MSc Artificial Intelligence
Master Thesis

Synthetic Caption Enrichment for Fine-Grained Classification in Low-Resource Domains

by
Jelle van der Lee
12538019

October 1, 2025

36 EC
01/04 - 01/10/2025

Daily Supervisor:
J.E. van Woerden

UvA Supervisor:
A. Bhowmik

Examiner:
Dr C. Snoek

Universiteit van Amsterdam

Contents

Abstract

Over the last years, Vision-Language Models (VLMs) have shown impressive zero-shot classification capabilities. However, their performance significantly decreases in low-resource, fine-grained domains where data is scarce and visual distinctions between classes are subtle. Directly fine-tuning models on such limited data is often impractical, as it typically leads to overfitting, causing the models to not generalize well to new inputs. State-of-the-art information-augmentation methods, such as retrieval-based enrichment, often fail in these scenarios due to the under-representation of specialized concepts in public web-crawled databases. Existing generation-based methods, on the other hand, are typically designed for high-resource settings and perform one-sided enrichment, missing the opportunity to align both image and class representations simultaneously.

This thesis addresses this gap by proposing **SynCE: Synthetic Caption Enrichment**, a training-free, generation-based framework designed to improve zero-shot classification in truly low-resource domains with a focus on military vehicles. SynCE overcomes the bottleneck that retrieval methods introduce when the dataset is under-represented in the retrieval database, by synthesizing attribute-focused descriptions for both the candidate class labels and the query image, using a Large Language Model (LLM) and a caption-capable VLM. At inference time, these generated captions are encoded and fused with the original embeddings to create more discriminative representations that contain information the classifier might lack in these low-resource domains.

The main evaluations of our method are run on a truly low-resource dataset of military vehicles, images which are often classified and not presented online, causing there to be a limited amount of publicly available training data. The results demonstrate that SynCE outperforms standard zero-shot baselines, and recent retrieval-based and generation-based approaches. Furthermore, we conclude a model-dependent effect of the enrichment strategy: the SigLIP classifier benefits most from a combination of image- and class-side enrichment, whereas the PerceptionEncoder classifier achieves peak performance with class-side enrichment alone. This finding suggests an interaction between the VLM's feature alignment and the optimal enrichment technique. Ultimately, this work establishes dual-sided synthetic enrichment as an efficient and flexible strategy for improving VLM classification performance in specialized domains where other methods fall short. While promising, our qualitative analysis also reveals the limitations of our proposed method and enrichment methods in general, highlighting that the method's effectiveness is ultimately capped by both the quality of the generated captions and the classifier's inherent ability to interpret them.

The code for this project can be found at `SynCE`.

Chapter 1

Introduction

1.1 Problem Description

Low-resource settings refer to domains in which the available data and annotations are scarce, such that AI models lack sufficient examples for effective training, causing performance to suffer. In Natural Language Processing, extensive research has addressed this challenge by developing techniques to mitigate limited training data such as cross-lingual transfer learning [4], data augmentation [31], and meta-learning [13][15]. These approaches have been impactful on increasing model performance in low-resource domains, that often involve languages with far smaller online text corpora compared to English or other widely-spoken languages.

In contrary to NLP, the exploration of low-resource scenarios in Computer Vision and, more recently, in Foundation Models remains relatively underdeveloped, despite the growing importance of these models. Recent work by Zhang et al. (2024) [41] defines the challenges of these domains as a combination of severe data scarcity, subtle fine-grained differences, and a significant domain shift from natural images. Consequently, most large-scale vision models are trained on large, publicly available datasets like ImageNet [12] or LAION [32], making them less performant in specialized domains where images are scarce, restricted or highly fine-grained. One of these specialized domains is the classification of military vehicles, where much of the imagery is classified and therefore inaccessible for model training. Moreover, this data contains subtle fine-grained differences between different visually similar vehicle types and variants, making the classification task particularly challenging.

The recent shift in image classification tasks concerns the application of Vision-Language Models (VLMs), such as CLIP [27], BLIP [20], or DINO [7], which jointly embed images and textual descriptions into a shared embedding space. These models achieve impressive performance in open-domain and high-resource (classification) tasks, but they tend to lose significant capability when applied to low-resource, fine-grained domains where the pretrained distributions do not align well with the target data.

To overcome the issues of VLM classification tasks that follow from limited data in low-resource domains, various strategies have been explored. One of such includes synthetic data generation for model fine-tuning, which can help in some cases but often fails in truly low-resource domains. This is due to the fact that image generation models themselves also suffer from limited training data in low-resource domains, producing images that are either unrealistic or too similar to existing samples, which is being caused by the fact that the data distribution of rare domains is not well-represented in the generative models latent space [34][25]. Fine-tuning directly on the scarce available data often results in overfitting, causing models to not

generalize well across new inputs not seen in their training data [33].

An alternative perspective is to augment model inputs with additional textual information using either retrieval- or generation-based methods. Retrieval-based approaches, such as the 'Combination of Retrieval Enrichment (CoRE)' [11] method, enrich VLM representations in a training-free manner by retrieving external captions from a large web-crawled database to improve classification. While promising, its effectiveness is strongly tied to the quality and presence of the dataset images in the retrieval database, showing degradation in performance or only slight improvements over other methods in underrepresented domains like military vehicles. Alternatively, generation-based methods like 'What Do You See? (WDYS)' [3] synthesize the external information 'on the fly' by prompting a Multimodal Large Language Model to describe the query image directly. However, these approaches are often designed for high-resource settings and struggle to capture the subtle, fine-grained details necessary for classification in specialized low-resource domains, highlighting the need for an alternative method specifically designed for low-resource and fine-grained settings.

1.2 Research Objective

This thesis aims to address the research gap stated above, by investigating whether enriching the class and image embeddings of VLMs with additional generated descriptions can improve fine-grained classification in low-resource domains. A key limitation of retrieval-based enrichment is that in certain low-resource datasets, it often returns descriptions that are irrelevant or misaligned with the input images, which can degrade performance. In contrast, synthetic captions leverage the broad knowledge of the most recent LLMs and VLMs to produce more accurate and contextually aligned descriptions, and allows for precise control over which details to emphasize which is a key requirement for distinguishing subtle, fine-grained differences. Building upon these advantages we introduce the method of **SynCE (Synthetic Caption Enrichment)**, which is a training-free, generation-based enrichment method that requires no large external index (unlike retrieval-based methods) and is simple to deploy out of the box, while staying flexible in terms of selecting the captioning or classification model. Unlike existing generation-based methods that perform one-sided enrichment via simple averaging, SynCE employs a dual-sided and dynamically weighted fusion strategy to align both the image and class representations. From its ability to select the most recent models, it follows that as these models improve, SynCE's performance improves with it. Specifically, SynCE utilizes Large Language Models (LLMs) and Vision-Language Models (VLMs) to generate descriptive captions for both the class labels and input image. At inference time, we leverage the enriched embeddings, formed by fusing the original inputs with the generated captions, to enhance zero-shot classification in low-resource domains, with a particular focus on military vehicles.

As shown in Figure 1.1, this thesis focuses on handling image classification on truly scarce data, where only a few examples are available and the differences between classes are very subtle and fine-grained. In contrast to high-resource domains where large-scale and diverse datasets allow for efficient model training, we focus on the challenges that are presented by low-resource, specialized domains, with military vehicles serving as a representative study. This focus is motivated by the task to develop practical applications at the Netherlands Organization for Applied Scientific Research (TNO), by which the research is proposed.

The general research question guiding this thesis is:

To what extent can a training-free framework, which enriches both image and class embeddings with descriptive information, improve zero-shot, fine-grained classification accuracy in low-resource domains?

From this, we derive the following sub-question:

To what extent does the effectiveness of enrichment with synthetically generated captions depend on the interaction between the enrichment strategy (image-side, class-side, or combined) and the chosen VLM classifier?

These questions will be investigated through systematic experimentation, including the evaluation of different prompt tuning strategies, enrichment methods and VLM classifiers.

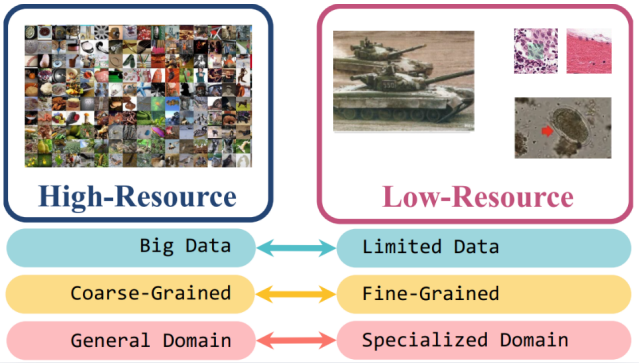


Figure 1.1: **High-resource vs. Low-resource image domains**. High resource vision tasks focus on Big Data collected at scale, coarse-grained difference between classes and a more general domain. In this thesis, we focus on low-resource tasks, which encompass the difficulties of handling environments with limited data, fine-grained differences between classes and specialized domains. Adapted from Zhang et al. [41].

1.3 Contributions

To summarize, our contributions are:

- We highlight the problem setting by analyzing the limitations of existing state-of-the-art approaches in low-resource, fine-grained domains. We explore the impracticality of simple prompt tuning techniques and reproduce two key methods from recent research: the generation-based 'What Do You See?' (WDYS) and the retrieval-based 'Combination of Retrieval Enrichment' (CoRE). Our analysis demonstrates why these methods are not well-suited for specific specialized datasets, thereby establishing the research gap that our work addresses
- We propose the training-free generation-based method of SynCE, that adresses zero-shot image classification in low-resource domains while overcoming the issues present in WDYS and CoRE
- We establish a benchmark for zero-shot low-resource image classification on military vehicles, on which we evaluate our reproduced and proposed methods

- We show that our method outperforms existing methods on fine-grained image classification in low-resource domains while remaining training-free and simple to implement
- We conduct a detailed qualitative analysis that identifies key limitations in our method and enrichment strategies in general, establishing a clear foundation for future work

1.4 Overview

The remainder of this thesis has the following structure:

- **Chapter 2** provides a review of the relevant literature on low-resource image classification, VLM prompting and existing methods for representation enrichment.
- **Chapter 3** describes the concepts of the most important methods used in this paper, such as Vision-Language Models and Classification Enrichment.
- **Chapter 4** details the methodology, covering the preliminary prompt tuning study, the implementation of the baseline methods (WDYS and CORE), and the design of our proposed method, SynCE.
- **Chapter 5** outlines the experimental setup, including the datasets, specific models, implementation details, and the evaluation metrics used to assess performance.
- **Chapter 6** presents and analyzes the results from all experiments.
- **Chapter 7** discusses the broader implications and limitations of the findings presented in the previous chapter.
- **Chapter 8** concludes the thesis by answering the research questions, highlighting implications for future work and discussing remaining challenges.

Chapter 2

Related work

2.1 Low-resource Settings

Low-resource settings refer to domains where the available data and annotations is too scarce to effectively train AI models, often causing their performance to suffer. These domains present a critical challenge for many state-of-the-art deep learning architectures, which are data-hungry and rely on massive datasets to learn the data's patterns and gain reliable performance. When trained on limited data, models often fail to generalize to new inputs and are prone to overfit, where the training examples are memorized instead of truly learning the underlying task [33]. The challenge of low-resource data has been addressed with varying levels of detail across different AI domains, creating a contrast between the well-established techniques in Natural Language Processing and the more recent, underexplored, problems in Computer Vision.

2.1.1 NLP

In Natural Language Processing (NLP), low-resource challenges are well-recognized and have been largely mitigated. These settings typically involve languages that lack the large online text corpora available for languages like English. The field has developed several effective techniques to overcome these data limitations. One often utilized approach is cross-lingual transfer learning, which uses available labeled data from a high-resource 'source' language to fine-tune a model for a 'target' language that lacks such data [4]. Another common technique is using data augmentation, where the limited data is overcome by creating new training instances made from transformations on the existing instances in a way that does not change the label, such as replacing words with synonyms or paraphrasing via back-translation. These effective augmentation methods are neatly highlighted by Sahin et al. (2022) [31], who provide a comparative study on the different text augmentation techniques. A third strategy is meta-learning, which is a multi-task learning method that trains a model on a set of high-resource tasks and learn it how to best approach and adapt to a new, low-resource target task [13][15].

2.1.2 Vision with VLMs

In contrast to the progress in NLP, truly low-resource settings in computer vision and vision-language tasks, where only a few hundred images are available, remain underexplored. In their paper 'Low-Resource Vision Challenges for Foundation models' [41], Zhang et al. (2024) directly address this gap by building a benchmark of datasets from the specialized domains of circuit diagrams, historic maps and mechanical drawings. Their study shows that existing foundation models, despite their impressive zero-shot classification results on natural images, fail to generalize well under conditions of severe data scarcity, subtle fine-grained differences

and substantial domain shifts. To tackle these challenges, the authors propose three simple yet effective baselines that outperform standard transfer-learning and data-augmentation methods, and prove that there are still open low-resource vision challenges within foundation models that require continued research. A critical area of this continued research focuses on enhancing the textual input to VLMs, while altering the language prompt is a direct and effective way to improve the model performance in both challenging and more general domains.

