Crop performing slightly worse than CLIP-Full with an F1 of 80.7%. The similar performance displayed by both models suggests that only a minor performance improvement can be achieved by providing the whole image as context for entity linking over only the most relevant portion of the image. The results obtained by ZegFormer and VELCRO further indicate that an alignment scheme that leverages targeted RoI information without removing whole-image awareness is more conducive to the task of VEL than either strategy separately, though VELCRO outperforms ZegFormer in F1 by 12.6%. This suggests that the application of open-domain semantic models like CLIP as classification heads to specialized domains like medicine requires additional text encoder training.

Among $\mathcal{O}$ models, VELCRO and ZegFormer display comparable segmentation performance and outperform both YOLO models by a significant margin. In particular, it was observed that 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 entities with small mention regions, like cells. We note that this behaviour can be adjusted by tuning hyperparameters like detected object confidence to increase the amount of predicted masks, but this also leads to increased incidence of false positives.

Finally, both LLMs exhibited poor zero-shot performance. Although GPT-4o-mini achieved overall better results with an F1 of 17.9%, this still fell considerably short of all non-LLM baselines. Off-the-shelf LLM application to the task of VEL thus fails to

Table 6.2: Comparison between entity description approaches for VELCRO and both 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	-
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	-

efficiently leverage their reasoning capabilities even in condensed linking scenarios where only a few candidate entities exist. While zero-shot LLM use has become a popular method to take advantage of their large reserves of accumulated knowledge without additional training costs, exercising caution in high-stakes scenarios like medicine remains warranted. More direct methods of adaptation to the VEL task, effective prompting mechanisms such as few-shot prompting, and, in the case of domain-specific applications, domain knowledge via finetuning, are needed.

6.2 Ablations

6.2.1 KB Ablation

To study the effect of information-dense KB entity representations on entity linking in VELCRO and in the LLMs, which displayed poor initial task performance, linking performance was further evaluated with shortened entity descriptions. Descriptions were simplified such that new entity descriptions took the form ‘‘An image of a {entity}’’.

As shown in Table 6.2, overall linking performance was relatively similar between the two approaches, with full-caption VELCRO yielding slightly higher performance over its shortened counterpart. However, VELCRO’s negligible decrease in query refinement performance further suggests that variances in entity captioning were minimally influential. For $\mathcal{G}$ models, the KB provided in the model’s prompt was only composed of entity names. LLaVA-Next demonstrated overall improved linking performance under these simplified conditions with a particularly large recall increase from 17.3% to 31.2%, meaning that it

Table 6.3: VELCRO performance at varying model sizes.

Model	Components	Param#	F1	Precision	Recall	IoU
VELCRO _{B/V}	SAM _{B/V} + BERT _B	113M	0.844	0.812	0.945	0.691
VELCRO _{B/T}	SAM _B + Text Proj.	119M	0.937	0.914	0.985	0.793
VELCRO _B	SAM _B + BERT _B	203M	0.952	0.934	0.981	0.806
VELCRO _T	SAM _B + BERT _L	429M	0.950	0.933	0.978	0.783
VELCRO _V	SAM _L + BERT _B	422M	0.923	0.932	0.944	0.801

likely suffered from over-contextualization during initial full-KB inference. GPT-4o-mini was less sensitive to KB variances and displayed relatively consistent performance between settings, but still only achieved a maximum F1 of 17.9%. For competitive zero-shot performance, larger, more complex models would likely be needed.

These results are particularly interesting in light of recent findings in related VEL work on images of celebrities, which have demonstrated the efficacy of dense entity descriptions on improved mention disambiguation and linking [70]. Given our seemingly contradictory results, this benefit may be modulated by VEL domain or entity distributions within our dataset. For example, because some entities only appear on specific imaging modalities with distinct visual characteristics, like optic cups and disks on coloured fundus images or breast masses on mammograms, information from the images and the subtle variations in words alone (e.g. optic *cup* versus optic *disk*) were likely sufficient to produce well-separated latent spaces for our data. The curation of additional data and design of experiments to further explore this are left as future work, to be discussed in Chapter 7.

