Grounding Concepts to Vision via Descriptions

by

Michael Ogezi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

This thesis introduces a new approach for grounding concepts to vision using visual descriptions[1], which are text-based descriptions of visual attributes. We hypothesize that these descriptions can enhance the grounding of concepts to vision, thereby improving performance in vision-language tasks. We also suggest that these descriptions can be effectively produced using pre-trained language models. Toward validating our hypotheses, we conduct two studies.

In the first study, we address the task of visual word sense disambiguation. This task aims to select the image that best represents the meaning of a word in context. Here, we demonstrate that augmenting the original context with rich visual descriptions produced by a language model significantly improves performance.

In the second study, we attempt to produce visual descriptions for arbitrary, concrete concepts, focusing on two downstream tasks: zero-shot image classification and zero-shot class-conditional image generation. Primarily, we demonstrate that conditioning a large language model with lexico-semantic knowledge from a semantic knowledge base produces richer, and better grounded visual descriptions than previous methods. Furthermore, these visual descriptions result in substantial empirical improvements in the aforementioned downstream tasks.

Overall, this thesis confirms our initial hypothesis and demonstrates that visual descriptions offer a robust mechanism for grounding concepts to the visual domain.

[1] https://en.wikipedia.org/wiki/Visual_description

Preface

The work presented in Chapter 2 is published as M. Ogezi, B. Hauer, T. Omarov, N. Shi, and G. Kondrak "UAlberta at SemEval 2023 Task 1: Context Augmentation and Translation for Multilingual Visual Word Sense Disambiguation" (Ogezi et al., 2023b). The author of this thesis has implemented the methods and conducted all experiments described in the chapter.

Chapter 3 is adapted from the research article M. Ogezi, B. Hauer, and G. Kondrak "Visualy-Grounded Descriptions Improve Zero-Shot Image Classification" (Ogezi et al., 2023a) in submission. The author of this thesis has implemented all methods and performed all experiments described in the chapter.

“Vision is the art of seeing what is invisible to others.”

-Jonathan Swift

To my parents, who loved me unconditionally, not merely as a biological imperative, but as a profound act of devotion that shaped my life. Your unwavering support and belief in me have been the bedrock of my journey. This work is a testament to your love and sacrifice.

Acknowledgements

I express my gratitude to my supervisor, Greg Kondrak, for his guidance. I also appreciate Alona Fyshe and Lili Mou for warmly welcoming me into their group meetings and for creating a conducive environment for research discussions.

In no particular order, I owe a debt of gratitude to Sam, Uduak, Ning, Talgat, Eric, Abhishek, Bryan, Siting, Revan, Sacha, Justin, Daniela, Aidan, Dawn, Tales, Jiayi, Abdul, Sambo, Francesco, Sumedh, Shreya, Anna, Prabhat, Shibhansh, Farzane, Bradley, Alan, Gabriel, and Runor for their invaluable emotional and research support during my degree. I have made many great friends here and appreciate all the amazing memories I've shared with them.

I extend my profound gratitude to my cherished family friends in Edmonton: Bunmi and Niyi, as well as, Funmi and Gbenga. They have not only made my stay delightful but have also been my family away from home.

I extend my gratitude to my siblings Susan, Oja, Enoche, Eunice, and Tina, as well as my parents Eunice and Agbaji. Their steadfast support has been instrumental to all I have achieved.

I am thankful to the Alberta Machine Intelligence Institute (Amii) for their unwavering support towards me and all AI graduate students.

This thesis was made possible through the funding provided by Amii.

Table of Contents

List of Tables

List of Figures

Glossary of Terms

Image Classification A task in computer vision that involves predicting the class or category of an image.

Language Model (LM) A statistical model that predicts the likelihood of a sequence of words in a sentence.

Large Language Model (LLM) A language model parameterized by a large neural network and trained on extensive text data.

Visual Concept Description (VCD) A novel task that involves describing the visual attributes of a class or concept with text.

Visual Description A text description that specifies the visual attributes of a class or concept.

Visual Word Sense Disambiguation (V-WSD) A vision-language task that involves selecting the image from a set of candidates that best depicts a word's meaning in context.

Word Sense Disambiguation (WSD) A task in natural language processing that involves determining the correct meaning of a word based on its context.

WordNet A lexical semantic knowledge base.

Zero-Shot A machine learning paradigm where a model is able to perform tasks it has not been trained on.

Zero-Shot Class-Conditional Image Generation (ZSCIG) A vision-language task where a model generates images based on a dynamic set of classes that it has not been trained on.

Zero-Shot Image Classification (ZSIC) A vision-language task that involves categorizing images into a dynamic set of classes that it has not been trained on.

Chapter 1

Introduction

The intersection of natural language processing (NLP) and computer vision (CV) represents a critical frontier in the field of artificial intelligence. The ability to understand and process both language and vision is of paramount importance in creating systems that can interact with the world in a more human-like manner. This involves not only interpreting language and vision as separate entities but also connecting both modalities to create a more holistic understanding of the real world.

In this thesis, we aim to contribute to this fast-growing field by conducting studies that focus on two language-vision tasks: visual word sense disambiguation (V-WSD) and visual concept description (VCD). V-WSD involves selecting, from a set, the image which best represents the contextual meaning of a word, while VCD involves producing visual descriptions for concepts. These tasks have significant real-world applications. For instance, in text-based image retrieval, understanding the visual context to which text queries relate can result in more accurate and relevant results. Similarly, in zero-shot image classification and generation, the ability to accurately represent concepts and associate them with relevant images can greatly enhance the performance of solution systems.

The first study focuses on the visual word sense disambiguation (V-WSD) task. Here, we augment the *very short* context already provided. During augmentation, we extend the context with a rich, visual description of the original. This approach yields strong improvements in performance.

The second study focuses on the novel task of visual concept description (VCD). Visual descriptions produced by VCD can subsequently be used for downstream tasks such as zero-shot image classification and generation. This is a valuable upstream task as it allows systems to produce and understand descriptions of concepts that they have not been explicitly trained on. Our two-part approach leverages semantic knowledge bases and transformer-based, pre-trained large language models to produce rich and well-grounded descriptions of concepts. We earn strong improvements in our downstream tasks, demonstrating the benefits of improved visual descriptions in language-vision systems.

Both studies show the value of rich and well-grounded visual descriptions in bridging the gap between concepts and the images depicting them. We enhance the capabilities of language-vision models and pave the way for a deeper understanding of concepts across both modalities.

1.1 Background

This section provides an overview of the key ideas that underpin our work. We discuss language grounding, semantic knowledge bases, concepts, and language models, all of which play a crucial role in our proposed methods.

1.1.1 Language Grounding

Language grounding involves connecting linguistic symbols (language) to perceptual experiences and actions (Harnad, 1990). Thus, language grounding enables us to associate everyday language with common experiences, knowledge, or emotions.

1.1.2 Semantic Knowledge Bases

Semantic knowledge bases (SKBs) are repositories of knowledge that capture information about the world. This knowledge is structured as concepts and the relationships between them and the relationships between them. Prime examples of SKBs, in this work, are wordnets such as Princeton WordNet (Miller, 1998), BabelNet (Navigli and Ponzetto, 2012), and the Open Multilingual Wordnet (OMW) (Bond and Paik, 2012). SKBs are composed of basic semantic units called synsets.

Synsets and Semantic Relationships

A synset is a set of words that are interchangeable in some context without changing the truth value of the proposition in which they are embedded. For instance, consider the synset, `happy.1` with the constituent words "happy", "elated", and "felicitous." These words can be used interchangeably in a sentence like "She is [happy/elated/felicitous] about the news," without altering the overall meaning.

SKBs model various relationships between synsets. The relationships we focus on in this work are hypernymy and hyponymy, which model the is-a relationship. More specifically, hypernyms are supersets, while hyponyms are subsets. For example, an apple is a kind of fruit; thus, fruit is a hypernym of apple, and apple is a hyponym of fruit.

Lexico-Semantic Knowledge

SKBs also store lexico-semantic knowledge about synsets. This includes glosses (brief definitions), usage examples, and semantic relationships. For instance, the gloss for the synset `apple.1` might be "a round fruit with red or yellow or green skin," providing a brief definition of the concept represented by the synset.

1.1.3 Concepts

A concept is an abstract notion that generalizes a class or category of objects, events, or experiences (Spitzer, 1975). In the context of SKBs, we can think of a synset as a group of words that can express the same concept. Hence, the relationship between synsets and concepts is direct, and a synset is a representation of a concept.

For example, consider the word “apple” in the sentence “I ate an apple.” Next, consider an image of an apple. Although these two examples involve a lexical representation and a visual representation, respectively, they both indicate the same concept, which can be expressed both lexically and visually.