2.2 VLM Prompting

Vision-Language Models rely on the combination of visual and textual inputs. In classification tasks, the manner in which the textual input is written has proven to be an important factor in the performance of the model. Recent works have done extensive research into the area of VLM prompting, ranging from learning the best prompt to leveraging LLMs to generate descriptive or hierarchical-aware descriptions for the classifier input.

2.2.1 Prompt Learning

Much research has been done on tuning the input prompt into CLIP's model to increase the classification accuracy on the images. Zhou et al. (2022) introduce the methods of **CoOp** [43] and **CoCoOp** [42] that both learn the most effective prompt for a dataset to utilize as textual input into CLIP-like models. These prompts are however not interpretable or readable by humans, but are rather a collection of the learned context vectors that serve as optimized input tokens without any semantic meaning. In order to make the prompt optimizer contain semantic meaning, not overfit to base classes seen during training and be interpretable by humans, Du et al. (2024) introduce the method of **IPO (Interpretable Prompt Optimizer)** that utilizes LLMs to generate textual prompts dynamically [14]. While effective, this lack of interpretability in the learned prompts has resulted in another alternate line of research focused on generating descriptive and semantically meaningful text to guide the underlying classifier.

2.2.2 Descriptive and hierarchical prompts

An alternative to prompt learning involves enriching textual inputs with descriptive information generated by LLMs. The **Classification by Description** approach [24] prompts an LLM to list visual attributes for each class (e.g., "two legs" or "a beak" for a "hen") and classifies based on the average similarity between the image and these descriptor embeddings. This improves both accuracy and model explainability, as predictions can be traced back to specific features.

However, as these descriptions can be too similar for closely related but different classes, additional works create more discriminative prompts using comparative and hierarchical structures. Some methods dynamically construct a class hierarchy via clustering, by prompting an LLM to recursively compare and group classes [28]. Other works use LLMs to generate *comparative prompts*, strengthening the difference between the leaf class and its related classes in the hierarchy, and *path-based prompts*, that focus on the unique characteristics of the leaf class. For each leaf in the query, these two groups of language prompts are constructed and are aimed to reduce misclassification severity by "making better mistakes" [21]. These strategies demonstrate that using structured, comparative text is a powerful training-free method to improve fine-grained VLM classification performance.

Alternatively, some recent research examines the true source of these performance gains when utilizing LLMs to build descriptive prompts. Roth et al. (2023) showcase that replacing

LLM-generated descriptions with random character or word sequences, independent from the input class, can achieve comparable zero-shot classification performance. With their method of *WaffleCLIP*, they conclude that much of the benefits from the performance gains might arise from the averaging effect of structured noise in stead of the relevant semantic context that the descriptions bring, questioning whether providing prompts with true knowledge has any positive effects [29].

2.2.3 CLIP input extension

While CLIP's text encoder does not allow for input sequences that are longer than 77 tokens, several recent works explore increasing the input size or utilizing more expressive text encoders in order to overcome this issue. These methods overcome the hindrance of performance on tasks requiring longer descriptions, such as extensive and long input prompts with specific prompt tuning techniques. The methods of **Long-CLIP** [39], **TULIP** [26], and **LLM2CLIP** [18] propose mechanisms to handle longer descriptions or distill richer LLM text into CLIP-compatible embeddings, allowing class prompts to include additional context that would otherwise not be possible with the maximum of 77 input tokens. These methods address the limitations of the input prompt length, but the principle of enhancing the VLM inputs can be extended beyond just the class-side prompts and static prompt tuning. For example, a more general framework of information-augmented classification could be utilized, where external knowledge is used to enrich the representations of both images and classes at inference time.

2.3 Information-augmented zero-shot classification

To improve the performance of Vision-Language Models (VLMs) in zero-shot classification tasks without fine-tuning, recent work has focused on introducing external information to augment the image- and class-embeddings at inference time. The core method of this research contains approaches to enrich the representations in the embedding space of either the input image, the candidate class labels, or both, using the additional textual information. The external knowledge helps the VLM produce more discriminative embeddings and tries to pull the image and its correct class embedding more aligned in the embedding space, which is particularly useful for fine-grained and low-resource domains where it introduces information that the model's pretrained knowledge might lack. The textual knowledge is typically sourced by either **(i)** retrieving it from existing databases or by **(ii)** generating it at inference time by using the generation capabilities of VLMs or LLMs.

2.3.1 Retrieval-based enrichment

Retrieval-based enrichment methods retrieve external knowledge by fetching relevant captions from large, web-crawled databases. The core idea for this approach stems from the field of Natural Language Processing where methods like **Retrieval Augmented Generation** combine pre-trained models with a retrieval method in order to access a vast corpus of external knowledge [19]. The thought behind using this approach is to ground the VLM's representations in real-world, human-written text, which can provide valuable domain context and improve performance on knowledge-intensive tasks.

As briefly mentioned in section 2.1.2, a notable example of this approach is the method of **Combination of Retrieval Enrichment (CoRE)**. CoRE is a training-free approach that retrieves the captions most semantically similar to the candidate classes and the image to be classified, using them to enrich the representations of both inputs in the embedding space at

inference time.

The method uses the external databases of CC12M [8] and COYO-700M, and collects a set of datasets that covers diverse, truly low-resource datasets, including medical imaging [35], rare plants [16] and mechanical circuits [41]. The method uses an architecture where the image embeddings are used for image-to-text retrieval and the class name embeddings are used for text-to-text retrieval, to retrieve captions from the external database. By fusing the original embeddings with those of the retrieved captions, CoRE creates more context-aware representations for classification. The proposed architecture of CoRE is shown and further explained in Figure 2.1

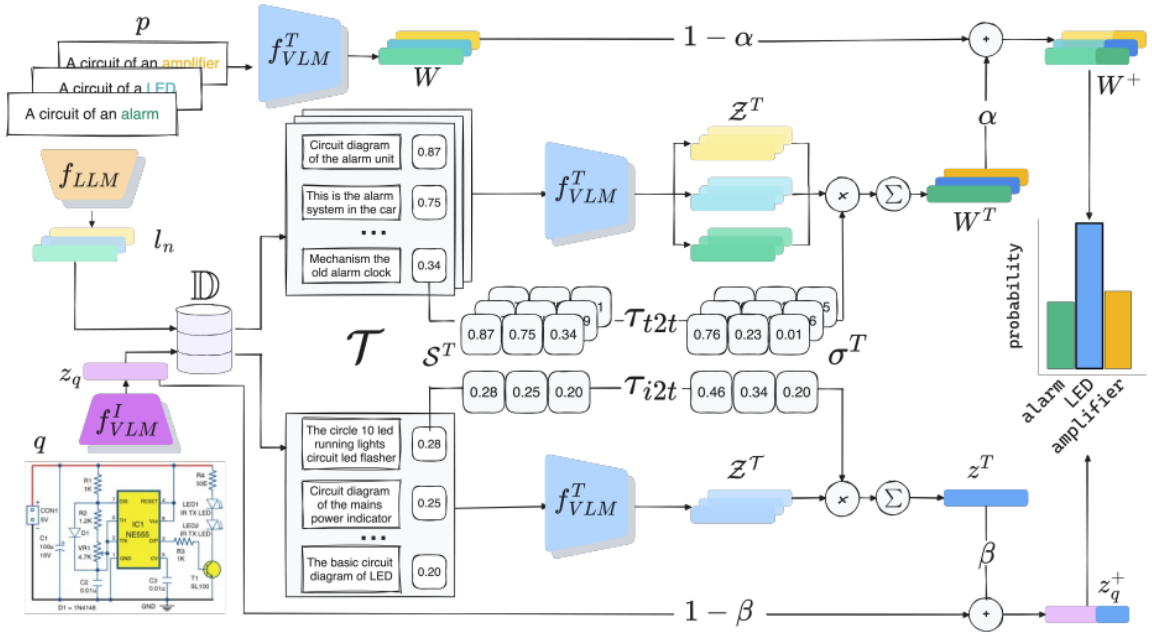


Figure 2.1: CoRE's proposed architecture, enriching both the image embedding z_q and the class prompts p with retrieved captions from a large-scale web-crawled database $\mathbb{D}$. They weight the retrieved captions $\mathcal{T}$ with their similarity scores S^T, which are skewed with controllable temperatures τ_{i2t} and τ_{t2t}. By combining the retrieved captions embedding with the original representations W and q through α and β, they obtain enriched representations W^+ and z_q^+ which are employed for zero-shot classification [11].

Other works have also explored retrieval-based enrichment, where one notable method aims to remove the constraint of a predefined vocabulary. The paper of **Vocabulary-Free Image Classification (VIC)** [10] formalizes a novel task, with the same name, where a model is not given a set of candidate classes at test time but instead it must identify the most fitting class from an unconstrained, language-induced semantic space. Based on this task, the authors propose the method **CaSED**, that first performs image-to-text retrieval to find the captions most relevant to the input image and from these retrieved captions, it extracts a set of potential candidate class names. Finally, it scores these candidates using a multi-modal approach that considers both the image-to-text similarity and the text-to-text similarity to make the final class prediction. These similarity scores are computed by comparing the candidates to the centroid of the retrieved captions.

Even though the retrieval-based methods are quite powerful, a key limitation is that for certain low-resource datasets the data is underrepresented, or completely absent, in the retrieval database, causing performance to suffer. This under-representation causes retrieved descriptions to be irrelevant or misaligned, creating a significant bottleneck for this approach that is

also specifically mentioned by the authors of CoRE [11].

2.3.2 Generation-based enrichment

When retrieval fails or is not feasible, an alternative is to generate descriptive text using the vast knowledge encoded in LLMs and M-LLMs. This approach can synthesize contextual information even for concepts that are rare or absent in retrieval databases, while operating in a training-free manner at inference time.

The recent method of **Classification by Description** prompts an LLM like GPT-3 to generate a list of descriptive visual attributes for each class label (e.g., 'has feathers' for a 'hen'). The image is classified based on its similarity to these sets of descriptors, improving both the accuracy and the model explainability [24]. More recent work by Abdelhamed et al. (2025) [3] utilizes an M-LLM to enrich the image-side of the VLM classifier. Their method prompts an M-LLM to generate a detailed description and an initial class prediction from the input image. These generated texts are then encoded and fused with the original image feature, in order to create a more robust query representation for the classifier.

In order to boost performance even futher, some methods incorporate a training step that utilizes generated text to adapt the VLM classifier. Even though it is not strictly zero-shot at inference time, these approaches use generated data to improve the model's general zero-shot capabilities. One example is the method of **AdaptCLIPZS**, that fine-tunes CLIP on a 'bag' of LLM-generated class descriptions (on the class's appearance, habitat, etc.) combined with large, existing image datasets [30]. An alternate and more parameter-efficient approach is **VDT-Adapter**, which freezes CLIP and then trains a lightweight self-attention adapter to learn how to select, and aggregate, the most relevant sentences from a pool of generated descriptions [23].

Eventually, the performance of all generation-based methods hinge on the quality of the generated text. Following this conclusion, Chen et al. (2024) [9], with their **ShareGPT4V** dataset, highlight that highly detailed and comprehensive captions will lead to significantly better modality alignment and model performance when compared to brief or generic descriptions.

One common limitation of these works is their focus on high-resource and coarse-grained benchmarks. When applied to truly low-resource domains, whose images contain subtle, fine-grained differences, the generated descriptions might lack the necessary specificity to correctly distinguish between visually similar classes, pointing to an important research gap that remains to be addressed. As our reproduction work in Section 4.2.2 will highlight, the generation-based method's performance significantly degrades in truly low-resource domains such as military vehicles, where fine-grained distinctions are crucial. Additionally, most of these approaches perform *one-sided enrichment* only. They either focus on enhancing the class-side prompts (like **Classification by Description** or **AdaptCLIPZS**) or the image-side representations (like **'What Do You See'** from Abdelhamed et al. (2025)). Although Abdelhamed et al. (2025) explore a form of class-side enrichment by averaging over a limited number of class prompts, their method misses the opportunity to dynamically align both the image and class representations with consistent and context-aware information. This coordinated, dual-enrichment strategy is a key feature of the CoRE framework, which is specifically designed to improve performance in low-resource domains.

2.4 Discussion and Research Gap

The literature review shows a dilemma in information-augmented zero-shot classification methods, specifically for low-resource domains. On the one hand, retrieval-based methods like CoRE offer an efficient dual-enrichment architecture that improves both the image and class embeddings with external knowledge that the VLM encoders might lack in low-resource domains. However, these methods are constrained by the coverage of a dataset in the external database, making them unreliable for specific specialized domains like military vehicles, where the data is available in a very limited amount on the public web.

On the other hand, generation-based methods overcome this external database underrepresentation by synthesizing knowledge directly from powerful LLMs and VLM captioners. However, the existing approaches are either **(i)** not designed for challenges that arise when working with truly low-resource domains and fine-grained differences, causing poor performance, or **(ii)** they typically focus on one-sided enrichment only, not covering the possibility of aligning both the image and class representations of the embedding space in a combined manner at inference time.

This conclusion shows a clear research gap and the need for a method that combines the strengths of both approaches named above. Specifically, an ideal solution for zero-shot classification in truly low-resource domain, where the data is under-represented in external databases, should be **generation-based**, **dual-sided**, **aligned**, focusing on drawing the image embedding towards its correct class embedding in the representation space, and **explainable**, using the generated captions to give insights into the model's decision-making process.