6.2.2 Model Size Ablation

VELCRO’s minimal query refinement performance drop under ablated captions prompted further evaluation into the effect of different submodel component sizes on overall task performance. We hypothesized that the predictive power of larger, more complex individual components would enhance the overall model’s performance in the associated task. E.g. when VELCRO’s vision model size was increased, query refinement performance would increase, while text model size increases would yield improved linking performance. We compared our base model VELCRO_B with 203M parameters to VELCRO_T, a variant which 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 6.3 indicate that larger model sizes did not necessarily yield increased performance. Nearly all models were able to achieve comparable performance in both linking and segmentation, suggesting that VELCRO_B is sufficient for the task on this dataset. We further dove into this by freezing different parts of the model architecture to explore the impact of various model components on performance. In VELCRO_{B/V}, SAM’s image encoder, ViT [19], was frozen, resulting in a smaller model with 113M trainable parameters, while in VELCRO_{B/T} BERT was frozen such that only the text projection layer was learnable, resulting in a smaller model with 119M parameters. We observed that performance was only mildly affected by freezing BERT, as evidenced by VELCRO_{B/T}’s F1 of 93.7% in comparison to VELCRO_B’s 95.2%, but that freezing SAM’s image encoder yielded a 2.401% performance drop in linking F1 and a 13.6% drop in segmentation. This implies that the learning of good visual representations is the leading factor in region localization and, consequently, linking.

6.3 Qualitative Study

The variation in performance between methods that leverage generated masks (YOLOv12-obj, YOLOv12-seg, ZegFormer, and VELCRO) prompts us to perform a qualitative analysis across several cases evaluated by all three models. Fig. 6.1 shows two examples. In case (a), 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, indicating that imprecise visual mention localization likely contributed to this failure. While YOLO-seg does predict the correct entity, its generated segmentation mask is also poor, suggesting that this model may in fact have failed to learn the morphological distinction between the smaller optic cup and larger optic nerve. Only VELCRO performs both high-quality object localization *and* correct entity linking, demonstrating harmony between RoI refinement and semantic alignment.

Case (b) visualizes VELCRO’s superior linking performance even in scenarios with comparable detected RoIs. In this instance, despite all models producing similar segmentation masks, only VELCRO was able to correctly link 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 partial understanding of entity *location*, if not entity semantics.

As demonstrated in these case studies, segmentation masks support the interpretation

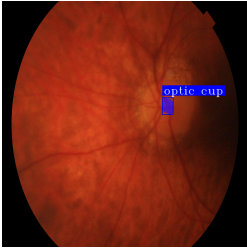
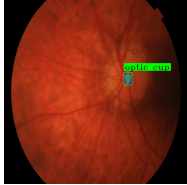
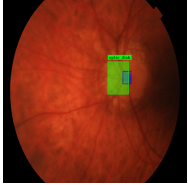
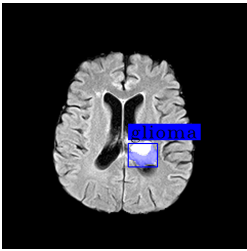
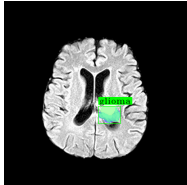
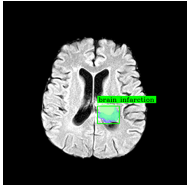
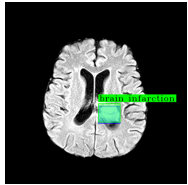
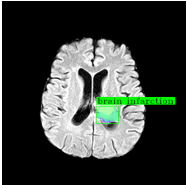
Ground Truth	VELCRO	ZegFormer	YOLO-obj	YOLO-seg
	 Optic Cup A double walled invagination of the op...	 Optic Disk The portion of the optic nerve seen in the fund...	 Optic Disk The portion of the optic nerve seen in the fund...	 Optic Cup A double walled invagination of the op...
	 Glioma Benign and malignant central nervous syste...	 Brain Infarction Tissue necrosis in any area of the brain, incl...	 Brain Infarction Tissue necrosis in any area of the brain, incl...	 Brain Infarction Tissue necrosis in any area of the brain, incl...