1.1.4 Language Models

Language models (LMs) are probabilistic, computational models that predict the likelihood of a sequence of words (Jurafsky and Martin, 2023). They play a crucial role in many NLP tasks, such as machine translation, sentiment analysis, and question answering. In this work, we describe two types of language models: n-gram language models and neural language models.

N-Gram Language Models

N-gram language models are a type of probabilistic language model used to predict the next word in a sequence given the previous $n - 1$ words (Jurafsky and Martin, 2023; Charniak, 1996). For example, a bigram model (2-gram) predicts the next word based on the preceding word. Given the word “New”, a bi-gram model might predict “York” as the next word if “New York” appears frequently in its training data.

Neural Language Models

Neural language models (NLMs) use neural networks to predict the next word in a sequence. Unlike n-gram models, which estimate probabilities based on the frequency of sequences in the training data, NLMs learn continuous vector representations, or embeddings, of words and use these embeddings to compute the probability of a sequence of words (Bengio et al., 2000).

Transformer-Based Language Models The Transformer model (Vaswani et al., 2017a) is a type of NLM that uses a self-attention mechanism to consider all words in the input simultaneously. This allows the model to capture both local and global dependencies in the data. Transformer-based models, such as BERT (Devlin et al., 2018) and GPTs (Radford et al., 2018, 2019; Brown et al., 2020), have shown remarkable success in a wide range of NLP tasks.

We have described the key ideas of our work: language grounding, semantic knowledge bases, concepts, and language models. Next, we will briefly describe the link between these ideas. In essence, concepts and the relationships between them are represented within semantic knowledge bases as synsets, and we use the lexico-semantic information linked to these synsets to prompt a language model to produce text that improves language-to-vision grounding.

1.2 Contribution

This section presents our hypothesis, thesis statement, and the tasks we undertake. We also outline our specific scientific contributions.

1.2.1 Hypothesis

Our hypothesis is twofold:

1. *Visual* descriptions focusing on the visual attributes of a concept perform better in zero-shot language-vision tasks.

2. These descriptions can be reliably produced using relevant lexico-semantic knowledge and a pre-trained large language model.

To validate our hypothesis, we compare descriptions produced by our method to the baselines and previous methods on a range of language-vision tasks. This thesis contributes significantly to the language-vision field, particularly in the novel tasks of visual word sense disambiguation and visual concept description, and their relevance to language grounding.

1.2.2 Thesis Statement

This thesis demonstrates that *language models conditioned with lexico-semantic knowledge from a semantic knowledge base can produce visual descriptions that better ground lexical concepts to vision.*

Take the term *brambling*, for instance. This term refers to a small bird from the Finch family. In WordNet (Miller, 1998), it is simply defined as a “Eurasian finch”. However, our method generates a more detailed description: “Small brown bird with a black head and a white patch on its chest.” This description is not only richer but also more visually descriptive than the WordNet baseline. The empirical evidence presented in this thesis demonstrates that our descriptions are more effective in grounding concepts to vision, as they outperform baselines and previous methods in zero-shot language-vision tasks.

Our research not only deepens the theoretical comprehension of the tasks under consideration but also offers practical solutions that enhance performance across various applications.

From our work, we make three primary contributions.

Contribution #1: We demonstrate that prompting a language model with lexico-semantic knowledge can produce visual descriptions that improve language-to-vision grounding and enhance performance on language-vision tasks We focus on two general tasks: visual word sense disambiguation and visual concept description, demonstrating that both benefit from visual descriptions produced using our method.

Contribution #2: We introduce contrastive prompting, a novel prompting method that induces a language model to produce visual descriptions that are particularly useful in distinguishing similar concepts We demonstrate that contrastive prompting induces our language model to produce visual attributes that differentiate one concept from another. For instance, when contrasting alligators and crocodiles, contrastive prompting produces descriptions such as “gray, round snout” for alligators and “green, v-shaped snout” for crocodiles.

Contribution #3: We create a silver dataset of visual descriptions for the 1,000 classes in ImageNet and demonstrate that it outperforms previous alternatives for vision-language tasks We produce a single, best description for each class in ImageNet. We then extrinsically evaluate this dataset based on its performance on a suite of language-vision tasks.

1.2.3 Main Tasks

Below is a summary of the tasks we undertake and the similar methods we employ to address them. Both strategies utilize transformer-based, pre-trained large language models to some extent. Our reliance on these models is based on two key observations. Firstly, they generate contextually relevant text, thanks to the attention mechanism (Vaswani et al., 2017b). Secondly, their extensive pre-training on a comprehensive dataset equips them with a wealth of world knowledge (Brown et al., 2020), which proves beneficial for the tasks at hand.

Visual Word Sense Disambiguation This task, which involves selecting the most suitable image that represents the meaning of a word within a given context, is a challenging problem due to the inherent ambiguity of words. Our approach addresses this challenge by using a language model to augment the context already provided in the task. We construct an algorithm that scores images based on their similarity to both the augmented context and the potential senses of the focus word. This novel approach yields strong improvements in performance, demonstrating the potential of pre-trained large language models in enhancing our understanding of word meanings in the vision domain.

Visual Concept Description This task involves producing visual descriptions for classes or concepts. These descriptions can then subsequently be used for downstream tasks such as zero-shot image classification. Our two-part approach leverages semantic knowledge bases and transformer-based, pre-trained large language models to produce rich, well-grounded visual descriptions of concepts. This approach yields strong improvements in our downstream tasks, demonstrating the power of visually-grounded descriptions in enhancing the performance of language-vision systems.

In both studies, we show the value of using visually-grounded descriptions to bridge the gap between concepts and images. By leveraging transformer-based, pre-trained large language models and semantic knowledge bases, we may enhance the capabilities of language-vision models and pave the way for a deeper understanding of concepts within both modalities. This work represents a significant step forward in integrating NLP and CV, contributing to the development of more robust, effective, and useful language-vision systems.

1.3 Outline

The remainder of this thesis is organized as follows: Chapter 2 focuses on the task of visual word sense disambiguation, detailing our novel approach and the experimental results that demonstrate its effectiveness. Chapter 3 introduces the novel task of visual concept description. It presents our unique approach that leverages semantic knowledge bases and pre-trained language models and discusses the significant improvements observed in downstream tasks. Finally, Chapter 4 concludes the thesis. It provides a summary of our findings, discusses their implications for the field of language-vision tasks, describes some limitations, and suggests potential directions for future research.

Chapter 2

Visual Word Sense Disambiguation

2.1 Introduction

This chapter addresses our work on SemEval-2023 Task 1: Visual Word Sense Disambiguation[1] (Raganato et al., 2023). The visual word sense disambiguation (V-WSD) task is closely related to the word sense disambiguation (WSD) task, and similarly involves understanding and classifying the meaning of a polysemous word in context. The distinction is in how classes are defined: In WSD, a system has access to a sense inventory that enumerates the possible senses of each word, and the task is to classify the focus word according to the sense that best corresponds to its intended meaning. In V-WSD, a system is given a set of candidate images, and the task is to select the image which depicts the intended meaning of the focus word. Despite the apparent similarities, the relationship between WSD and V-WSD is not as straightforward as it might seem. On closer observation, *V-WSD* appears to closely resemble text-based image retrieval. Furthermore, one could argue that a more intuitive V-WSD task might involve using both images and text as context, with the goal of selecting the most appropriate meaning of a word from a sense repository, considering both contextual elements. However, in this work, we set aside these alternative perspectives and concentrate on the task at hand, even if its naming may seem somewhat misleading.

The multi-modal nature of V-WSD introduces challenges not encountered in WSD.

[1] https://raganato.github.io/vwsd/

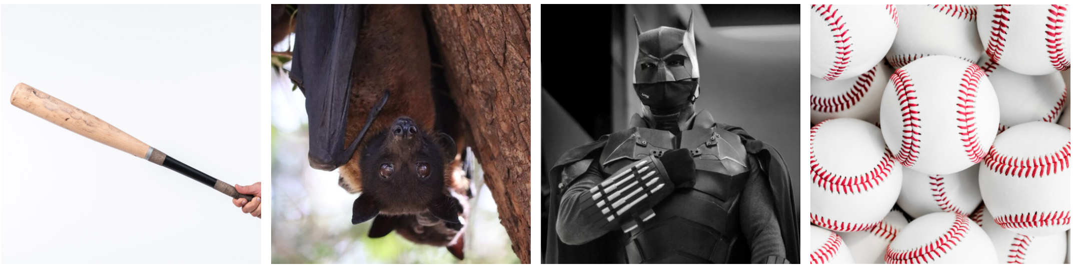


Figure 2.1: The task is to select the image that best represents the meaning of the focus word (e.g., *bat*) in the context (e.g., “baseball bat.”)