This thesis aims to fill the research gap by proposing a method fulfilling the characteristics above, specifically designed to provide efficient and explainable classification for datasets where retrieval-based methods fail and existing generation-based methods under-perform, while remaining completely training-free.

Chapter 3

Background

In this chapter, we highlight the necessary background useful for understanding our method. Section 3.1 introduces Vision-Language Models and showcases how they map images to text into a shared space, why this enables efficient zero-shot classification and how prompt design affects the performance. Additionally, we briefly outline the specific VLMs used in this thesis (CLIP, SigLIP and PerceptionEncoder). Section 3.2 highlights the functionality of classification enrichment and builds intuition for how the approach operates in embedding space, describing the functionalities of Inter- and Intra-Class distance, while showcasing a visual sketch of what enrichment might look like. Together, these sections establish the foundation on which our method of SynCE is built.

3.1 Vision-Language Models

Vision-Language Models (VLMs) are multimodal models that learn from both images *and* text, embedding them into a shared representation space. In this space, semantically similar image–text pairs lie close together, while unrelated pairs are pushed apart. This property makes VLMs particularly effective for image classification: an image can be classified by simply comparing its embedding to text embeddings that describe candidate classes, without requiring task-specific training. As illustrated in Figure 3.2, this enables **zero-shot classification**: encode the image once, encode a short prompt for each class (e.g., `"a photo of a {class}"`), and select the class whose text embedding is most similar to the image embedding.

VLMs are introduced to combine the domains of images and text, and they represent a shift from classic image classifiers such as Convolutional Neural Networks (CNNs). These traditional models are trained on a **closed set** of classes and being trained to map an image to a fixed output (such as the 1000 classes in ImageNet), where the labels do not contain any semantic meaning to the model and it is not possible to classify concepts not seen during training [17]. Instead of having these disadvantages, VLMs are considered as **open-vocabulary**: their knowledge is grounded in natural language, meaning the labels contain semantic meaning to the model, and their classifier is constructed at inference time using textual prompts, which allows the models to generalize to new classes unseen during training [40].

For most domains, VLMs are a promising starting point due to their large web-scale pretraining. While they are trained on billions of image-text pairs, they serve as powerful feature extractors with strong capabilities in generalization. This extensive knowledge provides a much more effective foundation in low-resource domains than training a classical model from scratch on the few hundred available images, which would lead to severe overfitting.

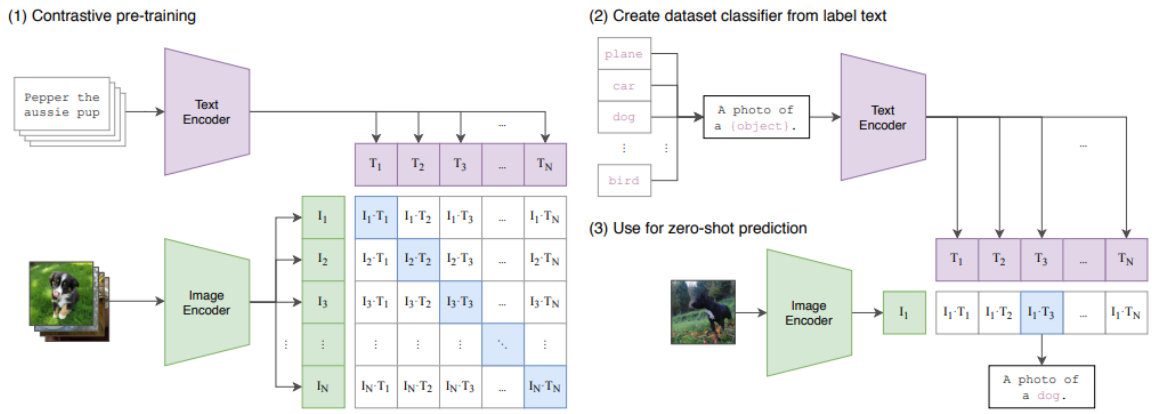


Figure 3.1: CLIP's architecture. The model jointly trains an image and a text encoder to predict the correct pairings of a batch of training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes [27].

Despite their impressive capabilities across a wide range of tasks however, VLMs still face limitations when applied to truly low-resource domains. As briefly mentioned in section 2.1.2, Zhang et al. (2024) accurately identify the three main issues that VLMs struggle with in these domains: **(i)** Data Scarcity, **(ii)** (Subtle) Fine-Grained Differences and **(iii)** Specialized Domain Shift. As a result of these issues, the authors conclude that existing foundation models often fail to generalize well in these low-resource tasks and that the need for additional methods to bridge this performance gap is high.

3.1.1 CLIP

The most well-known VLM classifier is the model of CLIP (Contrastive Language-Image Pre-training) from OpenAI [27], highlighted in Figure 3.2. **CLIP** follows a dual-encoder design: an image encoder (often a Vision Transformer) and a text encoder (a Transformer) map inputs into the same space. During pretraining, matched image-text pairs are pulled together while mismatches are pushed apart in the embedding space. At inference time, zero-shot classification is implemented by comparing an image embedding to a small set of text embeddings derived from the candidate class prompts. An important factor from CLIP is that we're able to *prompt tune* the input text: template wording, attribute phrasing and small tweaks in the semantic input can significantly alter the model performance. Practices such as prompt ensembling (averaging scores over multiple templates) or prompt learning (learning the most efficient input template) often yield additional gains. In this thesis, CLIP serves as a general baseline, mainly for our prompt tuning experiments, and a reference point for later enrichment.

3.1.2 SigLIP

SigLIP is a VLM classifier built upon CLIP, maintaining the dual-encoder architecture but replacing the standard contrastive learning with softmax normalization by a simple pairwise Sigmoid loss. SigLIP models were the cause for an improvement in classification performance and provide stronger image-text alignment on many zero-shot benchmarks while they maintain a similar architecture to CLIP. For this thesis, SigLIP is used as a drop-in replacement to CLIP that in most cases delivers a stronger zero-shot baseline, which is valuable in our case where the data is scarce. SigLIP is also used as the backbone of our reproduced CoRE method [11], meaning that it additionally gives us a fair comparison point to the baseline.

3.1.3 PerceptionEncoder (PE)

The PerceptionEncoder is a recent state-of-the-art vision encoder by Meta for image and video understanding, which is pre-trained via simple vision-language learning and is introduced in the paper of *'Perception Encoder: The best visual embeddings are not at the output of the network'* [6]. An important finding from the authors is that the most useful representations do not live in the final layer but that the best features for downstream tasks often emerge in intermediate layers. The components of the models introduced in their paper deliver state-of-the-art results on multiple benchmarks, such as zero-shot classification, visual Q&A and captioning. While our method of SynCE requires a VLM with capabilities of both zero-shot classification and image captioning, by letting it describe what is inside an image, we utilize this model in order to experiment whether using the same VLM for both stages will have a positive fact on the performance. Apart from this, the model is utilized while the PE-Core image classifier provides a strong starting point for classification in our low-resource domain.

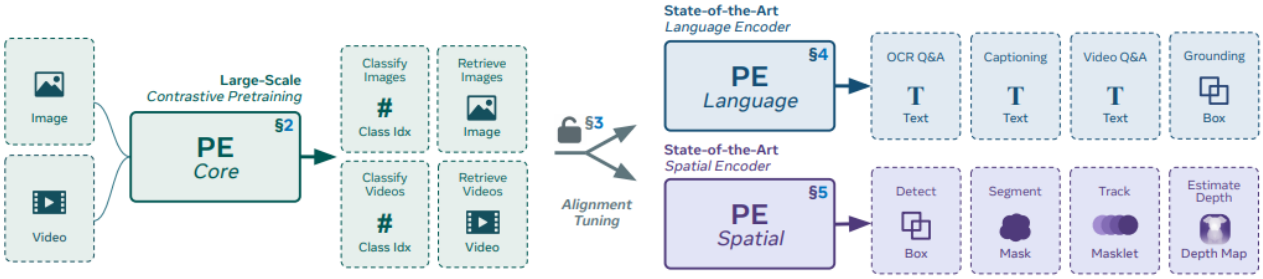


Figure 3.2: The PerceptionEncoder (PE) is a family of large-scale vision encoder models with state-of-the-art performance on a large variety of vision tasks, including image classification and captioning [6].

3.2 Classification Enrichment

Even though Vision-Language Models provide a powerful foundation for zero-shot image classification, their effectiveness is limited in low-resource domains where subtle fine-grained differences between classes are common. In such scenarios, the VLM's embedding space often becomes 'semantically crowded' around the candidate classes, where the representations of visually similar classes (such as different variants of the same military vehicle) are located very close to each other, making it difficult for the classifier to reliably distinguish between them based on the input image, which often leads to misclassifications.

The term **classification enrichment** refers to the method designed to address this challenge in low-resource domains. It aims to augment the original, simple representations of the classes and/or the images in the embedding space with additional descriptive textual information, often in a training-free manner at inference time. Instead of relying on the ambiguous, difficult to distinguish, embedding, more discriminative and context-aware representations are created, that highlight the subtle features that differentiate the candidate classes. Additionally, the textual descriptions introduce very specific knowledge about the classes that the VLM classifier might lack in these low-resource domains due to its limited training data.

The intuition behind the approach can be understood by highlighting the effects on the geometry of the embedding space. The process aims to achieve the following goals:

1. **Increase Inter-Class Distance:** By enriching each class prompt with unique and detailed descriptive attributes, their resulting text embeddings are pushed further apart in the geometrical space. For example, augmenting the "T-72" class prompt with text about its *smooth turret design* and the "T-90" class prompt with details about its *upgraded explosive reactive armor* creates more distinct representations in the space. This improved separation between the class prototypes creates a clearer decision boundary for the VLM classifier. A simple sketch of the intuition behind this inter-class distance improvement is presented in Figure 3.3.

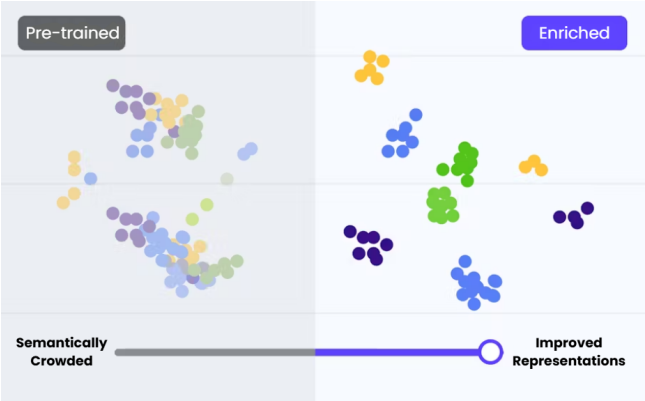


Figure 3.3: Sketching the intuition behind classification enrichment by increasing the inter-class distance in the embedding space. By introducing textual descriptions, similar class embeddings are pushed further apart to allow for more accurate predictions by the VLM classifier. This sketch is based on a relevant blog-post, explaining the intuition behind enhancing geo-spatial embeddings [1].

2. **Decrease Intra-Class Distance:** Additionally, by generating a detailed description of the attributes in the input image itself, the image's representation is pulled closer to its correct, enriched class prototype, while the introduced textual descriptions should be similar. An image of a T-72, combined with the encoded textual description mentioning its unique features, will have a higher similarity to the enriched "T-72" class embedding, that describes similar features, and a lower similarity to the "T-90" class embedding, that focuses on other specific attributes. By fusing each image embedding with this textual representation, all individual embeddings for that class are drawn towards a more compact, central point in the embedding space, thus aiming to tighten the cluster and reduce the intra-class distance.

In essence, classification enrichment is used as a method that utilizes descriptive text to realign the local embedding space by pushing similar class representations apart while pulling the query image toward its correct class label. This approach makes the classification prediction more robust and accurate, especially when dealing with very subtle visual differences that images in low-resource domains often contain. Additionally, it aims to introduce external textual knowledge that the VLM classifiers might otherwise lack in these domains, due to the under-representation of the images in the pre-train data.

Chapter 4

Methodology

4.1 Preliminary Study: Prompt Tuning

We began our initial experiments by trying different methods of Prompt Tuning as a simple and lightweight way to potentially boost the zero-shot accuracy of CLIP. First we scraped web-pages from the archived website of MilitaryToday.com, which contains detailed, and grounded since it is written by experts, information about different military vehicles. Using these scraped pages, we constructed a json containing all metadata from each vehicle using an LLM, which was then used to build metadata-driven descriptions for each of our candidate classes. The experiments were designed to evaluate the impact of three key factors on classification performance:

1. **Caption Length:** Varying the length of the generated descriptions (e.g., Short, Medium, Long).
2. **Knowledge Source:** Whether the LLM's used knowledge was grounded to the scraped metadata or whether it relied on its own world knowledge about the dataset.
3. **Semantic Focus:** Directing the prompt's focus on specific types of phrasing, such as: technical specifications, purely visual descriptions or highlighting the differences between different vehicle types.

In Figure 4.1 we showcase some examples of prompts generated using this technique for different class inputs. The classification performance of these templates is reported in Section 6.1.