Figure 6.1: **Qualitative Case Studies:** Ground-truth segmentation masks and named entities (left), followed by the top 1 retrieved entities and detected entity locations for all evaluated region-aware models (VELCRO, ZegFormer, YOLO-obj, and YOLO-seg). The first row visualizes case (a), while the second row corresponds to case (b).

of model results by allowing users to visualize which features the model determined belonged to queried objects. This explains to users why certain decisions may have been made, and ultimately why particular entities were retrieved. In case (a), for example, Zegformer and YOLO-obj’s segmentation masks allow users to trace the incorrect entity predictions made by these models to their selected visual features, thus revealing where failures during the linking process likely occurred. For end users like radiologists and clinicians, segmentation mask visualization can prompt critical thinking about the correctness of model outputs by allowing users to review how well the model’s selected features aligned with their information need. A segmentation mask that extends far beyond the bounds of the user’s supplied visual prompt, for example, could indicate to the user that the model may not have understood their query and that they should exercise more caution when reviewing its results.

Model explainability is particularly important in high-stakes domains like medicine, where model failures can lead to missed or incorrect diagnoses, prevent medical professionals from obtaining accurate medical information in a timely manner, and result in poorer

patient outcomes [2]. For tools like VELCRO, which can be used to augment a variety of image analysis tasks influential to the patient care process, appropriately calibrating user trust in its outputs through explainability is essential. To this end, we note that segmentation masks alone may not be sufficient to explain all model outputs. As shown in case (b), poor linking can still result from good object localization. If users were to exclusively rely on mask quality as a proxy for model trustworthiness, less-experienced end users still in the process of developing their clinical knowledge may fail to recognize these more subtle failures. Thus, additional measures of model uncertainty, or the model’s confidence in its decision, and more sophisticated explanation techniques may be needed to ensure safe model use in practice.

Chapter 7

Conclusion

We presented VELCRO, a new model for visual entity linking with user-supplied bounding boxes as queries. Using a segmentation model (SAM) as a backbone, we adapted this pure segmentation model for visual entity linking by directly aligning mask embeddings containing information about an entity of interest to text candidates in a knowledge base. VELCRO’s ability to perform both segmentation and entity linking negates the need for dual-model pipelines and collapses the two primary requirements of VEL—mention detection and linking—into a single, unified model capable of generating high-quality, semantically-oriented segmentation masks.

Evaluation against alternative VEL methods based on the popular multi-modal model CLIP revealed VELCRO’s improved performance in entity linking against standard fine-tuning approaches and VEL-specific CLIP models using RoI embeddings. Our results indicate that exploring segmentation model extension to semantically oriented tasks is a promising research avenue for fine-grained image understanding tasks such as VEL.

Based on these findings, many avenues of future work are open for exploration. Our experiments were conducted on a small subset of the information in the UMLS Metathesaurus. Expanding our image and entity datasets would allow us to evaluate the scalability of our approach and further explore VELCRO’s stability under more challenging task conditions. This would be especially valuable for better analyzing the influence of dense KB descriptions on linking performance (discussed in Chapter 6), as our results suggested a heavier reliance on image contextual data than deep text-based semantic knowledge. Introducing more closely-related entities with subtle variances best captured through free-form descriptions would allow future work to better evaluate the potential benefit of entity descriptions in more cluttered latent spaces.

As discussed in Chapter 5, the image dataset used in this work is a subset of a pre-existing medical image segmentation dataset and does not contain all modalities (such as videos) and entities featured in the parent dataset. Future work can thus explore the pre-processing and incorporation of unused parent dataset components to create even more diverse test-beds for medical VEL application. In particular, the use of prompting-based VEL applications to continuous sequential data like surgical or internal exploratory videos could help clinicians track particular artifacts throughout the video’s duration, drive targeted video retrieval, and assist in interactive video question answering or analysis, and would be an interesting path of future inquiry.