First, image processing is generally more computationally intensive than text processing. Second, a V-WSD system must represent the meanings of both images and text, and must have mechanisms to compare these multi-modal semantic representations. Last, since the candidate images in V-WSD are not restricted to a sense inventory, they may exhibit highly variable levels of sense granularity.

The V-WSD task is motivated by cases where textual context alone is insufficient to disambiguate a word. In such cases, visual context may be available to facilitate disambiguation. For example, the word *play* is ambiguous in the context “That was a good play,” as it may refer to a theatrical performance or an action in a sport. However, an associated image of a stage or a sports field would enable a V-WSD system to disambiguate *play*.

We propose a novel V-WSD algorithm that ranks candidate images by embedding images and words-in-context in a shared semantic space, while also taking advantage of lexical knowledge bases commonly used in WSD. In particular, our method uses sense glosses of the focus word to create representations of the possible meanings that word may have. Our algorithm is flexible and includes several optional modules, as well as hyper-parameters that facilitate customization, optimization, and detailed analysis.

We test various configurations of our method and analyze their performance. We draw three principal conclusions. First and foremost, the augmentation of the original textual context plays a crucial role in improving performance. Second, there is a considerable gap between English and non-English performance, indicating that bias towards English models extends to the multi-modal setting. Third, we observe a major distribution shift between the train and test sets, which is confirmed by our ablation study.

2.2 Related Work

The field of Word Sense Disambiguation (WSD) has made considerable strides in recent years, with research primarily diverging into two main approaches: supervised and knowledge-based systems. Supervised WSD methods, which hinge on extensive training corpora where content words are tagged with their correct senses, have been the subject of numerous studies (Blevins and Zettlemoyer, 2020; Barba et al., 2021). Conversely, knowledge-based methods draw on other linguistic knowledge sources, providing a distinct perspective on the problem (Wang and Wang, 2020). Despite the contrasting approaches, contemporary supervised methods have generally outperformed knowledge-based ones (Pasini et al., 2021).

The evolution of WSD saw the integration of visual information, leading to the emergence of the early precursors of Visual Word Sense Disambiguation (V-WSD). The then groundbreaking work by Barnard et al. (2003) signaled the onset of this new direction, proposing a statistical model that links image regions with words to predict word senses, thereby setting the stage for future V-WSD research. Building on this groundwork, Loeff et al. (2006) employed spectral clustering to group similar images corresponding to the same senses, further honing the methodologies within this area. In a different vein, Saenko and Darrell (2008) used an unsupervised method to assign senses to images using surrounding texts and dictionary definitions, subsequently training a visual SVM classifier to disambiguate unseen images. The field was further broadened with the introduction of the task of visual verb sense disambiguation by Gella et al. (2019). This task, which involved selecting an image based on a given context, bore a striking resemblance to V-WSD. Vascon et al. (2021) further evolved this concept by proposing a graph-based semi-supervised transductive learning method for visual verb sense disambiguation.

The advent of multi-modal foundation models (Bommasani et al., 2021), such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), has paved new paths for V-WSD and, more broadly, for research areas incorporating visual knowledge into lexical semantics. These models, capable of representing both text and images in a shared embedding space, have revolutionized the field. Recent work[2] (Bianchi et al., 2021; Sajjad Ayoubi, 2022) has further enhanced the text encoder by bootstrapping off pre-trained text-only encoders like BERT (Devlin et al., 2018), thereby further bridging the gap between text and visual information in WSD. Our method builds upon these recent advances, as well as text-based language models (Devlin et al., 2018; Brown et al., 2020; Ouyang et al., 2022), to address the task of V-WSD.

2.3 Task & Dataset

Task Definition: Given a focus word w in a short context c, and a set of candidate images I, the task is to select the image $i^* \in I$ which best represents the meaning of w in c. For example, when given the context "baseball bat" with *bat* as the focus word., a V-WSD system should choose the image that depicts the bat used in baseball (Figure 2.1).

Dataset: The training data provided for this shared task consists of a silver dataset with 12,869 V-WSD instances. Each sample is a 4-tuple $\langle f, c, I, i^* \in I \rangle$ where $|I| = 10$. The contexts are generally very short, often just a single word in addition to the focus word. We randomly select 10% of the training data for the development set. The test dataset consists of 968 instances, of which 463 are English, 200 are Farsi, and 305 are Italian. We observe that many of the incorrect candidate images in the training data have nothing to do with any sense of the focus word. However, in the test data, we

[2] https://github.com/moein-shariatnia/OpenAI-CLIP

observe that this is less often the case, making the test set considerably more difficult.

Evaluation Metrics: The primary metric is the hit rate, which is equivalent to top-1 accuracy, or simply accuracy. This is the proportion of instances for which the system selects the correct image. We also compute the mean reciprocal rank (Voorhees and Tice, 2000) which represents how highly V-WSD systems rank the ground-truth image, on average.

Language	%	Example
English	82	waxflower wildflower
Latin	15	shorea genus
German	2	truppenübungsplatz workplace
French	1	brumaire month

Table 2.1: The languages observed within a sample of 100 instances from the training set. The focus word is underlined.

Language Distribution: We observed some instances where the context contained non-English words. To estimate the prevalence of this phenomenon, we randomly selected 100 instances from the training set and manually identified the language of each. For example, the focus word *shorea* in “shorea genus” is derived from new Latin, and refers to a genus of mainly rainforest trees. Table 2.1 shows the frequency of each language in our sample.

2.4 Method

In this section, we describe the key components of our systems, including an algorithm that combines text and image similarity measures.

2.4.1 Algorithm

We propose an algorithm to select a single image from a set of candidates that best matches the context. To reiterate the problem, we are given a context c containing a focus word w and a set I of candidate images. We assume that we also have a non-empty set G containing possible glosses of w; in practice, we obtain G from BabelNet using the freely available API.[3]

Our algorithm makes calls to two similarity functions: The first is sim^L, a *written language* similarity function, which takes as input two text strings and returns a value indicating the semantic similarity between them. The second is sim^{VL}, a *vision-to-written language* similarity function, which takes as input an image and a text string and returns a value indicating the similarity between what the image depicts and what the text describes.

With these functions, for each candidate image $i \in I$, and for each gloss $g \in G$ of the focus word w, we compute the pairwise similarity between:

1. The image and context: $s_{ic} = sim^{VL}(i, c)$
2. The image and gloss: $s_{ig} = sim^{VL}(i, g)$
3. The context and gloss: $s_{cg} = sim^{L}(c, g)$

This allows us to identify the pair of a candidate image i^* and gloss g^* that maximizes a weighted average of these three similarity scores. Algorithm 1 shows the pseudocode for this algorithm.

Hyperparameters: Our algorithm depends on three weight hyperparameters: w_{ic}, w_{ig}, and w_{cg}. They represent the weights for image-context, image-gloss, and context-gloss similarity, respectively. Table 2.2 shows the results of the hyperparameter binarized grid search performed on a 500-sample of the training set. Based on our development experiment results, we decided to set all hyperparameter weights to 1 for simplicity, except where otherwise noted. We discuss the hyperparameters further in Section 2.7.5.

[3] https://babelnet.org/guide

Algorithm 1 Candidate Image Scoring

```
$c \leftarrow$ the context of the focus word
$G \leftarrow$ list of glosses for the focus word
$I \leftarrow$ list of candidate images
for $i$ in $I$ do
    $s_g \leftarrow 0$
    for $g$ in $G$ do
        $s_{ig} \leftarrow w_{ig} \cdot sim^{VL}(i, g)$
        $s_{cg} \leftarrow w_{cg} \cdot sim^{L}(c, g)$
        $s_g \leftarrow \max(s_g, s_{ig} + s_{cg})$
    $scores[i] \leftarrow s_g + w_{ic} \cdot sim^{VL}(i, c)$
return $scores$
```

Context Augmentation: For each instance, we prompt InstructGPT (Brown et al., 2020; Ouyang et al., 2022) to generate a definition for the context phrase. We use the following prompt template: "For each line, define the phrase:" followed by the contexts, one per line. For example, the context "baseball bat" is augmented to become "baseball bat: a bat used to hit a baseball during the game of baseball." The use of this additional context is described in Section 2.5.3

Supplementary Training Data: We speculate that the size of the training dataset may be a limiting factor in the accuracy of our method. We, therefore, experiment with augmenting the training data with additional data derived from BabelPic (Calabrese et al., 2020), a multi-modal resource that maps a subset of BabelNet synsets to sets of one or more images. For each pair of a synset and an image, we enumerate

w_{ic}	w_{ig}	w_{cg}	Accuracy (%)
1	1	1	79.2
1	1	0	79.2
1	0	1	72.2
1	0	0	72.2
0	1	1	68.4
0	1	0	68.6
0	0	1	11.0

Table 2.2: Binarized grid search results for weight hyperparameters.

a lemma from the base synset and a lemma from a related synset. The two lemmas are concatenated, starting with the lemma from the related base synset, to form a two-word context. We then select nine other random images from BabelPic, forming an instance comparable to those in the training set: a two-word context with a single focus word, with ten images, one depicting the correct sense of the focus. We create 54,968 instances this way and experiment with adding this dataset to the training data at training time.