4.2 Baseline Methods

To provide more context to the performance of our proposed method, SynCE, we implemented and evaluated two state-of-the-art baseline methods for information-augmented zero-shot classification: one retrieval-based (**CoRE**) and one generation-based (**What Do You See?**).

4.2.1 Combination of Retrieval Enrichment (CoRE)

The CoRE method, introduced by Dall'Asen et al. (2024), is a training-free, retrieval-based approach designed for low-resource domains [11]. Instead of generating new text, CoRE enriches representations by retrieving existing, relevant captions from a large, web-crawled database (e.g., CC12M or COYO-700M).

```
[Baseline]
    An image of a T-72, a military vehicle
```

```
[Country]
    An image of a Russian T-72, a military vehicle
```

```
[Metadata + LLM knowledge]
    An image of a T-72. It is a Soviet MBT introduced in the early 1970s.
    It features a low-profile, dome-shaped turret ...
```

```
[Visual Summary]
    An image of a T-72.  It features a low-profile, dome-shaped turret with
    a prominent IR searchlight mounted ...
```

```
[Visual & Differences]
    An image of a T-72b3. It has the classic T-72 low-profile turret but
   with updated armor blocks  ...
```

Figure 4.1: Input prompts for the classes "T-72" and "T-72b3", generated using an LLM and web-scraped data. The marked text highlights the specific semantic focus of each prompt, from simple country-of-origin context to detailed visual summaries and comparative descriptions between different vehicle types, and based on whether or not it utilized the web-scraped metadata as knowledge source.

The method operates through two parallel enrichment streams at inference time:

1. **Image Enrichment:** The input image is encoded and used to perform image-to-text retrieval, finding semantically similar captions from the database.

2. **Class Enrichment:** The textual class labels are used to perform text-to-text retrieval, gathering captions relevant to each class.

The embeddings of the retrieved captions are then fused with the original image and class embeddings, respectively, to create enriched representations for the final classification decision.

Implementation Steps. Reproducing the CoRE method involved several key steps mentioned on the official github page, where we first constructed retrieval indices for the CC12M and COYO-700M datasets. Following the original method, this involved encoding the text corpora with two separate models: a **SigLIP** encoder for the image-to-text index and an **SFR-Embedding-Mistral** encoder for the text-to-text index. We then used FAISS to index these embeddings for efficient similarity search.
During inference, we used these indices to retrieve the top-k captions for both the input image (via image-to-text search) and each candidate class label (via text-to-text search). The embeddings of these retrieved captions were weighted by similarity scores and then fused with the original image and class embeddings using tunable hyperparameters (α, β, and temperature) to create the final enriched representations for classification. A detailed visualization and explanation of CoRE's architecture is highlighted in Figure 2.1. To validate our implementation, we successfully reproduced the results on the paper's original benchmark, as detailed in Section 6.2.1.

4.2.2 What Do You See? (WDYS)

The WDYS method, proposed by Abdelhamed et al. (2025), is a training-free, generation-based approach that enhances zero-shot classification by leveraging the Multimodal Large Language Model (M-LLM) of Gemini [3]. The method aims to augment the query image's representation with textual information generated directly from the image itself by asking the M-LLM the question of "What do you see (in the image)?". For a given input image, the M-LLM is prompted to perform two tasks: **(i)** generate a detailed textual description of the image's content, and **(ii)** provide an initial prediction of the image's class from the dataset's list of candidate classes. The original image, the generated description, and the initial prediction are then independently encoded into a shared embedding space using a pre-trained Vision-Language Model classifier like CLIP. Finally, these three feature vectors are fused, by averaging, to create a single, enriched query feature, which is then used for classification against the class label embeddings using a linear classifier. In Figure 4.2 we show their proposed architecture. In this work, we reproduce their method to serve as a state-of-the-art generation-based baseline and to investigate whether providing the classifier with an initial prediction made by the M-LLM would increase performance in low-resource classification.

Implementation steps. In order to reproduce their method, we closely followed the author's implementation on Github: `WDYS`. The process involved formatting our custom dataset and utilizing our selected M-LLM Qwen2.5-VL to generate both an initial class prediction (Prediction Feature) and a detailed image description (Description Feature) for each image in our test set. The generated features were then fused with the original image features and used in the final classification to evaluate the method's performance, of which the results are presented in Section 6.2.2. In addition to their main method of the WDYS architecture, we also utilized the author's perturbation-based attention method for visualizing the feature contributions most important for the model's made prediction. This technique was used in our preliminary prompt tuning experiments to analyze the model's focus on different parts of the image and text inputs. The results from these interpretability experiments are highlighted in Section 6.1.

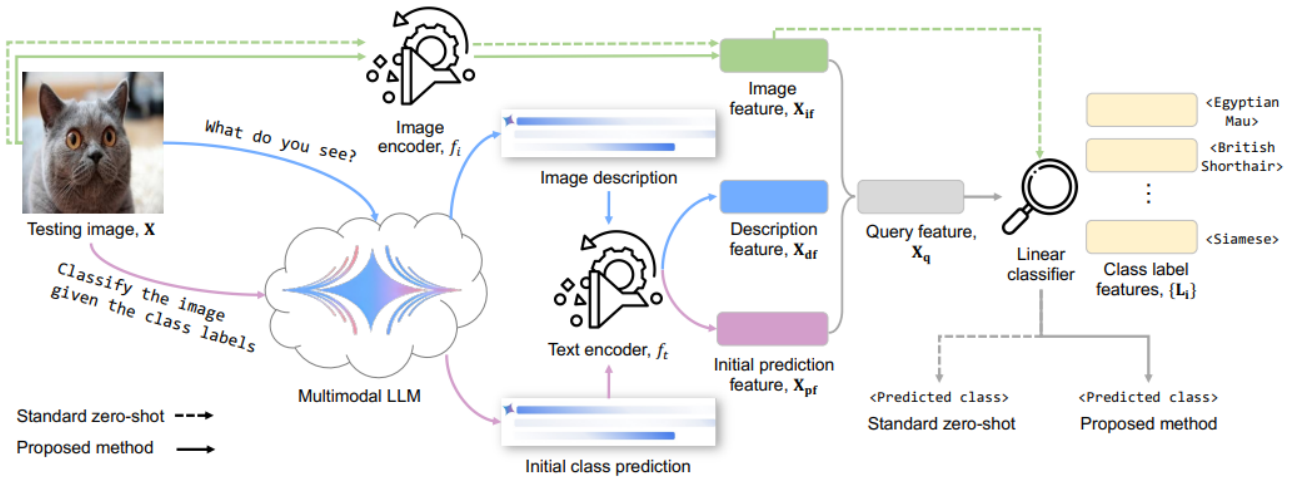


Figure 4.2: Zero-shot image classification method proposed by Abdelhamed et al. (2025), leveraging a multimodal Large Language Model (M-LLM) to generate a description and initial class prediction for a query image. They encode this data along with the input testing image using a cross-modal embedding encoder to project the inputs into a common feature space. Finally, the generated features are fused to produce the final query feature which is utilized by a standard zero-shot linear image classifier to predict the final class [3].

4.3 Proposed method: SynCE

To overcome the coverage bottleneck of retrieval and the architectural imbalance of existing generative methods that perform one-sided enrichment, we replace static, index-dependent captions with *synthetically generated* ones. We call this training-free, generation-based approach **SynCE** (Synthetic Caption Enrichment).

Overview. SynCE synthesizes concise, attribute-focused descriptions on both sides of the classifier: (i) **class-side captions** using an LLM, and (ii) **image-side captions** using a VLM with captioning capability. These captions are encoded by the text encoder of our VLM classifier and fused with the original image and class embeddings at inference time, causing the method to maintain its zero-shot capabilities.

Pipeline.

1. **Class-side generation.** For each certain dataset, we prompt an LLM, such as GPT-4o or Qwen-Instruct, to produce a list of attributes useful for identifying images in the specific domain, following the approach of Saha et al. (2024) [30] and Ren et al. (2023) [28] who find that prompting an LLM for a list of visual attributes results in the best performance. Following this, we ask the LLM to produce short sentences with values of the attributes for each of the candidate classes. The following prompts are given to the LLM:

```
What are the attributes useful for identifying images of military
vehicles (tanks, APC, armoured fighting vehicle)

Provide a list of all attributes.
```

Prompt 1: Attribute Generation

```
Provide short sentences with values of these attributes for the
following vehicles. Give {num_captions} short sentences, making sure
each sentence is different.
```

Prompt 2: Description Generation

2. **Image-side generation.** For each query image, a caption-capable VLM, such as Qwen2.5-VL [36] or PerceptionEncoder [6], is prompted to generate a description of its most important visual features. To ensure this description is structurally similar to the class-side captions, we utilize a few-shot prompting technique that provides the VLM with a small number of examples from the previously generated class-side captions within its prompt. By providing the model with these examples, we ensure it produces image captions that follow the same attribute-focused format. The thought behind this is that this structural similarity will cause the description of an image to be more similar to the description of its correct class, causing it to draw these embeddings more together and thus strengthening the classification performance. The prompt given to the VLM is approximately structured as follows:

```
You are a military vehicle expert
Here are some example captions, follow the same structure:

+ {ex} for ex in fewshot)
+ Generate {num_captions} DIFFERENT short captions
({max_words} words each) for the vehicle in the image below.
```

VLM Prompt: Caption Generation

3. **Similarity scores.** The generated captions for each side are combined via a dynamic weighting scheme. The weights are built with the similarity scores of each caption, conducted by taking the cosine similarity between the original input (class or image) and each caption, which are then converted into a probability distribution by using a softmax function controlled by a temperature parameter (τ). This allows us to control the influence of each description on the final fusion, with separate temperatures for the class- and image-side inputs (τ_{t2t} for text-to-text generation and τ_{i2t} for image-to-text generation). A low temperature skews the distribution and gives more weight to the most similar captions, while a high temperature provides a more uniform distribution across the generated captions.

4. **Encoding.** We pass all captions through the same text encoder of the used VLM classifier to keep everything in a shared space.

5. **Fusion at inference.** The final step combines the generated captions with the original representations to create the enriched embeddings. This is done by merging the embeddings controlled by the fusion hyperparameters α and β. For each class candidate, the enriched class representation C^+ is calculated with:

$$C^+ = \alpha C^G + (1 - \alpha)C$$

where C is the original class prompt embedding, C^G is the weighted embedding of the generated class captions, and $\alpha \in [0, 1]$ is the class-side fusion weight parameter. In the same manner, the enriched representation for the query image, I^{q+}, is calculated with:

$$I^{q+} = \beta I^G + (1 - \beta)I^q$$

where I^q is the original query image embedding, I^G is the weighted embedding of the generated image captions, and $\beta \in [0, 1]$ is the image-side fusion weight parameter. The final prediction is then computed by performing simple zero-shot classification, using the similarity scores between these enriched representations.

Why this helps.

- **Domain alignment.** Captions are specifically generated for the candidate classes and input images, which avoids the mismatch and noise seen in the retrieved web captions.
- **Fine-grained focus.** Prompts allow us to focus on specific subtle fine-grained details of different classes, which are the cues needed to discriminate between near-duplicate variants. The flexibility we gain by being able to alter the prompt input allows us to focus on different specifics we deem important.
- **Simplicity and Scalability.** SynCE is training-free and deploys out of the box, while no large external index is required. When applying the method on a new domain, we only need to specify this domain in the input prompt while no data collection is needed.

- **Model flexibility.** SynCE is compatible with a wide range of different captioning models (e.g. Qwen, PE-Lang) and classifiers (e.g. CLIP, SigLIP, PE). As the underlying quality of the LLM & VLM generators and classifiers improve, SynCE benefits accordingly.

Architecture. In Figure 4.3 we highlight the architecture of SynCE. Note the differences between CoRE's architecture shown in Figure 2.1 and our proposed method: the former retrieves captions from a fixed external database $\mathbb{D}$, while the latter synthesizes the captions on the fly and fuses them with the original representations.