All three splits (training, testing, and validation) of our dataset also feature marked class imbalances, with a heavy preference towards common structural entities like lungs in chest imaging views, entities which occur in multiple slices in 3-D scans, and unique instantiations of the same entity within the same 2-D image, as with cells and paired organs like lungs. Though reflective of real-life entity distributions in some scenarios (lungs will almost always appear in chest X-rays, for example, while an infection like pneumonia will not), balanced validation and testing data splits, the inclusion of additional data augmentation strategies, and dataset expansion to images with diverse entity presentations should be explored to reduce potential frequency and appearance-based biases during model development.

Furthermore, the backbones (CLIP, YOLO, ZegFormer and SAM) used in our experiments are pretrained on different resolutions of imaging data, with SAM requiring the highest and consequently learning from more detailed features. The effect of image resolution on linking performance is thus another avenue of future work. Exploring VELCRO adaptations to lower-resolution imaging data to increase computational efficiency is particularly interesting, which could enhance model accessibility in resource-constrained medical centres.

Finally, all of our experiments focus on the entity linking subtask of image-text entity linking, but other variations such image-image+text entity linking are also applicable, especially to medical imaging data. While the development of purely visual prompt-based linking models is valuable for ensuring robustness when text prompts are sparse, uninformative, and lacking, we acknowledge that visual prompts alone may not sufficiently capture user intent in all downstream tasks or may not be preferred by all users. Future work can evaluate the design of more universal medical VEL systems which support a variety of user prompting preferences, including visual prompts, textual prompts, and hybrid visual+textual prompts. To this end, additional evaluations of LLM performance in few-shot, where the model is provided with examples of visual mentions and expected outputs, and finetuning settings would further illuminate methods to leverage the reasoning power

and interactivity of these models for medical V2TEL.

Guided systems like VELCRO are designed to support the workflows of end users and should accurately reflect their needs. The work presented in this thesis provides meaningful insights into a focused, vision-first user-centric VEL formulation in the medical domain. However, because clinical utility is dependent on real-world usefulness and user trust, system evaluations must extend beyond metric-based performance measures to include the direct involvement and feedback of medical professionals. Stakeholder collaboration can help model developers optimize model design for clinical workflows, identify effective explainability techniques for medically-oriented audiences, and suggest domain niches in which these tools can most benefit medical practice. We therefore emphasize healthcare professional engagement from system development to deployment, alongside model explainability, as significant paths of future work and essential pillars of responsible AI design. Ultimately, this thesis aims to establish a quantitative foundation for future research to build upon through direct stakeholder collaboration towards mature, expert-validated image analysis and management systems in the medical domain.

References

- [1] Mamatjan Abdurxit, Turdi Tohti, and Askar Hamdulla. An efficient method for biomedical entity linking based on inter- and intra-entity attention. *Applied Sciences*, 12(66):3191, January 2022.
- [2] A.S. Albahri, Ali M. Duhaim, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96:156–191, August 2023.
- [3] Anton Alekseev et al. Medical crossing: a cross-lingual evaluation of clinical entity linking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4212–4220, Marseille, France, June 2022. European Language Resources Association.
- [4] Jason Ansel et al. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024.
- [5] Nils Barlaug and Jon Atle Gulla. Neural networks for entity matching: A survey. *ACM Trans. Knowl. Discov. Data*, 15(3), April 2021.
- [6] 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 Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang, editors, *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 186–194, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

- [7] Yonatan Bitton, Raphael Cohen, Tamar Schifter, Eitan Bachmat, Michael Elhadad, and Noémie Elhadad. Cross-lingual unified medical language system entity linking in online health communities. *Journal of the American Medical Informatics Association*, 27(10):1585–1592, October 2020.
- [8] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, January 2004.
- [9] Mu Cai, Haotian Liu, Dennis Park, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts, 2024.
- [10] Mathilde Caron, Alireza Fathi, Cordelia Schmid, and Ahmet Iscen. Web-scale visual entity recognition: An llm-driven data approach. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 34533–34560. Curran Associates, Inc., 2024.
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.
- [12] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023.
- [13] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 5152–5161, New York, NY, USA, 2022. Association for Computing Machinery.
- [14] 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.
- [15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin,