Glosses: For each instance, we enumerate the BabelNet (Navigli and Ponzetto, 2012) glosses corresponding to each sense of the focus word. If there are multiple glosses for a single sense, we pick the first and add it to the set G. This prevents senses from being over-represented due to the number of glosses in BabelNet.

2.5 Systems

In this section, we describe our systems for the V-WSD task, Our official system submissions are based on our primary systems: Tr and LangSpec. We also describe two alternative systems, which do not use Algorithm 1. Both perform worse than the primary systems, but their results are nevertheless valuable for analysis. We also present a supplementary method, which can be optionally used in combination with our other systems. Non-English instances are translated using DeepL[4] for Italian and ChatGPT[5] for Farsi.

2.5.1 Primary Systems

Tr: Image Scoring with Translations If the input instance is not English, we translate it into English. Then we apply Algorithm 1. We compute sim^{VL} using embeddings from CLIP (Radford et al., 2021), an English-only model which encodes text and images in a shared embedding space. We compute sim^{L} using BERT (Devlin

[4] https://www.deepl.com/translator

[5] https://openai.com/blog/chatgpt/

et al., 2018) as an English-only text encoder. We set the weight parameters: w_{ic}, w_{ig}, and w_{cg} to 1 in this specific case.

LangSpec: Image Scoring with Language-Specific Models This system is similar to Tr, except that non-English instances are not translated into English. This is our only system that directly operates in other languages. Given a non-English instance, we replace CLIP and BERT with language-specific models to compute sim^{VL} and sim^{L}. For English instances, this method is the same as Tr. For Italian, we use CLIP-Italian (Bianchi et al., 2021) to compute sim^{VL} and Italian BERT[6] to compute sim^{L}. For Farsi, we use CLIPfa (Sajjad Ayoubi, 2022) to compute sim^{VL} and ParsBERT (Farahani et al., 2021) to compute sim^{L}.

2.5.2 Alternative Systems

Gen: Generative Image Model This method takes a different approach compared to Tr and LangSpec; it does not use Algorithm 1. Instead, we provide the context (translated into English, if needed, as outlined above) as input to Stable Diffusion (Rombach et al., 2022), a generative model which takes a text prompt as input and produces candidate images to depict what the text describes. For each context, we generate 10 images using 20 diffusion steps each. We set the guidance scale hyperparameter to 7.5. For each candidate image, we compute its cosine similarity with each generated image based on embeddings produced by CLIP. The candidate with the highest similarity to the generated images is chosen as the output.

Seg: Text-Conditioned Image Segmentation As with Gen, this method does not use Algorithm 1. Instead, we use a zero-shot image segmentation system (Lüddecke and Ecker, 2022) to segment images based on the provided context. This system produces a *mean mask value*, which we use as a measure of similarity between the

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

	EN	IT	FA	Avg
Baseline	60.5	22.6	28.5	37.2
Tr*	61.1	59.3	**43.0**	54.5
Tr+Def	**69.1**	**63.3**	40.0	**57.5**
LangSpec*	56.8	37.7	14.5	36.3
Gen	51.6	45.9	39.0	45.5
Gen+Def	58.1	48.5	34.5	47.0
Seg	31.5	29.8	20.5	27.3
Seg+Def	34.1	36.7	20.0	30.3

Table 2.3: Accuracy for English, Italian, and Farsi, along with the macro average for all languages. We indicate our official system submissions with *.

context and the segmented image; we return the image with the highest mean mask value, given the context.

2.5.3 Supplementary Method

Def: Generating Additional Context Tr, Gen, and Seg make use of the input context, translated to English as needed. However, the contexts provided in the official dataset for this task are extremely short. With Def, we generate additional context by using the original context to prompt InstructGPT for a more extensive description, as described in Section 2.4.1. We then concatenate the generated text to the context and pass this augmented context to Tr, Gen, or Seg. We refer to the methods using this supplementary method as Tr+Def, Gen+Def, and Seg+Def, respectively. We do not combine Def with LangSpec, as we observe that InstructGPT is less robust to short non-English contexts.

For Tr+Def, we set w_{ig} and w_{cg} to 0, as the improved context obviates the need for their corresponding terms in Algorithm 1. Gen+Def and Seg+Def, being based on Gen and Seg, do not depend on Algorithm 1.

2.6 Results

Table 2.3 shows our performance on the test set. We find that accuracy has a 99.46% Pearson correlation with mean reciprocal rank, and so for conciseness, we report accuracy alone. The translation-based systems, Tr and Tr+Def, yield the best results. One explanation for this outcome is the disproportionate amount of English training data available to the models we build upon: CLIP and BERT. The higher performance of these models on English appears to compensate for the noise introduced by the translation process. We discuss this further in Section 2.7.3.

An interesting trend is the benefit of context augmentation, (Section 2.5.3). Between Tr and Tr+Def, we observe a 3% average improvement in accuracy. We observe a similar trend in Gen versus Gen+Def and Seg versus Seg+Def.

We further observe that accuracy on English instances is highest, accuracy on Farsi instances is lowest, while accuracy on Italian instances is in between both. This corresponds to the quality and quantity of resources available for each language. We undertake more thorough analyses in the next section.

2.7 Discussion

2.7.1 Distribution Shift

As shown in Table 2.4, we observe a clear disparity in polysemy, and the proportion of focus words which are nouns, between the training and test sets. This difference is especially notable when considering the performance gap between the sets. As previously mentioned, this discrepancy stems from the fact that the training dataset was generated automatically, while the test dataset was curated by humans.

Zero-shot vs. Fine-tuning: We observe that fine-tuning on the training set leads to a drop in performance on the test set (Figure 2.2). This may be due to the divergence between the training and test datasets outlined above.

	Train	Test		
	EN	EN	IT	FA
Polysemy	6.8	23.1	13.6	10.7
Nouns (%)	74.7	88.1	91.5	92.5

Table 2.4: Distribution shifts between the training and test sets. Polysemy indicates the average number of senses each focus word in the set has.

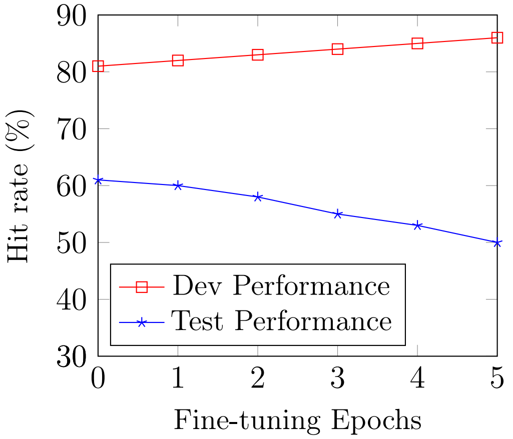


Figure 2.2: As we fine-tune on the training set for more epochs, we see an increase in dev set performance, but a drop in test set performance. This indicates that both the training and dev sets are similar to each other but different from the test set. Epoch 0 refers to using the model zero-shot.

2.7.2 Traditional WSD

Although both the V-WSD and WSD tasks have some similarities, we found that some ideas drawn from WSD prove ineffective for V-WSD.

Using Glosses: We observe empirically worse performance when using glosses in our algorithm. Specifically, with Tr on the English test set, we obtain a hit rate of 61.1% when we do not use glosses and 56.8% when we do. Such a steep drop (4.3%) is surprising, especially since most state-of-the-art WSD systems explicitly use glosses in their methods.

We posit that sense disambiguation in V-WSD is more focused on homonymy than polysemy and, as a result, can be less nuanced than in WSD. For example, *apple* could refer to a fruit or a tree. In an image depicting both, the focus may be unclear. In WSD, this distinction is critical since *tree* and *fruit* are distinct senses. In V-WSD, however, we can make a correct prediction without deciding between both senses. As a result of this lower granularity, glosses become less important.

Performance of WSD Systems on Context: We manually disambiguated the sense of the focus word in a randomly-selected set of 16 instances from the training set. We then applied a state-of-the-art WSD system, ConSec (Barba et al., 2021), to these instances. We observe that ConSec sense predictions were accurate 50% of the time, falling considerably below its reported accuracy of 82%.