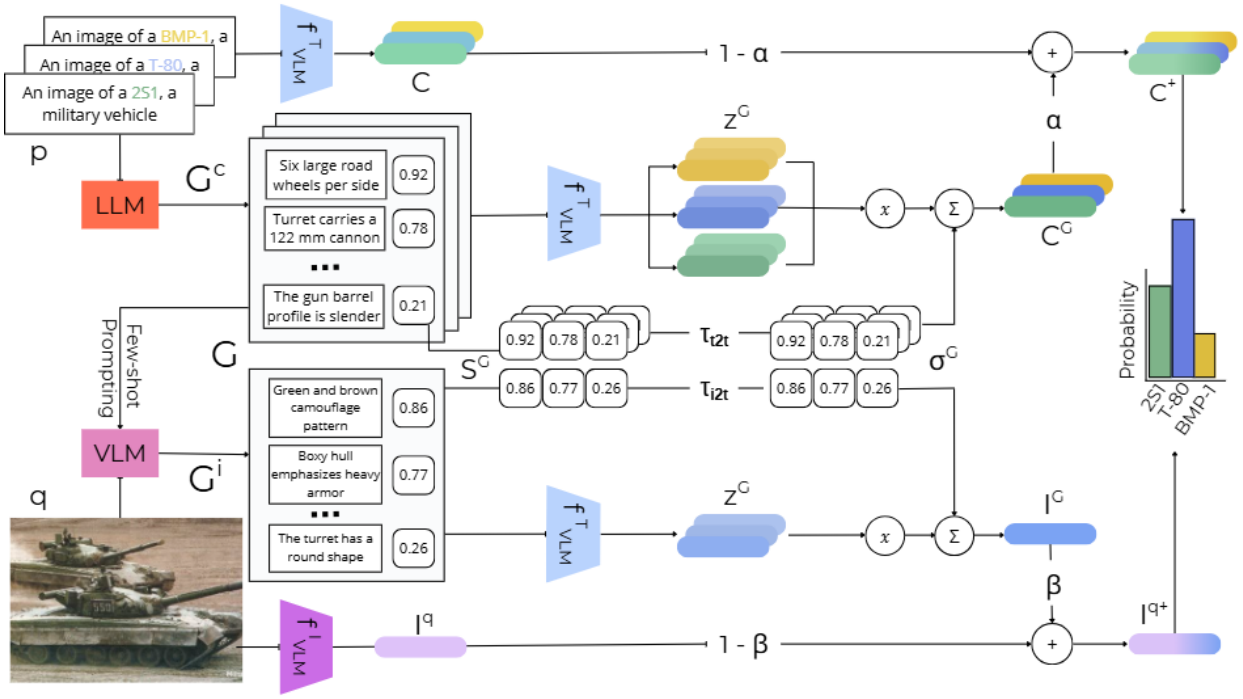


Figure 4.3: Architecture of the generation-based enrichment method of SynCE, which replaces static, index-driven captions with synthetic, domain-aligned descriptions. The method enriches both the image embedding I^q and the class prompt embeddings C with the synthetically generated captions G. The class-side captions G^c are produced by an LLM, while the image-side captions G^i are generated by a VLM using a few-shot prompting strategy with examples from G^c to ensure structural alignment. We weight these generated captions with their similarity scores S^G, which are skewed with controllable temperatures τ_{i2t} and τ_{t2t}. By combining the generated caption embeddings C^G and I^G with the original representations C and I^q through the parameters of α and β, we obtain the enriched representations C^+ and I^{q+} which are employed for zero-shot classification. Closely based on the architecture proposed by Asen et al. (2024) [11].

Practical notes. Both the manner of prompting for caption generation, by using short, structured prompts, and the weighting techniques turned out to be of high importance. When captions are low-confidence or off-topic, the similarity scores down-weight the less relevant captions based on the temperature parameter, causing the model to not overvalue the captions that are misaligned or contain noise. We evaluate SynCE against zero-shot baselines, WDYS and CoRE in the following sections.

Chapter 5

Experimental Setup

5.1 Datasets

In this thesis we consider two primary, custom-built low-resource datasets focused on military vehicles. Additionally, we utilize several public datasets in order to validate our baseline implementation and report our method on slightly different domains.

5.1.1 Primary Datasets

Our main experiments are centered on two specialized, military focused, low-resource datasets that reflect real-world challenges in fine-grained classification tasks. These datasets are the following:

Military-Vehicles-18.

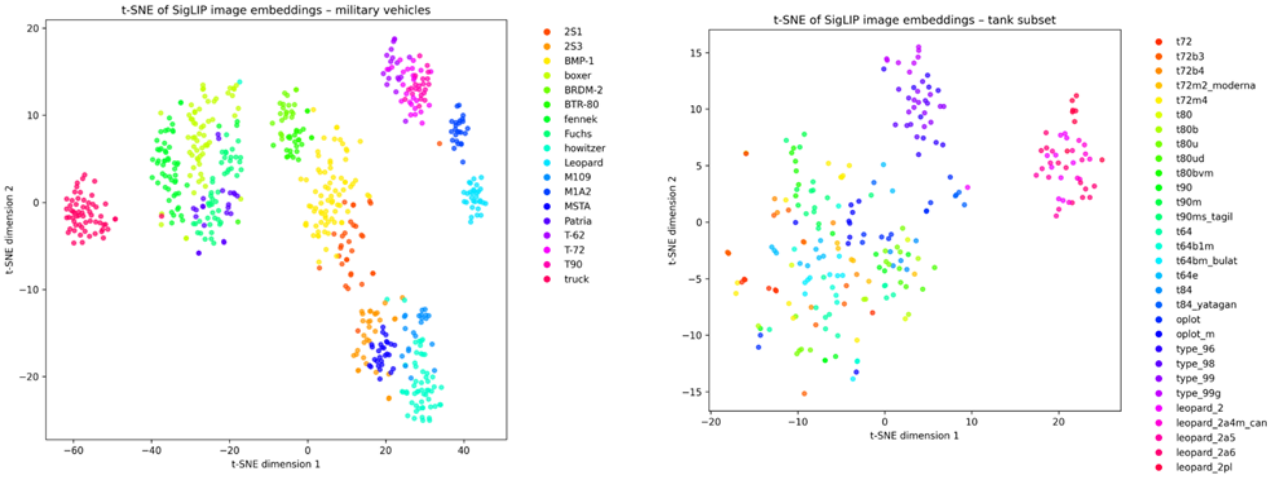
Our main evaluation is reported on our **military-vehicles-18** dataset, which consists of 18 distinct military vehicle classes. This dataset was chosen based on its variety of both fine-grained and more coarse-grained classes, the differences in country of origin (useful for highlighting practical challenges) and the fact that it is used as a benchmark for image classification at TNO. It contains a mix of vehicles with both subtle, fine-grained differences (e.g., between the T-62 and T-72 main battle tanks) and more coarse-grained distinctions (e.g., between a Leopard tank and a BTR-80 armored personnel carrier). The t-SNE plot of the image embeddings in Figure 5.1a shows clusters of relatively distinct and similar classes, making it a suitable test set for measuring the performance of our method. Important to note is that much of the imagery in this domain is not widely available on the public web due to it being military imagery and thus being classified, making it an ideal low-resource scenario and being the cause for retrieval-based methods performing worse.

Tanks-30.

For our preliminary prompt tuning experiments, we utilized the more challenging **tanks-30** dataset. This set of images contains 30 classes of main battle tanks, causing it to be an extremely fine-grained classification task where distinctions often rely on subtle variations in chassis, turret design, or specification of road-wheels. The difficulty of this dataset is illustrated by the t-SNE visualization in Figure 5.1b, where the significant overlap between the image embeddings of different classes highlights the challenge for the VLM classifier.

5.1.2 Data Partitioning

For both the **military-vehicles-18** and **tanks-30** datasets, we use a fixed data split to ensure consistent and reproducible evaluations. Given their low-resource nature and our limited data, we allocate 75% of the images to the **test set** for final evaluation. The remaining data selected as the 'training' set (25%), where all model hyperparameters are tuned exclusively on the training set and the final performance is reported on the unseen test set. This partition was chosen such that we don't overfit the hyperparameters on our data, causing it to not generalize well when new classes are introduced. Note that while it is a truly low-resource task, the **military-vehicles-18** dataset contains 660 images, whereas the **tanks-30** dataset contains only 250 images in total.



(a) t-SNE plot of the military-vehicles-18 image embeddings

(b) t-SNE plot of the tanks-30 image embeddings

Figure 5.1: t-SNE plots of the military datasets introduced in this thesis.

5.1.3 Public Benchmark Datasets

In order to validate our reproduction implementation of CoRE [11], which is used as a baseline to our proposed method, we reproduced their results on the public benchmark mentioned in their paper. This contains datasets where retrieval-based enrichment is effective, such as the Circuits dataset from Zhang et al. (2024) [41], the HAM10000 medical imaging dataset [35] and the subset of rare plants from iNaturalist [16]. For the primary evaluation of our proposed method, SynCE, we then focused specifically on the datasets known to be challenging for retrieval-based approaches due to poor coverage in the database: Patch-Camelyon [22], a medical dataset, and Parasitic Egg Classification [5], a dataset of rare egg parasites. The performance of our proposed method on these public benchmarks is presented in section 6.3

5.2 Models and Implementation Details

This section outlines the specific models, software libraries, and hardware resources used to conduct the experiments in this thesis.

5.2.1 Encoder Models

We employ several encoder architectures for classification and retrieval, selected based on their relevance to each experimental stage.

- For the preliminary prompt tuning experiments, we utilized two publicly available CLIP variants: **CLIP_H_DFN** and **CLIP-L_datacomb**.
- For the reproduction of the CoRE baseline, we used the **SigLIP** [38] model as classifier and image-to-text retrieval, and **SFR-Embedding-Mistral** [2] for text-to-text retrieval, following the original paper's implementation. The SigLIP model was also employed as one of the two main classifiers for our proposed SynCE method.
- For the reproduction of the WDYS baseline and for our proposed SynCE method, we employed the *Large* variant of the **PerceptionEncoder** [6]: PE-L. This model served as the second main classifier for our SynCE method, allowing us to evaluate our enrichment strategy across different VLM architectures.

5.2.2 Description Generation Models

For the generation of textual descriptions, we utilized LLMs for class-side generation and VLMs for image-side captioning. To determine the optimal models for our framework, we evaluated several candidates.

- **Class-Side (LLM):** To generate attribute-based descriptions for the class labels, we evaluated two models: OpenAI's **GPT-4o** (via its API) and the open-source **Qwen2.5-32B-Instruct**. Based on its improved performance in preliminary experiments, GPT-4o was selected as the LLM for the main results reported in this thesis.
- **Image-Side (VLM):** To generate captions for the input images, we compared two state-of-the-art captioning VLMs: **Qwen2.5-VL-32B-Instruct** and **PerceptionEncoder-Lang-L14-448**. Following the results of our preliminary comparison, Qwen2.5-VL was chosen as the primary image captioner for our main experiments due to its more effective performance on our target dataset.

5.2.3 Implementation and Hardware

All experiments were implemented in Python using PyTorch and the Hugging Face Transformers library for model access. The computations were performed on a single server equipped with an **NVIDIA A40 GPU** containing 45GB of VRAM.

5.3 Evaluation Metrics

To assess performance of our implemented methods, we use a standard set of evaluation metrics. Our primary metric is **Top-1 Accuracy (Acc@1)**, which measures the percentage of instances where the model's highest-probability prediction is correct. We also report **Top-5 Accuracy (Acc@5)**, which is particularly relevant for fine-grained tasks, counting a prediction as correct if the true label is among the top five predictions. To account for potential class imbalances, we additionally compute **Macro Precision** and **Macro Recall**, which evaluate the model's completeness by averaging the per-class scores, ensuring each class contributes equally.

Chapter 6

Results and Analysis

6.1 Preliminary Study: Prompt Tuning

Following the methodology described in section 4.1, we summarize the outcomes of our prompt tuning experiments in Table 6.1. From this table we can conclude that CLIP's zero-shot accuracy proved to be highly **prompt-sensitive**: small wording changes were often the cause for large swings in the classification performance. This conclusion aligns with Roth et al. (2023) [29], who in their paper show that adding random, non-semantically relevant words to the caption can increase CLIP's performance, and with Yan et al. (2023) [37], who show that using LLM generated attributes in a large quantity perform almost the same as random words. In both cases, it caused the authors to question whether improvements in accuracy truly rely on the semantic understanding of the model or if it might actually come from an ensembling effect, where averaging over multiple noisy or random prompt variations creates a more robust representation of the class than truly semantically relevant inputs. Notably, simpler prompts frequently outperformed more complex, meta-driven ones. For example, the basic "Country" augmentation, where only the country of origin is added to the input prompt, often outperforms longer or metadata-heavy captions, meaning that more detailed text is often the cause for decreasing performance. These results suggest that it is difficult to craft a single, robust description for each class that improves classification across models.

To better understand these large fluctuations in performance, we analyzed the model's attention on both the image and text inputs using the visualization method proposed by Abdelhamed et al. (2025) [3]. This attention technique is a perturbation-based approach that identifies the most important factors in the made prediction by masking out different parts of the input. For the image, they utilize a sliding kernel that masks out different patches of the input, noting whether the model's prediction is changed when a certain patch is masked out. Masked out patches with the highest drop in performance are considered to have the highest contribution to the prediction, giving us a relevant attention heatmap. A similar process is applied to the textual input prompt, where a sliding window masks words to determine their importance in the made prediction. As shown in Figure 6.1, the model's attention on the image-side consistently highlights relevant and logical regions of the image, such as the vehicle's turret, hull and and wheels. In contrast, the caption-side attributions seem unreliable and vary widely across different prompts. The model seems to draw its prediction from incidental words or random correlations rather than features that are semantically relevant to the vehicle it is describing.

From these fluctuating results and inconsistent attention patterns on the different parts of the input text, we can conclude that simple prompt tuning is an unreliable path to receive robust performance gains. These findings caused us to explore more systematic and reliable

strategies for performance increase, such as the methods of information enrichment.

Table 6.1: Zero-shot Top-1 accuracy (%) of two CLIP variants under different description types on our tank-30 dataset. "Country" adds only country-of-origin, "Medium/Long" vary caption length and the last three rows involve the semantic focus of the descriptions or the knowledge source.

Description type	**CLIP_H_DFN**	**CLIP-L_datacomb**
None	13.31	15.32
Country	**20.56**	18.95
Medium	17.74	**21.37**
Long	17.74	18.95
Visual Summary	18.55	15.73
Visual & Differences	18.95	12.10
Metadata + LLM Knowledge	18.55	20.97

(a) Contribution visualization for an image of `t72` with the caption of the (incorrectly) predicted `t64m_bulat`.