- P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875. Curran Associates, Inc., 2021.
- [16] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.
 - [17] 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.
 - [18] 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.
 - [19] 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.
 - [20] 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.
 - [21] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.
 - [22] Xiao Fang, Yi Lin, Dong Zhang, Kwang-Ting Cheng, and Hao Chen. Aligning medical images with general knowledge from large language models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 57–67, Cham, 2024. Springer Nature Switzerland.
 - [23] Ryan Faulkenberry, Saurabh Prasad, Dragan Maric, and Badrinath Roysam. Visual prompting based incremental learning for semantic segmentation of multiplex immuno-flourescence microscopy imagery. *Neuroinformatics*, 22(2):147–162, April 2024.

- [24] Evan French and Bridget T. McInnes. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, 137:104252, 2023.
- [25] 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.
- [26] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends. Comput. Graph. Vis.*, 14(3–4):163–352, December 2022.
- [27] Álvaro García-Barragán, Ahmad Sakor, Maria-Esther Vidal, Ernestina Menasalvas, Juan Cristobal Sanchez Gonzalez, Mariano Provencio, and Víctor Robles. Nssc: a neuro-symbolic ai system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Medical & Biological Engineering & Computing*, 63(3):749–772, Mar 2025.
- [28] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] Zhiwei Hu, Victor Gutiérrez-Basulto, Ru Li, and Jeff Z. Pan. Multi-level matching network for multimodal entity linking. In *Proceedings of the 31st ACM SIGKDD*

Conference on Knowledge Discovery and Data Mining V.1, KDD '25, page 508–519, New York, NY, USA, 2025. Association for Computing Machinery.

- [33] S. K. M. Shadikul Islam, Md Abdullah Al Nasim, Ismail Hossain, Dr. Md Azim Ullah, Dr. Kishor Datta Gupta, and Md Monjur Hossain Bhuiyan. *Introduction of Medical Imaging Modalities*, pages 1–25. Springer Nature Switzerland, Cham, 2023.
- [34] 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.
- [35] Davood Karimi, Haoran Dou, and Ali Gholipour. Medical image segmentation using transformer networks. *IEEE Access*, 10:29322–29332, 2022.
- [36] Hee E. Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69, April 2022.
- [37] 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.
- [38] 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.
- [39] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pages 467–484. Springer, 2025.
- [40] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023.
- [41] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: Language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):96–107, 2024.

- [42] 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)*, Oct 2017.
- [43] Xueling Lin, Lei Chen, and Chaorui Zhang. Tenet: Joint entity and relation linking with coherence relaxation. In *Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21*, page 1142–1155, New York, NY, USA, 2021. Association for Computing Machinery.
- [44] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. Kbpearl: a knowledge base population system supported by joint entity and relation linking. *Proceedings of the VLDB Endowment*, 13(7):1035–1049, 2020.
- [45] 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.
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [47] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022.
- [48] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, March 2024.
- [49] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.
- [50] Jun Ma et al. Efficient medcams: Segment anything in medical images on laptop, 2024.
- [51] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024.

- [52] 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.
- [53] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.
- [54] 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.
- [55] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023.
- [56] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, July 2024. Accessed: 16 Jul 2025.
- [57] OpenAI et al. Gpt-4 technical report, 2024.
- [58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [59] K Praveena, Anandhi R J, Shubhi Gupta, Alok Jain, Ashwani Kumar, and Aseel Muhammad Saud. Application of zero-shot learning in computer vision for biodiversity conservation through species identification and tracking. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, pages 1–6, 2024.
- [60] Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid, Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352, 2023.