2.7.3 English-Langauge Bias

Natural language processing research often focuses on the English language, at the expense of other languages (Magueresse et al., 2020). The relative performance of Tr and LangSpec reflects this phenomenon: Translating non-English text to English, in order to apply an English encoder, can be expected to introduce some noise due to translation errors and information loss. However, we observe that this pipeline approach produces better results than using an Italian or Farsi encoder directly. This suggests that the field's focus on English has yielded (multi-modal) English encoders that perform well, but do so at the expense of other languages such as Italian or Farsi. Advancing the state-of-the-art for non-English encoders may potentially further improve performance in those languages, by avoiding the need to translate to English.

2.7.4 Image Generation

As shown in Figure 2.3, when applying our image generation system (Gen), we observe an increase in performance as we generate more images. Although the performance jump when transitioning from 1 to 5 images is most pronounced, we see benefits from scaling until a certain point, 10 images, where the trend becomes unreliable.

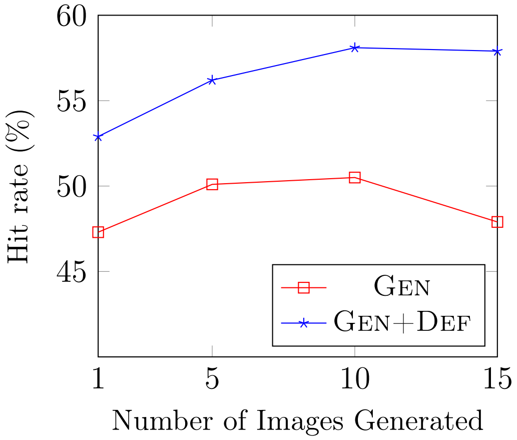


Figure 2.3: Hit rate (%) vs. number of images generated for GEN and GEN+DEF.

2.7.5 Text-Conditioned Image Segmentation

With SEG, we can sometimes robustly segment images and predict masks indicating the correct image, conditioning on the full context. However, this method sometimes forms incorrect semantic representations. Appendix A details more examples of SEG's usage. In addition to Figure 2.4, we present more extensive examples in Appendix A.

2.7.6 Algorithm Hyperparameters

Algorithm 1 uses three weights hyperparameters to balance pairwise similarities. We set all weights to 1 based on Table 2.2. Comparison of results with w_{cg} set to 0 or 1 suggests that $sim^L(c,g)$ does not improve performance. Two reasons support this finding. Firstly, images encode richer representations, producing more precise $sim^{VL}(i,g)$ and $sim^{VL}(i,c)$, while both context and glosses are discrete textual features, introducing uncertainty to $sim^L(c,g)$. Secondly, we use CLIP and BERT to calculate sim^{VL} and sim^L, respectively. CLIP's multi-modal pre-training may offer better similarity scores, fitting this task better. Understanding these findings more deeply is an interesting avenue for future research.

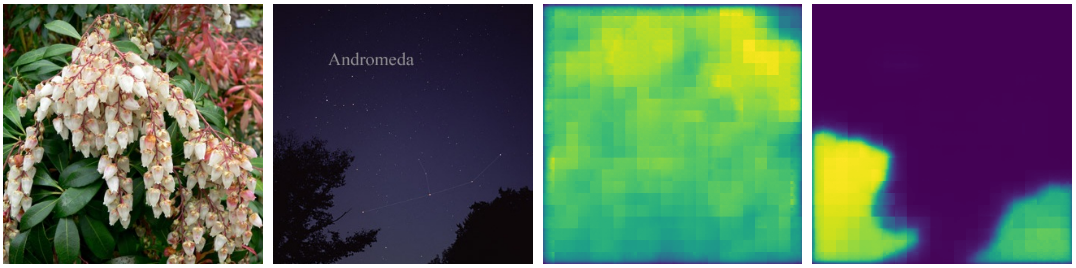


Figure 2.4: Original images from the dataset depicting ANDROMEDA (Japanese plant), ANDROMEDA (galaxy) and their two masks conditioned on “andromeda tree.”

2.8 SemEval-2023 Shared Task Reflections

This section highlights common strategies employed by the best methods in the Visual Word Sense Disambiguation (V-WSD) SemEval-2023 shared task (Raganato et al., 2023), then goes on to briefly describe the four top-performing methods.

2.8.1 Common Strategies

Here are the common strategies employed in work by various groups on the shared task. We employ all these strategies in our study.

Contextual Information Enrichment Strategies to enrich contextual information, such as establishing relationships between concepts and sentences, are commonly employed. This is closely related to our method and core thesis.

Pre-trained Vision-and-Language Models All methods utilize pre-trained models like CLIP in a zero-shot setting.

Generative Models and Data Augmentation Additional or augmented training data is often created to enhance model robustness.

Integration of External Resources Resources such as Wikipedia and semantic knowledge bases are frequently used to enrich models with linguistic and semantic knowledge.

Ensemble Techniques Top methods often combine different models or system modules to leverage their strengths.

2.8.2 Top-Performing Methods

TAM of SCNU (Yang et al., 2023) This system uses a Fine-grained Contrastive Language-Image Learning (FCLL) model and a new multilingual, multi-modal knowledge base for ambiguous words.

Samsung Research China - Beijing (Zhang et al., 2023) This model builds a reference sense inventory from definitions and synonyms of the target word and uses a bi-encoder architecture with SimCSE as the backbone.

Zywiolak (Dadas, 2023) This hybrid system combines multi-modal embeddings and knowledge-based approaches, integrating various system modules using a learning to rank (LTR) model.

Rahul (Patil et al., 2023) This system is an ensemble of different neural models, including CLIP models for English and text-to-text translation models for Farsi-to-English and Italian-to-English.

These methods demonstrate the effectiveness of combining pre-trained models, data augmentation, external resources, ensemble techniques, and contextual enrichment strategies in V-WSD.

2.9 Conclusion

In this chapter, we outlined our work on the recently-proposed task of visual word sense disambiguation (V-WSD). We found that many ideas from traditional WSD are difficult to adapt to V-WSD, and, that in general, WSD systems are not useful for V-WSD. We were particularly surprised to find that, unlike in WSD, glosses appear to be unhelpful for V-WSD. Contrariwise, our innovation of augmenting the context did yield substantial gains in accuracy. We posit that this happens because the process of augmentation more strongly grounds the concept being represented in the sample in the visual domain. Consequently, language-vision models like CLIP can form more robust semantic representations.

Further research will be needed to establish the connection between V-WSD and the broader field of lexical semantics. We speculate that developing systems for joint WSD and V-WSD may yield improvements in one or both tasks. Our work here serves as a proof-of-concept establishing the utility of language models and lexico-semantic resources in the developing task of V-WSD.