(b) Contribution visualization for an image of `t72` with the caption of the (incorrectly) predicted `t80b`.

Figure 6.1: Attention visualizations for two examples. Image-side highlights are generally logical. Caption-side contributions are unstable, making it hard to identify consistently helpful semantic cues.

6.2 Baseline Performance

In this section, we evaluate the reproduced baseline methods explained in section 4.2 on our military vehicle datasets to conclude their performance in our specific low-resource, fine-grained domain and be able to compare it as a baseline to our proposed method.

6.2.1 Combination of Retrieval Enrichment (CoRE)

In order to validate our implementation of CoRE, we reproduce the original paper's results on their low-resource datasets, shown in Table 6.2. The reproduced accuracies closely match the reported numbers, confirming that our implementation was done correctly.

Table 6.2: CoRE reproduction results on the Top-1 and Top-5 Accuracy metrics

Dataset	Original	Reproduced	Difference
Circuits @1	42.94	43.14	+0.2
Circuits @5	67.71	68.92	+1.21
iNat @1	19.10	20.34	+1.24
iNat @5	45.70	41.72	-3.98
HAM10000 @1	61.54	59.23	-2.31
HAM10000 @5	95.70	96.29	+0.59

Following this validation, we applied CoRE to our military vehicle datasets. As shown in Table 6.3 and Table 6.4, CoRE failed to improve upon the domain zero-shot baseline for our military-vehicles dataset and very slightly outperformed the zero-shot baseline for our tanks dataset.

Table 6.3: Performance of SigLIP zero-shot versus the CoRE method on the **military-vehicles-18** dataset. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle".

Method	Accuracy @1	Accuracy @5
SigLIP (Generic)	48.78	85.11
SigLIP (Domain)	**75.53**	**98.94**
CoRE (CC12M)	74.51	98.87
CoRE (COYO-700M)	74.29	98.55

Table 6.4: Performance of SigLIP zero-shot versus the CoRE method on the **tanks-30** dataset. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle".

Method	Accuracy @1	Accuracy @5
SigLIP (Generic)	17.74	52.02
SigLIP (Domain)	16.94	**55.24**
CoRE (CC12M)	17.56	52.65
CoRE (COYO-700M)	**18.55**	51.61

This outcome aligns with the limitations stated by CoRE's authors: retrieval-based methods are highly dependent on the coverage of the target domain within the retrieval database [11]. Military vehicle imagery, particularly for specific variants, is underrepresented in public web-crawled datasets due to the fact that images of certain military vehicles are often classified and only a very limited amount appears in public databases. Consequently, the retrieved captions were often too generic (e.g., "an army tank"), described the wrong vehicle type (e.g. "T-90" instead of "T-80"), or were off-domain entirely (e.g., describing a Tesla Roadster vehicle instead of a Military Tank). Enriching the embeddings with this irrelevant information diluted

the original representations, drawing them away from the correct class and causing a drop in accuracy. These results confirm that retrieval-based enrichment is not a suitable strategy for our domain, underscoring the need for an alternative approach.

Interestingly, for the dataset *military-vehicles-18* we see a decrease in performance when retrieving captions from the larger COYO-700M database compared to the smaller CC12M database. Even though a larger index would be generally expected to improve concept coverage, two factors likely explain this outcome. First, it is important to note that due to computational constraints, we only utilized a 10% subset of COYO-700M. This partial index might not properly represent the full database and could have even less coverage of our niche domain in that specific 10% than the complete CC12M database. Secondly, larger web-crawled databases could introduce even more noise into the retrieval method. The increased scale may have led to the retrieval of a higher proportion of irrelevant captions to the specialized dataset, causing the performance to suffer.

Figure 6.2 provides an example of class- and image-side retrieved captions for a specific class in our dataset. These captions highlight the issues stated above, where the marked text showcases that the retrieved captions often describe the wrong vehicle type, not mentioning the correct class at all due to its under-representation in the retrieval database.

```
"t80": {
  "captions": [
    "Side view of a military war tank. Vector stock illustration",
    "WWII Tanks-Art Print displayed on a wall",
    "1:72 Germany capture the soviet t34-76 medium tank model",
    "Vietnam is considering the purchase of 28 T-90MS modern tanks",
    "The Terrier a tracked armoured engineer vehicle"
  ]
}
```

Listing 6.1: Retrieved class-side captions for the class "t80".

```
t80
- Vietnam is considering the purchase of 28 T-90MS modern tanks
- The Russian T-90 tank interested in Indonesia
- A left front view of an M-551 Sheridan tank modified to look like a
   Soviet tank.
```

Listing 6.2: Retrieved image-side captions for an image of a T-80 tank.

Figure 6.2: Examples of irrelevant captions retrieved by the CoRE method for the T-80 tank class. The highlighted text (in red) shows how the retrieval process frequently returns captions describing different vehicle types (e.g., The Terrier, M-551 Sheridan) or related but incorrect variants (T-90), demonstrating the bottleneck of using a generic web-crawled database for a specialized domain.

6.2.2 What Do You See? (WDYS)

The WDYS method proposed by Abdelhamed et al. (2025) [3] proved to be ineffective in our domain. The approach heavily relies on the M-LLM's ability to provide an accurate initial

prediction and a descriptive caption. However, due to the specialized and fine-grained nature of our dataset, the M-LLM's initial zero-shot classification accuracy was below 50% as seen in Table 6.5. Even though recent M-LLMs like Qwen2.5-VL and Gemini are able to produce accurate and detailed captions when given an input image (even in the low-resource domain), it proved to be incapable of correctly identifying the class when given a list of candidate classes due to the subtle fine-grained differences between the images and the lack of online training data in this domain.

Table 6.5: WDYS ablation study on *Military-Vehicles-18* using the **PerceptionEncoder-Large** classifier, comparing combinations of the Image (IF), Description (DF), and Prediction (PF) features. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle"

Method	**Acc@1**	**Acc@5**
PE-L (Generic)	52.39	86.42
PE-L (Domain)	**73.28**	**96.96**
DF	47.15	74.50
PF	48.48	63.98
DF and PF	46.22	75.31
DF and IF	43.01	81.76
PF and IF	50.61	77.20
DF, PF and IF	47.87	77.96

Table 6.6: WDYS ablation study on *Military-Vehicles-18* using the **SigLIP** classifier, comparing combinations of the Image (IF), Description (DF), and Prediction (PF) features. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle".

Method	**Acc@1**	**Acc@5**
SigLIP (Generic)	48.78	85.11
SigLIP (Domain)	**75.53**	**98.94**
DF	40.50	71.12
PF	41.34	72.49
DF and PF	40.11	70.05
DF and IF	41.64	67.63
PF and IF	48.02	77.96
DF, PF and IF	45.14	74.01

It is important to highlight the difference between the findings of our domain on their method with the original results reported in the WDYS paper, where the combination of all three features (DF, PF and IF) resulted in significant performance improvements. A main reason for this difference comes from the performance of the M-LLM's initial prediction (PF), which is substantially higher on the author's high-resource benchmarks than on our low-resource dataset. While their initial prediction provides a strong and relevant signal, enhancing the embeddings and classification performance, our inaccurate Prediction Feature, with an accuracy below 50%, introduces noise that counteracts the benefits of the other features, leading to a degrading overall performance.

An additional important difference lies in the feature fusion strategy, where the authors of WDYS utilize simple averaging to combine the three feature inputs, assuming that each feature contributes equally. Even though it is effective in their context, it is not ideal in our domain while it allows the 'noisy' features to degrade the quality of the final input embedding. Following the approach of CoRE [11], we found that replacing the simple averaging strategy with a weighted fusion, including similarity scores for each description, causing us to have full control over the contribution of each features, yielded significant improvements over the WDYS baseline and highlighted the importance of adaptive feature weighting in our low-resource domain.

Due to the inability of the retrieval methods to capture relevant and accurate captions on the classes and images from certain datasets like ours, and the ineffectiveness of the WDYS method in low-resource domains, we were prompted to build a novel method that combines synthetically generating captions truly relevant to the inputs while maintaining focus on the low-resource domain by enriching both the image- and class-side representations in order to better align the image with its correct class in the embedding space via adaptive feature weighting.

6.3 Main Results: SynCE performance

We now evaluate our proposed method of **SynCE (Synthetic Caption Enrichment)**, which is a training-free framework that synthetically generates captions to enrich both the image and class representations. In the following sections, we compare the performance of three different configurations: enriching only the image-side, only the class-side, or a combination of both on different VLM classifiers and datasets.

6.3.1 Quantitative Comparison

Military Vehicles

The main results of our proposed method SynCE on the *Military-vehicles-18* datasets are shown in Table 6.7 for the PerceptionEncoder classifier and Table 6.8 for the SigLIP classifier. These tables highlight the performance of different SynCE configurations against the zero-shot and reproduction baselines. The configurations are based on whether the image-side, the class-side or a combination, is enriched at inference time.

The results show an interesting relationship between the performance of the enrichment strategy and the VLM's underlying architecture. Shown in Table 6.7, we see that enriching only the class-side for the PerceptionEncoder yields the highest Top-1 accuracy of **75.99%**, slightly outperforming the combination of dual enrichment **(74.01%)** and the method of CoRE **(73.45%)**. While both configurations outperform the already strong domain-specific baseline of **73.28%**, our results suggest that for this model the primary benefit comes from refining and enriching the class prototypes.

Shown in Table 6.8, the results for the SigLIP classifier present a different pattern. Here we see that the combination of both image- and class-side enrichment achieves the best performance across all metrics. We note a Top-1 accuracy of **76.44%**, which is an improvement over the domain-specific baseline of **75.53%** and the CoRE implementation of **74.51%**. Interestingly, we see also a performance gain when enriching only the image-side, whereas enriching only the class-side results in a slight degradation in performance compared to the baseline.

A possible explanation for the differences in the specific performance increase between the models may lie in the architectural properties of the VLM encoders. We note that the PerceptionEncoder is known for its powerful vision side, whose image features may already be tightly aligned with the class's textual concepts. Following from this, enriching the image-side representations provides degrading performance while it might introduce noise and draw the correct class away from the image embedding, whereas enriching the class-side prompts enhance the separation between the class prototypes in the embedding space without having impact on PE's already strong image features. SigLIP however may benefit more from the dual-enrichment strategy, while the approach pulls both the image and class embeddings towards a more aligned and context-aware representation, correcting for any weaker initial alignment between the embeddings.

Even though the SynCE method showcases a consistent improvement over the baselines, it is important to note that the performance gains are not substantial. Additionally, we observed a sensitivity in performance to the hyperparameters, highlighting that the robustness of the enrichment process could be optimized further. We hypothesize that the limitations could come from two main sources: a potential structural or semantic miss-alignment between the VLM-generated image captions and the LLM-generated class descriptions, which could be improved further, and possible noise in the form of factual inaccuracies or broad concepts on the class-side descriptions. These issues could have a negative impact on the enrichment method and in some cases prevent alignment between an image and its correct class, and are further discussed in our qualitative analysis in Section 6.3.3. The constraints show clear opportunities for further improvement and are highlighted for future work in Section 8.2.

Table 6.7: SynCE **PerceptionEncoder-Large** Results on *Military-Vehicles-18*. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle". The LLM and VLM generators used are GPT-4o and Qwen2.5-VL, respectively.

Method	**Acc@1**	**Acc@5**	**Precision**	**Recall**
PE-L (Generic)	52.39	86.42	53.30	55.61
PE-L (Domain)	73.28	96.96	70.56	71.76
CoRE (CC12M)	73.45	97.11	71.25	72.31
CoRE (COYO-700M)	72.99	96.82	70.88	71.90
WDYS	47.87	77.96	52.09	46.56
Image-side	70.82	98.28	68.41	69.52
Class-side	**75.99**	97.87	**74.13**	**75.20**
Combination	74.01	**98.33**	71.86	72.95

Patch-Camelyon

The Patch-Camelyon dataset presents a binary classification task: identifying the presence of metastatic tumor tissue in histopathologic scans of lymph nodes. For our baseline, we used simple prompts ("An image of a Tumor" or "An image of healthy tissue"). For SynCE, the class-side captions were generated to describe the key visual characteristics of tumorous versus healthy tissue, while the image-side caption described the content of the specific scan.

Table 6.8: SynCE **SigLIP** Results on *Military-Vehicles-18*. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a military vehicle". The LLM and VLM generators used are GPT-4o and Qwen2.5-VL, respectively.

Method	**Acc@1**	**Acc@5**	**Precision**	**Recall**
SigLIP (Generic)	48.78	85.11	53.54	54.19
SigLIP (Domain)	75.53	98.94	79.19	77.23
CoRE (CC12M)	74.51	98.87	78.50	77.88
CoRE (COYO-700M)	74.29	98.55	78.12	77.32
WDYS	45.14	74.14	58.89	45.17
Image-side	75.59	99.24	79.49	78.44
Class-side	74.38	98.94	78.90	76.82
Combination	**76.44**	**99.39**	**79.73**	**78.80**

As shown in Table 6.9, SynCE provides a clear improvement over the strong SigLIP baseline. While enriching the class-side alone yields a notable increase in accuracy, the best performance is achieved with the **Combination** of both image- and class-side enrichment. This dual-enrichment strategy achieves a Top-1 Accuracy of **78.18%**, a significant gain of over 8 percentage points compared to the baseline's 69.74%. This result demonstrates that even in a binary, medical imaging context, SynCE's method of generating and fusing descriptive captions can substantially enhance classification performance.