- [61] 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.
- [62] 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.
- [63] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, June 2024.
- [64] Ranjan Sapkota, Marco Flores-Calero, Rizwan Qureshi, Chetan Badgujar, Upesh Nepal, Alwin Poullose, Peter Zeno, Uday Bhanu Prakash Vaddevolu, Sheheryar Khan, Maged Shoman, Hong Yan, and Manoj Karkee. Yolo advances to its genesis: a decadal and comprehensive review of the you only look once (yolo) series. *Artificial Intelligence Review*, 58(9):274, Jun 2025.
- [65] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [66] Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. Entity linking meets deep learning: Techniques and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2556–2578, 2023.
- [67] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. Generative multimodal entity linking. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7654–7665, Torino, Italia, May 2024. ELRA and ICCL.
- [68] Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. In *Advances in Neural Information Processing Systems*, volume 37, pages 9153–9177. Curran Associates, Inc., 2024.

- [69] Ayush Singh, Saranya Krishnamoorthy, and John E. Ortega. Neighbert: Medical entity linking using relation-induced dense retrieval. *Journal of Healthcare Informatics Research*, 8(2):353–369, Jun 2024.
- [70] 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, Mar. 2024.
- [71] Camilo G. Sotomayor, Marcelo Mendoza, Víctor Castañeda, Humberto Farías, Gabriel Molina, Gonzalo Pereira, Steffen Härtel, Mauricio Solar, and Mauricio Araya. Content-based medical image retrieval and intelligent interactive visual browser for medical education, research and care. *Diagnostics*, 11(8):1470, August 2021.
- [72] Fabian M. Suchanek, Mehwish Alam, Thomas Bonald, Lihu Chen, Pierre-Henri Paris, and Jules Soria. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’24, page 131–140, New York, NY, USA, 2024. Association for Computing Machinery.
- [73] 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 Association for Computational Linguistics: ACL 2024*, pages 816–826, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [74] 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.
- [75] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, November 2023.
- [76] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.
- [77] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors, 2025.

- [78] Neha Tilak, Sunil Gandhi, and Tim Oates. Visual entity linking. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 665–672, 2017.
- [79] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.
- [80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [81] Yongquan Wan, Guobing Zou, and Bofeng Zhang. Composed image retrieval: a survey on recent research and development. *Applied Intelligence*, 55(7):482, Feb 2025.
- [82] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024.
- [83] Rebecka Weegar, Linus Hammarlund, Agnes Tegen, Magnus Oskarsson, Kalle Åström, and Pierre Nugues. Visual entity linking: A preliminary study. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [84] Wikidata. Introduction. <https://www.wikidata.org/wiki/Wikidata:Introduction>, Jun 2025. Accessed 2025-07-18.
- [85] 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.
- [86] 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.
- [87] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

- [88] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.
- [89] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.
- [90] 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, Apr. 2025.
- [91] Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in Oncology*, Volume 11 - 2021, 2021.
- [92] Barry Yao, Sijia Wang, Yu Chen, Qifan Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. Ameli: Enhancing multimodal entity linking with fine-grained attributes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2816–2834, St. Julian’s, Malta, March 2024. Association for Computational Linguistics.
- [93] Jiuxiang You, Zhenguo Yang, Qing Li, and Wenying 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.
- [94] Hang Yu, Laurence T. Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021.
- [95] Jing Zhang, Qiuge Qin, Qi Ye, and Tong Ruan. St-unet: Swin transformer boosted unet with cross-layer feature enhancement for medical image segmentation. *Computers in Biology and Medicine*, 153:106516, 2023.
- [96] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2025.

- [97] Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. Visual entity linking via multi-modal learning. *Data Intelligence*, 4(1):1–19, 02 2022.
- [98] Qiushuo Zheng, Hao Wen, Meng Wang, Guilin Qi, and Chaoyu Bai. Faster zero-shot multi-modal entity linking via visual-linguistic representation. *Data Intelligence*, 4(3):493–508, 07 2022.
- [99] Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. Question answering over knowledge graphs: question understanding via template decomposition. *Proceedings of the VLDB Endowment*, 11(11):1373–1386, 2018.
- [100] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.