Chapter 3

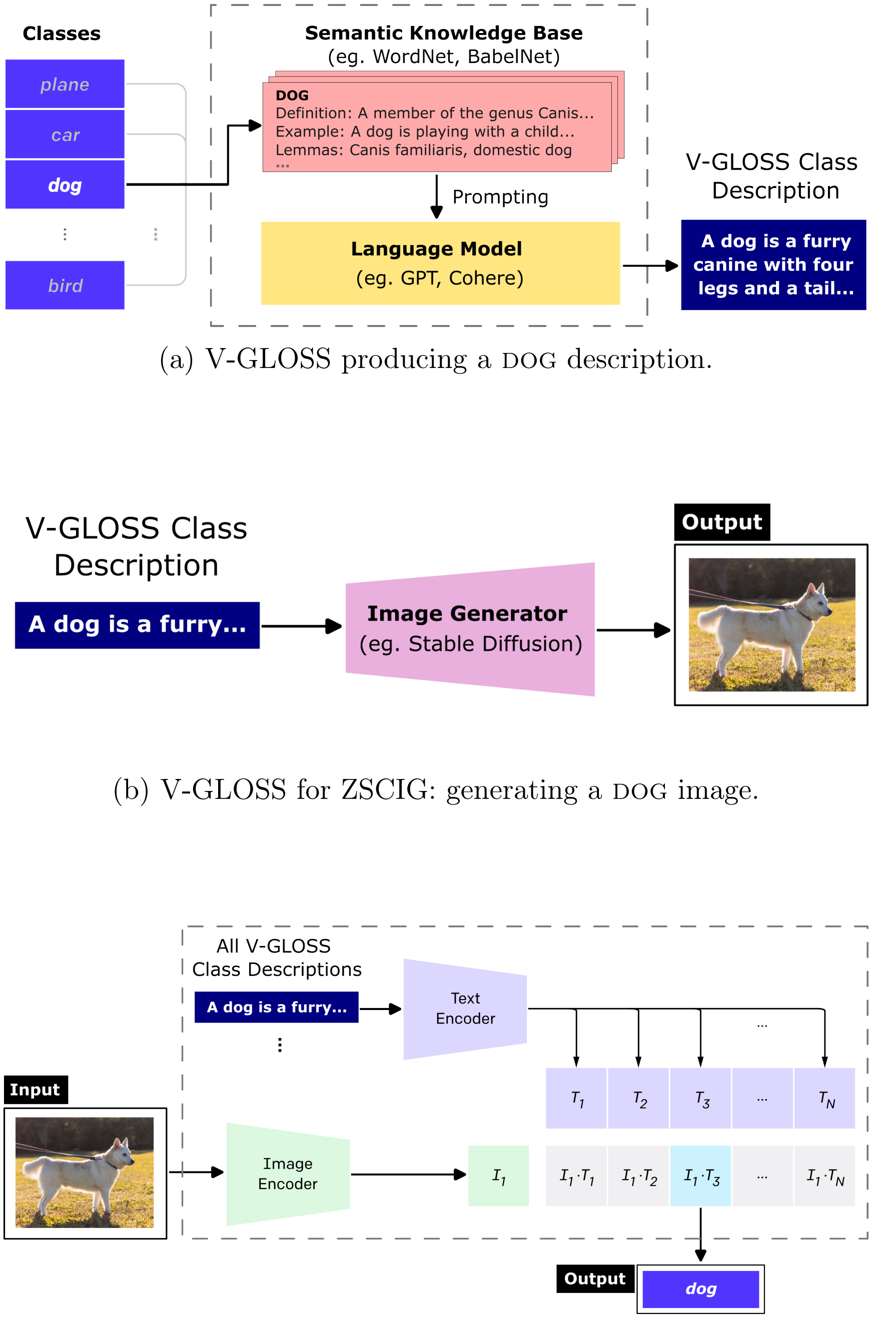
Visual Concept Description

3.1 Introduction

Language-vision models (Radford et al., 2021; Jia et al., 2021) have made significant progress in zero-shot vision tasks. However, we hypothesize that their accuracy is limited by a lack of visual descriptions that are both expressive and specific with regard to an intended concept, that is, glosses that describe what a concept or class looks like. In this work, we introduce the visual concept description (VCD) task, and we investigate our hypothesis by creating and testing a novel method for producing these descriptions.

Improving visual descriptions is crucial for enhancing system performance in zero-shot vision tasks. Such descriptions facilitate the creation of more useful representations. Additionally, being able to describe a concept in terms of its appearance is essential for developing more robust and adaptable methods incorporating diverse visual information across various domains, without the need for extensive re-training.

Existing approaches to generating class descriptions, such as those employed by CLIP (Radford et al., 2021) and CuPL (Pratt et al., 2022), involve directly plugging class labels into fixed templates (e.g., *a photo of* X), or using large language models such as InstructGPT (Ouyang et al., 2022) to generate descriptions based on class labels (e.g., *what does* X *look like?*). These methods suffer from two main issues: class granularity and label ambiguity. Class granularity refers to the difficulty in



(a) V-GLOSS producing a DOG description.

(b) V-GLOSS for ZSCIG: generating a DOG image.

(c) V-GLOSS for ZSIC: classifying a test image.

Figure 3.1: For the DOG class, we depict (a) V-GLOSS's architecture (Section 3.4.2), along with adaptations: (b) ZSCIG (Section 3.5.4) and (c) ZSIC (Section 3.5.4)

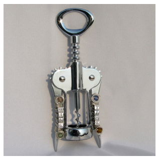
Class / Concept	WordNet Gloss	V-GLOSS (Ours)
CORKSCREW 	A bottle opener that pulls corks.	A **tool** with a **spiral blade** that is used to remove corks from **bottles**.
BRAMBLING 	Eurasian finch.	A **small brown** bird with a **black head** and a **white patch** on its chest.
BROCCOLI 	Branched green undeveloped flower heads.	A **green vegetable** with a **thick stalk** and **florets** that grow in a **dense** head.

Table 3.1: A qualitative comparison between WordNet concept glosses and V-GLOSS (Silver) class descriptions for some ImageNet classes. Our method describes what a class *looks like*, instead of what it *does* or *is*.

distinguishing between visually similar classes, such as ALLIGATOR and CROCODILE. Label ambiguity is caused by using polysemous words as labels for distinct concepts. For example, CRANE can refer to either a bird or a construction machine. These issues adversely affect the performance of existing models (Radford et al., 2021).

To address these challenges, we introduce V-GLOSS, a novel method that leverages transformer-based, pre-trained large language models (LLMs) and semantic knowledge bases (SKBs) to generate visually-grounded class descriptions – **v**isual **gloss**es. Table 3.1 shows some examples. By combining structured semantic information from SKBs such as WordNet (Miller, 1998), with a contrastive algorithm to distinguish similar classes, V-GLOSS is designed to mitigate the dual issues of granularity and ambiguity.

Our results demonstrate the effectiveness of V-GLOSS in improving the performance of ZSIC systems. We achieve state-of-the-art (SOTA) results on benchmark datasets such as ImageNet (Deng et al., 2009), CIFAR-10, and CIFAR-100 (Krizhevsky et al., 2009) in the zero-shot setting, and STL-10 (Coates et al., 2011) in both the zero-shot and supervised settings. In addition to this principal contribution, we also introduce and make available V-GLOSS Silver, a silver dataset constructed by V-GLOSS, consisting of a visual gloss for each ImageNet class. We show that V-GLOSS Silver is useful for vision tasks such as ZSIC and ZSCIG, comparing favorably to WordNet glosses.

3.2 Tasks

The main task, visual concept description (VCD), is to produce a visual description for a given class or concept. For example, if an image classification dataset has the class DOG, we aim to produce a description such as *"A dog is a furry, four-legged canine."* We consider such a description to be a specific kind of gloss.

We use two downstream tasks to compare methods of producing class descriptions: zero-shot image classification (ZSIC), and zero-shot class-conditional image generation (ZSCIG). In ZSIC, the goal is to classify an image based on a set of classes, without having seen any labeled images belonging to those classes. The set of classes depends on the dataset. For example, given an image depicting a dog, we aim to predict the class DOG. In ZSCIG, the goal is to generate an image that corresponds to a specific class, again without having seen any labeled examples. For example, given a class DOG, we aim to generate an image of a dog.

In short, ZSIC is the task of classifying a given image, while ZSCIG is the task of generating an image given a class. Both involve classes and images. VCD provides useful knowledge to facilitate both tasks, by making it easier to either recognize or generate images of each class. Therefore, by developing a novel method of producing such descriptions which focuses on the visual properties of a class, we hypothesize that performance on ZSIC and ZSCIG can be improved.

3.3 Related Work

Transformer-Based LLMs Transformer-based large language models have revolutionized many natural language processing tasks (Radford et al., 2018; Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020; Black et al., 2022; Ouyang et al., 2022). As these models are scaled up in terms of the number of parameters and the quantity of training data, abilities such as few-shot and zero-shot learning emerge (Wei et al., 2022).

Language-Vision Models Based on transformer-based architectures similar to those used by language models, language-vision models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have made rapid and significant progress, particularly in applying contrastive pre-training approaches on large image-text datasets. These advancements have led to better representation learning for both text and images, improving performance on several multi-modal tasks (Mokady et al., 2021; Song et al., 2022). Further progress has been achieved by scaling up pre-training with greater computational resources and larger datasets, as well as incorporating auxiliary training objectives (Pham et al., 2021; Yu et al., 2022).

Producing Visual Descriptions Following the rapid improvements in language-vision models, many zero-shot multi-modal tasks such as zero-shot image classification became more tractable (Radford et al., 2021). Toward tacking such tasks, Radford et al. (2021) introduce the template ensemble (TE) method, which employs a custom set of class labels as well as a fixed set of templates into which each label is inserted. The set of completed templates for each class is then aggregated into a single representation of the class. CuPL (Pratt et al., 2022) utilizes InstructGPT (Brown et al., 2020; Ouyang et al., 2022) to generate descriptions for ImageNet classes. Both TE and CuPL can be used for zero-shot image classification. Hao et al. (2022) also fine-tune GPT models (Radford et al., 2018, 2019) to rephrase image-generation prompts, resulting in improved images. In this work, we use LLMs as a foundation for generating visually grounded descriptions, while sourcing additional lexico-semantic knowledge from SKBs.

3.4 Method

We begin by describing how we map classes to concepts in a semantic knowledge base (SKB), in order to leverage the concept-specific lexico-semantic knowledge the SKB contains. We then introduce our novel method V-GLOSS, which has two variants, *normal* and *contrastive*. We conclude by describing the construction of V-GLOSS Silver, a set of visual concept descriptions produced using V-GLOSS which we will make available.