Table 6.9: SynCE SigLIP Results vs SigLIP Baseline, Patch-Camelyon

Method	**Acc@1**	**Precision**	**Recall**
SigLIP	69.74	70.24	69.74
Image-side	66.67	71.37	67.69
Class-side	74.00	74.78	73.50
Combination	**78.18**	**78.17**	**78.06**

Parasitic Egg

The Parasitic Egg dataset is an extremely fine-grained and low-resource challenge. Our results, presented in Tables 6.10 and 6.11, highlight the performance of our method on this data.

Among the training-free methods, our SynCE **Combination** approach again achieves the highest Top-1 Accuracy of **16.89%**, outperforming the SigLIP baselines and the CoRE method (15.14%). However, when compared to supervised fine-tuning, performed by the authors of CoRE [11], the results highlight the difficulty of this domain. As reported in their paper, a fine-tuned ImageBind model significantly outperforms our method with an accuracy of 33.27%.

Nonetheless, our training-free SynCE method achieves a competitive accuracy of 16.89%, which approaches the 17.59% performance of a fine-tuned SigLIP model. This highlights that for this particularly challenging dataset, supervised fine-tuning remains the most effective method if the labeled data is available. However, SynCE's ability to achieve comparable performance to a fine-tuned SigLIP model without requiring any training data underscores its strength as a powerful training-free alternative.

Table 6.10: Performance of SynCE SigLIP versus the SigLIP baseline on the Parasitic Egg dataset. The (Generic) baseline uses a standard prompt, while (Domain) adds the context "a parasitic egg".

Method	Acc@1	Acc@5	Precision	Recall
SigLIP (Generic)	12.61	54.94	13.83	12.61
SigLIP (Domain)	11.18	**61.04**	**21.27**	11.19
Image-side	15.63	55.53	14.04	15.27
Class-side	13.97	55.53	15.48	13.56
Combination	**16.89**	54.99	14.76	**16.50**

Table 6.11: SynCE SigLIP vs Fine-Tuned results found by CoRE's authors [11], Parasitic Egg Dataset

Method	Acc@1
ImageBind (FT)	**33.27**
SigLIP (FT)	17.59
ImageBind	9.18
SigLIP	12.72
CoRE (CC12M)	15.14
SynCE	16.89

6.3.2 Ablation Studies

To evaluate the design choices of our SynCE method and find the most optimal configuration, we ran several ablation studies on the parameters. First we investigate the sensitivity of our model to the fusion parameters of α, for the class-side enrichment, and β, for the image-side enrichment. We perform a parameter sweep with a step-size of 0.01, noting the impact on the Accuracy performance in our military vehicles dataset. In Figure 6.3 we highlight the search for the optimal enrichment parameters α and β. As seen in the figure, each value for α causes the performance to remain quite consistent, even though the lower values showcase a clear improvement on the classification performance. However, as the value of β increases, the classification performance drops rapidly, and the best value for the image-side enrichment parameter is found around the low value of 0.1. This is likely due to the trade-off between the specificity of the generated image captions and their accuracy. While the VLM-generated descriptions are designed to be attribute-focused and detailed, the semantic value of the descriptions might not have as big a positive impact on the VLM classifier as initially hypothesized, especially in a fine-grained domain with visually similar classes. Additionally, as highlighted in Section 6.3.3, the VLM can misidentify subtle details, leading to inaccuracies in the generated text and additional noise to the image signal. A high fusion weight β would amplify the impact of this noise and potential errors, corrupting the representation of the original embedding. A low value of β strikes a useful balance of utilizing specific textual details to provide a helpful corrective signal without overwhelming the fine-grained visual embedding.

Based on these and further analyses, we found the optimal values of each parameter, summarized in Table 6.12. These settings were used for all final evaluations reported in this thesis.

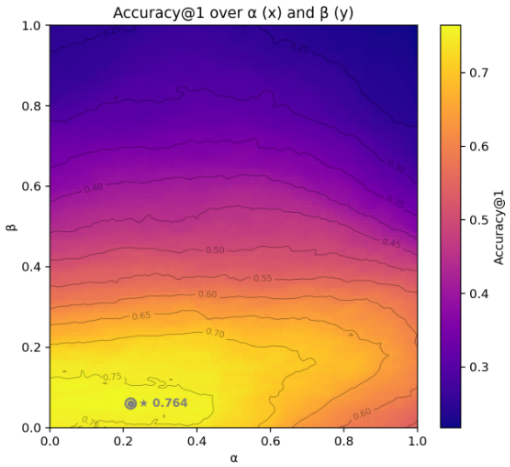


Figure 6.3: Top-1 Accuracy of SynCE SigLIP on military-vehicles-18 with a varying α and β. The best performance (0.764) is reached with relatively low values for both the image-side generated captions ($\beta \sim 0.1$) and the class-side generated captions ($\alpha \sim 0.2$).

Table 6.12: Optimal SynCE Parameters

Parameter	Optimal values
α	0.2
β	0.1
τ_{t2t}	0.04 - 0.1
τ_{i2t}	0.1
k	5
N	20 - 30
M	5

6.3.3 Qualitative Analysis

While the quantitative metrics demonstrate the effectiveness of SynCE, we additionally highlight a qualitative analysis of the generated captions to provide deeper insight into why the method succeeds and to reveal its inherent limitations.

In Figure 6.4 we compare the retrieved captions used by CoRE to the generated captions used by SynCE for the "T-90" class, where the advantage of generation in our specific low-resource domain becomes more clear. The captions retrieved with CoRE proved to be noisy and frequently irrelevant, where the class-side retrieval often returned descriptions of entirely different vehicles (e.g. "A9 tank", "M-60", "The Terrier"), while the image-side retrieval provided generic or completely off-topic sentences (e.g. "Even If You Fail While Driving A Tank – It's Still Badass"). Utilizing this irrelevant information in the enrichment process actively harms the classification performance while it dilutes the original embeddings instead of providing useful external knowledge. In contrast to these irrelevant captions, the descriptions generated by SynCE are very specific and discriminative, focusing on key visual attributes of the class such as "six road wheels per side," "Reactive armor tiles," and a "125 mm smoothbore cannon." These captions provide the VLM classifier with a relevant and powerful signal to help distinguish between the fine-grained classes. However, despite providing relevant and accurate captions, SynCE still falls short of perfect accuracy and leaves a performance gap of around 24%. This gap suggests that simply improving the caption quality for representation enrichment may not be enough to solve the entire problem.

This performance ceiling could be explained by the findings of the WaffleCLIP paper [29]. In this paper, the authors suggest that the performance gains from prompt improvements may not follow from the classifier's deep semantic understanding of the descriptive sentences, but rather from an ensembling effect of structured text or from the model latching onto specific keywords. Even for our dynamic enrichment method, this might be the case. If the VLM classifier is indeed behaving more like a "bag-of-words" model, then there could be a natural limit to how much performance can be improved by making prompts more semantically aligned with the classes, which might explain the remaining errors that our method still contains.

Figure 6.4: Comparison of Retrieved and Generated captions for "t90, a military vehicle".

```
Class
"t90": {
  "captions": [
    "Side view of a military war tank. Vector stock illustration",
    "A9 tank of the 1st Armoured Division, 2nd Armoured Brigade",
    "An M-60 battle tank moves down a road during training exercise",
    "The Terrier a tracked armoured engineer vehicle",
    "The Russian T-90 tank interested in Indonesia",
    "World of Tanks Museum of the History of the T-34 Medium tank"
  ] }
Image
"t90": {
  "captions": [
    "Vietnam is considering the purchase of 28 T-90MS modern tanks",
    "Even If You Fail While Driving A Tank -- It's Still Badass",
    "Nation heroes leading the sturdy Indian army tanks",
    "<PERSON> smash American cars with a tank! 2",
    "Air show during the Aviation and Space Salon MAKS-2019.",
    "Thick smoke lines the bottom of the sky and blocks the sunlight."
  ] }
```

Listing 6.3: Retrieved captions for the class "t90".

```
Class
"t90": {
  "captions": [
    "It has six road wheels per side",
    "The main armament is a 125 mm smoothbore cannon.",
    "Reactive armor tiles cover the turret front.",
    "Three return rollers support the upper track run.",
    "A welded turret mounts centrally on the hull."
  ] }
Image
"t90": {
  "captions": [
    "The main armament appears to be a 125 mm smoothbore gun",
    "Reactive armor tiles are attached to the front of the turret.",
    "The vehicle is equipped with advanced sensors and communication
   equipment mounted on top, including a stabilized  turret and a
   periscope.",
  ] }
```

Listing 6.4: Generated captions for the class "t90".

In Figures 6.5 and 6.6 we highlight failure cases for our SynCE method, which can be linked to two primary sources: **inaccurate caption generation** and **classifier misinterpretation**. The first occurs when the generative models produce inaccurate captions, on either the class- or image-side, where the inaccurate captions manifest differently for each side due to their distinct pre-training objectives. The class-side LLM is primarily at risk of 'hallucinating' due to the low-resource nature of the domain, while it is tasked with providing factual world knowledge. If a specific class is underrepresented in its text pre-training data, the LLM might lack the precise knowledge to generate accurate and detailed descriptions, leading to factual inaccuracies or hallucinations in the captions. In contrast, the image-side VLM generator is trained to describe "what it sees" in the image, which is a task less reliant on prior domain knowledge but rather on its ability to focus on specific details in an image. Its primary challenge is therefore not the low-resourceness but rather the fine-grained nature of the images. While for visually similar vehicles, an image taken from an unusual angle or with key features occluded, or missing, could be the cause for a VLM misidentifying a subtle but critical detail, resulting in an inaccurate caption. Captions with errors from either source could introduce noise into the enrichment process and mislead the classifier by causing the image embedding to be pulled towards the wrong class prototype in the representation space.

The second case where SynCE might fall short is when the VLM classifier misinterprets a correct signal. Even when both the class- and image-side captions are perfectly accurate and aligned, the VLM classifier's embedding space is not perfect. It is still possible for the model to make an incorrect prediction if the visual features of the images are very similar to a competing class, causing it to ignore the textual representations coming from the correct captions. These failure cases highlight the remaining challenges in fine-grained low-resource classification and show that there will be need for continued improvements in both the generative models and the VLM classifiers.

In Figure 6.5, we showcase an example of a failure case for SynCE, highlighting the parts of the captions that are incorrect and not aligned, which in this case is most likely due to the low-quality and obscureness of the image. This causes errors in the image description, diluting the classification signal. In the second example in Figure 6.6, we see that both the class- and image-side captions align, which makes us assume there exist cases of the VLM classifier misinterpreting a correct signal, causing the prediction to be incorrect. In Chapter A of the Appendix, we briefly note the results of SynCE when the class- and image-side captions are perfectly aligned, highlighting that even for correct signals, the VLM classifier can misinterpret the input based on the original image feature.

Ground Truth: `Boxer`
SynCE Prediction: `Leopard` (Incorrect)

Generated Class Captions (for `Boxer`):

- "The RWS can mount a 12.7 mm heavy machine gun."
- "It has eight road wheels per side."
- "Battlefield-management antennas sit on the roof."

Generated Image Captions:

- "Equipped with a remote weapon station featuring a 30 mm autocannon mounted on its roof."
- "The turret is designed for high mobility and firepower, with a visible thermal signature indicating the presence of advanced thermal imaging systems."

Figure 6.5: A failure case for SynCE. This misalignment in the class- and image-side captions is a potential cause for the final prediction error.

Ground Truth: `Fennek`
SynCE Prediction: `Patria` (Incorrect)

Generated Class Captions (for `Fennek`):

- "The hull is a welded steel-composite monocoque."
- "It has two road wheels per side."
- "The paint scheme is desert tan with green patches."

Generated Image Captions:

- "An armored personnel carrier with a monocoque hull design"
- "The undercarriage is smooth-sided, containing two road wheels per side for mobility"
- "The vehicle's camouflage pattern suggests it is intended for use in various environments, likely including urban combat scenarios"

Figure 6.6: A failure case for SynCE. Even though the captions are correct and aligned, the VLM misinterprets the input signal, causing an error in the prediction.

Chapter 7

Discussion

This thesis set out to investigate whether enriching the embeddings of Vision-Language Models with synthetically generated descriptions can improve fine-grained, zero-shot classification in low-resource domains. Our experiments evaluated preliminary prompt tuning, state-of-the-art baselines of retrieval-based enrichment (CoRE) and generation-based enrichment (WDYS), and our novel generation-based method of SynCE, built specifically for classification in these low-resource domains. In this chapter, we interpret and discuss the findings from these experiments presented in the previous chapter.