3.4.1 Mapping Classes to WordNet Synsets

The ImageNet classes are already mapped to WordNet synsets by the dataset's creators. For the other datasets, we employ a heuristic that starts by mapping each class to the most-frequent sense of the class label, as determined by WordNet[1]. For CIFAR-10 and STL-10, this heuristic is sufficient. However, for CIFAR-100, we manually re-map 18 classes. For instance, we needed to re-map RAY from *light* to *sea creature*, as the *light* sense is the most frequent according to WordNet, but the RAY images in the dataset depict sea creatures.

3.4.2 V-GLOSS

We discuss the two variants of V-GLOSS below, *normal* and *contrastive*. The *normal* variant involves generating independent visual descriptions for each concept, while the *contrastive* variant aims to create descriptions that distinguish between two concepts. In both, for each class, we produce multiple descriptions resulting in an ensemble.

[1] https://www.nltk.org/

What does a **platypus** look like?

A platypus looks like a beaver with a duck's bill

(a) CuPL (Pratt et al., 2022)

...

Concept name: eagle

Hypernyms: bird or prey

Hyponyms: bald eagle, eaglet, golden eagle, harpy

Gloss: any of various large keen-sighted diurnal birds of prey noted for their broad wings and strong...

Unique and expressive visual description: Eagles are large birds of prey with dark brown bodies and wings...

...

Concept name: ***platypus***

Hypernyms: ***duckbill, duckbilled platypus, ...***

Hyponyms: ***egg-laying mammal***

Gloss: ***small densely furred aquatic monotreme of Australia and Tasmania having a broad bill...***

Unique and expressive visual description:

Platypuses are water-dwelling mammals that have broad duck-like bills and hind legs with a foot web that has an intricate web of keratinised spongy hairs

(b) V-GLOSS (Ours)

Figure 3.2: Class descriptions for Platypus generated by two different methods that use LLMs. Input prompts, output descriptions, and **plugged values** are shown.

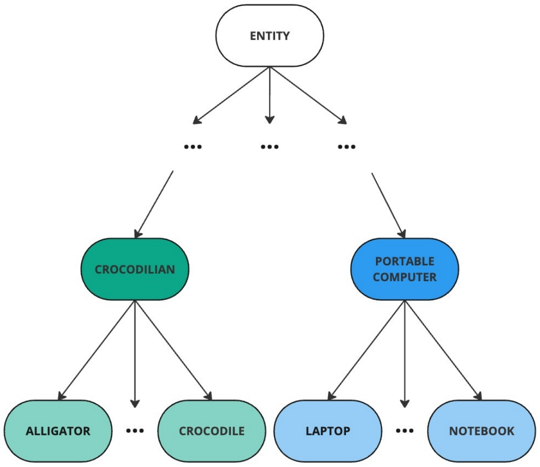


Figure 3.3: A sample of WordNet hypernym hierarchy. For *contrastive* prompting, we only distinguish classes that are semantically similar to the target class, like ALLIGATOR to CROCODILE.

Normal V-GLOSS

We generate normal descriptions via in-context learning with an LLM, beginning by providing the LLM with a description of the task to be performed, followed by multiple input-output examples. The examples are fixed, involving the concepts EAGLE, BAT (animal), BAT (baseball), and TELEVISION. We selected these to expose the model to ambiguous class labels (*bat*), a natural object (*eagle*), and an artificial object via (*television*). For each class, we obtain from WordNet the hypernyms, hyponyms, usage examples, synonyms, and glosses of the sense to which the class is mapped, and provide this to the LLM. Figure 3.2b shows a session with the LLM, beginning with the example of *eagle*, with the output generated for the class *platypus*. Table 3.1 compares our descriptions to WordNet glosses.

Contrastive V-GLOSS

During development, we observed that many errors were caused by false positives between visually similar classes. For example, the classes CROCODILE for ALLIGATOR refer to similar-looking animals, and are often confused with one another. The contrastive variant of V-GLOSS is designed to address this by using semantic similarity

Class / Concept	*Normal*	*Contrastive*
ALLIGATOR 	A large reptile with a long snout, a broad head, and a long tail.	A large, **dark-colored** reptile with a **rounded snout**, found in **freshwater**.
CROCODILE 	A reptile with a broad, flat snout, a long tail, and a long, pointed snout.	A **grayish-green** reptile with a **v-shaped snout**, found in **brackish** or **saltwater**.

Table 3.2: Two similar classes with **key differences** between their *normal* and *contrastive* descriptions.

between classes as a heuristic to estimate visual similarity. For each class, we search for other classes that are semantically similar, and if any are found, we add a negative instruction to the LLM prompt, e.g. we generate a description for an ALLIGATOR *but not* a CROCODILE, using a similar in-context prompt structure to normal V-GLOSS.

A similarity matrix M is created as follows.:

$$M_{i,j} = Sim(S[i], S[j]) \tag{3.1}$$

$Sim(s_1, s_2)$ is the Wu-Palmer path-similarity function (Wu and Palmer, 1994) comparing synsets s_1 and s_2; this similarity function uses the path between two concepts in the WordNet hypernym hierarchy (Figure 3.3) to measure semantic relatedness. S is the set of all classes in a dataset, $\mathcal{D}$, and i and j are indices ranging from 1 to $|S|$. Concisely, Equation 3.1 defines a similarity matrix containing similarity scores between all classes in a dataset. M is one of the inputs to our contrastive V-GLOSS variant, shown in Algorithm 2.

In Algorithm 2, λ is a threshold for minimum similarity. We only generate contrastive descriptions when classes have a similarity that exceeds or is equal to λ. N represents the maximum number of classes for which contrastive descriptions are gen-

Class	False Positives	Contrastives
African Elephant	Tusker (44), Asian Elephant (6)	Tusker, Asian Elephant
Notebook	Laptop (22), Desktop (10), Space bar (2)	Laptop, Desktop, Space bar

Table 3.3: False positives and their counts vs. classes selected by the contrastive algorithm (see Equation 3.1 and Algorithm 2). Hits and misses are shown.

erated. k is the number of distinct descriptions to generate for a class pair. LLM_c takes in the *target* class, a neighbor class, and k, then prompts the LLM to generate k descriptions that distinguish the *target* and neighbor classes. In summary, for each class, Algorithm 2 identifies the classes most similar to it, excluding itself, and generates descriptions that distinguish them. Table 3.2 compares the normal and contrastive descriptions for alligator and crocodile; note that distinguishing features of the two classes are included in the LLM's output. Table 3.3 shows examples of classes with high false positive rates, and the classes they are contrasted with.

Algorithm 2 Generate Contrastive Descriptions: We generate contrastive descriptions to help distinguish the most similar classes.

```
Require: $M$: Equation 3.1 result
Require: $\lambda$, $N$, $k$: Hyperparameters
Require: $S$: All classes in dataset, $\mathcal{D}$
Require: $LLM_c$: LLM prompted contrastively
        ▷ Returns a description to distinguish the second class from the first
$G \leftarrow$ empty $|S|$-list for class descriptions
for $i \leftarrow 0$ to $|S| - 1$ do
    $target \leftarrow S[i]$
    $S^* \leftarrow$ top $N$ classes : $\lambda \leq M_{i,*} \leq 1$   ▷ Select the classes that are most similar to the target class
    for $s^*$ in $S^*$ do
        $samples \leftarrow LLM_c(target, s^*, k)$
        $G[i].insert(samples)$
return $G$
```

3.5 Evaluation

Toward evaluating V-GLOSS, we describe our datasets, evaluation metrics, baselines, previous methods, and experiments.

3.5.1 Datasets

We evaluate our method on the test splits of four widely-used benchmark datasets: ImageNet (Deng et al., 2009) consists of 50,000 images equally distributed across 1,000 classes, and serves as our primary benchmark. CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) both comprise 10,000 test samples across 10 and 100 classes, respectively. Finally, STL-10 (Coates et al., 2011) comprises 100,000 test samples and is designed for unsupervised learning. For CIFAR-10, CIFAR-100, and STL-10, which are not pre-mapped to WordNet, we employ the two-step process detailed in Section 3.4.1 to map each class to a WordNet synset.

Experiment 1 (Section 3.5.4) involves ImageNet alone and covers both the ZSCIG and ZSIC tasks. In contrast, Experiment 2 (Section 3.5.5), which is our main experiment, tests the impact of various class description methods on the ZSIC task and uses all datasets. In Experiment 2, we allow methods to use ensembles of descriptions of each class, while in Experiment 1, we experiment with only a single description.

The datasets we selected to evaluate the following properties of V-GLOSS:

1. **Performance on benchmark datasets with varying numbers of classes.** Each dataset has its own set of classes, ranging from ImageNet with 1,000 classes, to CIFAR-100 with 100 classes, to CIFAR-10 and STL-10, each with 10 classes.