Limitations of Existing Methods. Our initial experiments presented in sections 6.1 and 6.2 confirmed that current approaches are often insufficient for challenges introduced when evaluating on truly low-resource, fine-grained domains such as military vehicles. Prompt tuning with an LLM and grounded data yielded unstable and only marginal performance gains. The results from this preliminary study were highly sensitive to minor changes in phrasing of the input prompt, which suggested that the improvements might be caused by incidental correlations between words rather than a truly semantic understanding of the input text. Similarly, the retrieval-based method of CoRE failed for our specialized domain due to the under-representation of this domain in the external web-crawled database. For military vehicles, this under-representation caused the method to retrieve irrelevant captions that caused the classification performance to decrease, concluding that retrieval is not a viable solution when the domain coverage is poor in the retrieval database, as mentioned by the authors who found similar results when testing it on two separate datasets [11].

As for WDYS, our experiments showed that their specific generation-based method is not well-suited for our low-resource domain since the method's performance is critically dependent on the M-LLMs ability to correctly predict an initial class prediction of the image, which it fails to do in over half the cases for our low-resource datasets, causing significant noise to be fused into the query feature and degrading the method's classification performance. Secondly, their specific method of image caption generation caused the M-LLM to lack descriptions of the fine-grained details that are important when distinguishing between similar class variants. Additionally, their method was mainly focused on enriching the image-side of the classifier, missing the opportunity to properly align it with the class-side descriptions. Because the method was developed and benchmarked on several high-resource and coarse-grained datasets, it fails to adapt to the specialized knowledge requirements of our low-resource task.

The Success of Low-Resource, Generation-Based Enrichment. Our proposed method of SynCE successfully demonstrates that synthetic descriptions can lift VLM classification performance in low-resource domains in a training-free and easy-to-use manner. The issue

of under-representation in the retrieval database is overcome by synthesizing knowledge directly, allowing the captions to be focused on the correct class or specific fine-grained details. However, our results also suggest that the optimal enrichment strategy is model-dependent and an important factor in the performance gains is the manner in which a VLM's architecture is built up. Additionally, we note that the performance gains are not substantial and that there is room for further improvement of the method, specifically in creating better alignment between the image and class descriptions, and reducing possible noise introduced by our generation models.

For the **PerceptionEncoder** classifier, we concluded that it benefited most from class-side only enrichment. This indicates that its vision encoder already produces highly descriptive image features that are well-aligned with its textual prompts, causing the additional image-side captions to slightly decrease performance in most cases. For this model, we found that the main challenge was separating the class prototypes which was effectively done by the generated candidate descriptions used to enrich the class-side. On the other side, we found that for the **SigLIP** classifier, the dual-enrichment strategy was most effective. These findings suggest that SigLIP's initial image and class representations became more aligned in the representation space by enriching the embeddings with the synthetic captions, helping to introduce external knowledge that the classifier might lack in this low-resource and fine-grained domain.

Generative Model Quality. A key limitation of SynCE is its reliance on the quality of the underlying generative models. The potential for LLMs and VLMs to produce factual inaccuracies or hallucinations in their descriptions poses a risk, as this can introduce noisy or misleading information into the enrichment process. Even though our results show a positive effect when applying our method, the reliability of the generated text is a critical factor for the success of this approach. Note that as the capabilities of LLM and VLM generators improve, these hallucinations or inaccuracies decrease, causing our method to see additional performance gains.

VLM classifier misinterpretation Finally, it is important to mention that the performance of any enrichment method, including SynCE, is ultimately capped by the capabilities of the underlying VLM classifier. Our work has focused on improving the quality of the textual input signal for improved enrichment performance, but our analysis reveals that the classifier can still fail even when this signal is perfectly accurate. As we highlight in Chapter A of the Appendix, providing the classifier with 'perfectly aligned' captions, where the image-side captions are identical to the correct class-side captions, does not result in a perfect 100% accuracy, but is currently capped at 89% for the brief experiments we conducted. This indicates that the VLM's embedding space is not flawless and that when a visually ambiguous image produces an embedding so close to an incorrect class prototype, the perfect textual cue can be overridden, resulting in an incorrect prediction. The performance ceiling when using these captions may also point to a more fundamental limitation regarding the role of truly semantic relevance in prompt improvement methods. The authors of "Waffling around for Performance" [29] found that replacing meaningful descriptors with random text can yield similar performance gains, which suggests that improvements might come from an ensembling effect of structured noise rather than from a deep semantic understanding of the text by the VLM classifier. This could explain why even our correct and highly descriptive captions do not fully resolve all misclassifications. These limitations show that even though enrichment methods can significantly improve performance by clarifying the semantic context, the underlying robustness of fine-grained classification highly depends on the continued advancement in the core VLM architectures themselves.

Chapter 8

Conclusion & Future Work

8.1 Conclusion

This thesis addressed the challenge of performing fine-grained, zero-shot classification in low-resource domains where VLM adaptation methods like prompt tuning, retrieval-based enrichment or coarse-grained focused generation-based enrichment fall short. We proposed **SynCE (Synthetic Caption Enrichment)**, a training-free method that utilizes the most capable generative models to create attribute-focused descriptions for both class labels and input images.

The results from our experiments presented in Chapter 6 demonstrated that SynCE effectively improves classification performance over the baselines. In order to answer our research question, this work confirms that synthetically enriching both image and class embeddings is a viable and efficient strategy for enhancing fine-grained classification in low-resource domains. The success of SynCE highlights that generation-based enrichment can overcome the coverage limitations of retrieval-based methods, providing a flexible and robust solution for image classification tasks in specialized domains. However, we have to note that the performance gains are not substantial and that there is room for further improvement of the method, which we mention in the next section.

8.2 Future Work

From the findings and limitations of this thesis, we can consider the following future works to extend the ideas presented by our method:

Tighter VLM-LLM Caption Alignment. Our results showed that the optimal enrichment strategy is model-dependent. Future work could explore methods to create a tighter alignment between the captions generated by the class-side LLM and the image-side VLM, improving upon our proposed few-shot captioning method. This could include techniques that force the captions to follow a shared vocabulary or a fine-tuning step of both sides on a small, relevant corpus, aligning the phrasing and structure of the captions. Achieving a more consistent captioning style across the models could lead to more robust performance gains.

Additional encoder & generative models. To further investigate the model-dependent nature of our enrichment strategy, future work could expand the evaluation to include a wider variety of VLM encoders. This would help identify which underlying architectures make a model perform better when introduced to dual-sided enrichment. Additionally, experimenting

with different generative models for captioning could reveal optimal VLM-LLM combinations that produce the most aligned and effective descriptions for both the class and image side.

Hybrid Generation-Retrieval Method. A novel approach, building upon our method and the method of CoRE, could combine generation and retrieval into a single enrichment framework. In this hybrid method, an LLM would first be used to generate a set of high-quality class-side captions for all candidate classes, which would then be utilized as the domain-specific retrieval database to retrieve the image captions from. For a query image, the image-side caption would then be retrieved from this set, which guarantees a perfectly aligned structural and semantic alignment between the image- and class-side descriptions that SynCE is currently missing.

Grounded Caption Generation, Reducing Inaccuracies. In order to reduce the risk of LLM hallucinations and inaccuracies in the class-side descriptions, future work could focus on grounding the generation process by utilizing expert-written text such as our web-scraped pages from MilitaryToday.com. Using the meta-data from these pages as a guide to generate the class captions would ensure the details to follow the factual expert data, which might prove to be more accurate than purely LLM-generated captions. Keep in mind however that humans also make mistakes, and that LLMs could also prove to be more accurate in generating details about the data.

Comparison with a Fine-tuned Baseline. To provide better evaluations on the performance of our training-free SynCE method, creating a strong, fine-tuned baseline would be useful. Comparing SynCE's zero-shot performance against the fine-tuned VLM would provide a direct comparison of the effectiveness of training-free enrichment versus traditional supervised training in a truly data-scarce environment.

Investigating the Role of Semantic Relevance. Finally, the qualitative analysis of our proposed method, informed by the findings from works like "Waffling around for Performance" [29], raises the important question about whether performance gains from enrichment stem from a truly deep semantic understanding of the descriptive text or rather from an ensembling effect of structured noise. A critical direction for future work could be to explicitly test the value of semantic relevance in our enrichment and dynamic weighting framework. This could be achieved by comparing the performance of SynCE's attribute-focused captions against a set of captions that match the same structure (e.g., length and sentence count) but contain random and semantically irrelevant content. These experiments could help provide deeper insights into whether VLM classifiers genuinely leverage the fine-grained knowledge provided to them by the generated captions or if they primarily benefit from the regularization that the structured text offers, which is often the case in simple prompting techniques.

Conclusively, the methods presented in this work not only offer insights and a practical solution for current low-resource challenges, but also pave the way for future research into the critical interactions between language, vision and data scarcity.

Appendix A

Analysis on perfectly aligned captions

As discussed in the qualitative analysis in Section 6.3.3 and Figure 6.6, some of SynCE's failure cases occur even when the generated class- and image-side captions appear to be correct and well-aligned. This conclusion suggests that the VLM classifier sometimes misinterprets a correct textual signal, likely due to the overpowering influence of the original image feature, which could be visually ambiguous due to the fine-grained differences between similar classes.
To briefly investigate the upper bound of a 'perfectly' aligned enrichment signal, we perform a short experiment where this ideal scenario is simulated by bypassing the image-side VLM generator and instead directly assigning the exact same set of captions to an image that were generated for its correct class label. This creates a 'perfectly aligned' environment where the textual information of both the class- and image-side is identical for an image and its correct class.

The results presented in Table A.1 highlight a significant increase in performance, where the Top-1 Accuracy jumps to **89.11%**. This shows the potential of having an increasingly aligned enrichment signal on both sides of the classifier. However, the fact that the accuracy is not 100% confirms our hypothesis from Section 6.3.3 that even with an ideal textual signal, the VLM classifier can still make errors, likely due to the original image embedding being so visually similar to an incorrect class that it overrides the textual information.
It is important to note that these experiments were conducted briefly as a conceptual exploration. We believe that developing a more robust method in perfectly aligning the class- and image-side captions could close this gap even further.

Table A.1: Best SynCE SigLIP results vs perfectly aligned captions

Method	**Acc@1**	**Acc@5**	**Precision**	**Recall**
SynCE (Best)	76.44	99.39	79.73	78.80
Aligned captions	**89.11**	**100**	**84.78**	**86.11**

Bibliography

[1] Fine-tuning vlm: Enhancing geo-spatial embeddings. `https://encord.com/blog/fine-tuning-vlm-enhancing-geo-spatial-embeddings/`. Accessed: 2024-04-04.

[2] Sfr-embedding-mistral: Enhance text retrieval with transfer learning. `https://www.salesforce.com/blog/sfr-embedding/`. Accessed: 2024-10-28.

[3] Abdelrahman Abdelhamed, Mahmoud Afifi, and Alec Go. What do you see? enhancing zero-shot image classification with multimodal large language models, 2025.

[4] Thejaswi Adimulam, Swetha Chinta, and Suprit Kumar Pattanayak. Transfer learning in natural language processing: Overcoming low-resource challenges. *International Journal of Enhanced Research In Science Technology & Engineering*, 11:65–79, 2022.

[5] Nantheera Anantrasirichai, Thanarat H. Chalidabhongse, Duangdao Palasuwan, Korranat Naruenatthanaset, Thananop Kobchaisawat, Nuntiporn Nunthanasup, Kanyarat Boonpeng, Xudong Ma, and Alin Achim. Icip 2022 challenge on parasitic egg detection and classification in microscopic images: Dataset, methods and results. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4306–4310, 2022.

[6] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023.

[10] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification, 2024.

[11] Nicola Dall'Asen, Yiming Wang, Enrico Fini, and Elisa Ricci. Retrieval-enriched zero-shot image classification in low-resource domains, 2024.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13773–13774, 2020.

[14] Yingjun Du, Wenfang Sun, and Cees G. M. Snoek. Ipo: Interpretable prompt optimization for vision-language models, 2024.

[15] Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

[16] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018.

[17] Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. Open-vocabulary vs. closed-set: Best practice for few-shot object detection considering text describability, 2024.

[18] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. Llm2clip: Powerful language model unlocks richer visual representation, 2025.

[19] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[21] Tong Liang and Jim Davis. Making better mistakes in clip-based zero-shot classification with hierarchy-aware language prompts, 2025.

[22] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge, 2023.

[23] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E. O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts, 2023.

[24] Sachit Menon and Carl Vondrick. Visual classification via description from large language models, 2022.

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

[26] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M. Asano, Nanne van Noord, Marcel Worring, and Cees G. M. Snoek. Tulip: Token-length upgraded clip, 2025.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[28] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. Chatgpt-powered hierarchical comparisons for image classification, 2023.

[29] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts, 2023.

[30] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions, 2024.

[31] Gözde Gül Şahin. To augment or not to augment? a comparative study on text augmentation techniques for low-resource nlp. *Computational Linguistics*, 48(1):5–42, 2022.

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[33] Marton Szep, Daniel Rueckert, Rüdiger von Eisenhart-Rothe, and Florian Hinterwimmer. A practical guide to fine-tuning language models with limited data, 2024.

[34] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2025.

[35] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018.

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[37] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition, 2023.

[38] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[39] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024.

[40] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024.

[41] Yunhua Zhang, Hazel Doughty, and Cees G. M. Snoek. Low-resource vision challenges for foundation models, 2024.

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022.

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, July 2022.