2. **Ability to represent diverse concepts at varying levels of granularity.** The datasets we use contain a wide range of concepts across various domains, rather than those targeting specific subareas such as pets (Parkhi et al., 2012),

foods (Bossard et al., 2014), cars (Krause et al., 2013), scenes (Xiao et al., 2010), or airplanes (Maji et al., 2013).

3.5.2 Evaluation Metrics

Top-1 Accuracy In ZSIC, this metric is the frequency with which the model's top prediction for an image matches the gold label.

Fréchet Inception Distance (FID) For ZSCIG, FID (Heusel et al., 2017) quantifies the divergence between ground truth and generated images, with lower scores signifying a better ability to produce images similar to the ground truth.

Inception Score Also for ZSCIG, the inception score (Salimans et al., 2016) uses an Inception model's (Szegedy et al., 2015) output probability distribution to assess the diversity and realism of generated images, with higher scores indicating more diverse and convincing images. Unlike the above metrics, this does not require ground-truth images to compare to.

3.5.3 Baseline & Previous Methods

In this section, we describe the methods that we compare V-GLOSS to. For methods that produce ensembles of class descriptions (i.e. multiple descriptions per class), a single representation of the class is obtained by averaging individual representations.

First, the **1-Template baseline** inserts a class label into a *single* specific template. For example, given the class DOG, the baseline produces *"A photo of a dog."*

Template Ensemble (Radford et al., 2021) generates an ensemble of descriptions for a class by inserting the class label into each member of a set of templates. For example, some descriptions for DOG are: *"A photo of a dog."*, *"A blurry photo of a dog."*, and *"An origami dog."* This method uses a modified list of class labels[2] designed to reduce ambiguity.

[2] https://github.com/anishathalye/imagenet-simple-labels

CuPL (Pratt et al., 2022) also generates an ensemble of descriptions for each class. The descriptions are generated by prompting an LLM, InstructGPT (Ouyang et al., 2022), with questions such as: *"What does a dog look like?"* and *"Describe an image of a dog from the internet."* CuPL uses the same class labels as Template Ensemble.

The authors of CuPL also combined their method with Template Ensemble. The resulting method, **CuPL + Template Ensemble**, combines the class descriptions from both methods.

3.5.4 Experiment 1: V-GLOSS Silver

This experiment evaluates V-GLOSS's ability to generate a *single* description for each class, without relying on ensembling. We then evaluate the V-GLOSS description of each class against its WordNet gloss.

To construct this set of class descriptions, which we view as a silver dataset of such descriptions, we generate a *single*, *normal* description for each ImageNet class via greedy decoding. We generate only *normal* descriptions because they outperform *contrastive* ones when only a single description is used. We call the resulting dataset *V-GLOSS Silver*.

We extrinsically evaluate V-GLOSS Silver by using it for the ZSIC and ZSCIG tasks, and comparing the results to those achieved using the 1-Template baseline, and WordNet glosses. We do not compare V-GLOSS Silver to CuPL or other previous methods which do not produce a single description for each class.

Technical Details

ZSIC We employ CLIP (Radford et al., 2021), which comprises an image encoder and a text encoder, as the ZSIC backbone model. Our procedure consists of three steps: First, we use the CLIP text encoder to create an aggregate representation for each class based on its description(s). Then, at test time, we employ the CLIP image encoder to generate a representation of the input image. Finally, we predict the class

	ZSIC	ZSCIG	
	Accuracy ↑	Inception ↑	FID ↓
Baseline (1-Template)	71.0	99.7	25.7
WordNet Glosses	44.7	58.5	30.0
V-GLOSS Silver	**72.3**	**109.6**	**20.0**

Table 3.4: Extrinsic evaluation on the tasks of ZSIC and ZSCIG. ↓ means that lower is better.

which maximizes the cosine similarity between the representation of its description(s), and the image representation (see Figure 3.1c). We evaluate the predictions using top-1 accuracy.

ZSCIG For ZSCIG (see Figure 3.1b), we condition Stable Diffusion (Rombach et al., 2022) on each class description before generating an image. We use a guidance scale of 7.5 and run 50 diffusion steps. We evaluate the generated images using Inception and FID scores.

Results

The results of Experiment 1 are shown in Table 3.4. Based on our extrinsic evaluation n the ZSIC and ZSCIG tasks, *V-GLOSS Silver* descriptions yield better performance compared to baseline and WordNet Glosses. On ZSIC, we improve accuracy by 1.3%; on ZSCIG, we improve Inception and FID scores by 9.9 and 5.7, respectively. This demonstrates the effectiveness and utility of V-GLOSS: the visually grounded descriptions V-GLOSS generates yield better results on ZSIC and ZSCIG.

Analysis

V-GLOSS Silver descriptions are considerably more detailed, more expressive, and better visually grounded than their WordNet gloss counterparts (see Figure 3.1). Specifically, we observe that V-GLOSS descriptions make greater use of descriptive words and phrases, e.g. *spiral, brown, green, thick, small*, etc.

Method	Model	Accuracy (%) on Datasets				# Parameters
		ImageNet	CIFAR-100	CIFAR-10	STL-10	
Baseline (1-Template)	ViT RN50	72.4 68.7	77.3 57.7	95.2 81.0	99.5 98.4	0
Template Ensemble	ViT RN50	76.2 73.2	77.9 61.3	96.2 86.8	99.4 98.3	0
CuPL	ViT	76.7	-	-	-	175B
CuPL + Template Ensemble	ViT RN50	77.6 75.1	- -	- -	- -	175B
V-GLOSS *(Normal-Only)*	ViT RN50	77.3 73.3	77.5 63.5	95.6 86.8	99.4 98.3	6.1B
V-GLOSS *(Normal + Contrastive)*	ViT RN50	**78.5** 74.5	**78.2** 64.6	**97.0** 87.8	**99.6** 98.8	6.1B

Table 3.5: Top-1 accuracy on ZSIC. ViT and RN are Transformer- and ResNet-based CLIP variants.

3.5.5 Experiment 2: ZSIC

Our second experiment assesses the effectiveness of V-GLOSS descriptions in facilitating ZSIC. The details for the ZSIC pipeline are largely similar to those described in Experiment 1 (Section 3.5.4), except that we generate an ensemble of descriptions per class, as opposed to only one description. We also experiment with two image encoder variants: ViT (Dosovitskiy et al., 2020) and RN50 (He et al., 2016). For all baselines and methods (Section 3.5.3, Section 3.4.2), we follow the same evaluation procedure after generating class descriptions.

Technical Details

We generate class descriptions using the 6.1B-parameter Cohere LLM[3]. We choose Cohere over alternatives due to its extensive cost-free availability, reducing the cost of our experiments. Cohere has comparable performance to the similarly-sized Instruct-GPT (Brown et al., 2020; Ouyang et al., 2022) variant, as demonstrated by Liang et al. (2022) across various benchmarks. Therefore, we do not gain any advantage by

[3] https://docs.cohere.com/docs/models

using Cohere instead of InstructGPT.

When generating class descriptions with *normal* V-GLOSS, we use a temperature of 2.5 to produce an ensemble of 50 descriptions per class. When generating *contrastively*, we use a temperature of 1.5 to generate an ensemble of 20 descriptions per class. Like Pratt et al. (2022), we observe that performance saturates around 50 descriptions for *normal* V-GLOSS, but we also observe saturation at around 20 descriptions for *contrastive* V-GLOSS. Based on tuning on development data, we set $N = 5$, $\lambda = 0.5$, and $k = 4$ (see Algorithm 2). In total, we obtain 70 class descriptions. During generation, we set the maximum number of tokens to 35, but also terminate generation when the *boundary parameter* or *newline* token is reached.

Results

The results of Experiment 2 are shown in Table 3.5. V-GLOSS yields better accuracy than the baseline by an average of 3.60% overall (2.22% with ViT and 4.98% with RN50). V-GLOSS also outperforms Template Ensemble and CuPL + Template Ensemble, by 1.21% and 0.15% respectively. This improvement is especially notable since the top 15 results on the ImageNet benchmark differ by less than 1% accuracy.[4] In addition, we make the following observations. (1) V-GLOSS (*Normal + Contrastive*) surpasses V-GLOSS (*Normal-Only*), by an average of 0.91% accuracy. (2) We outperform *CuPL + Template Ensemble* using an LLM with 28.7x fewer parameters. (3) The RN backbone (He et al., 2016), which is generally less capable than ViT (Dosovitskiy et al., 2020), sees a more significant benefit from the *V-GLOSS* method, on average 3.8%. (4) For STL-10, V-GLOSS matches the top-performing supervised system (Gesmundo, 2022) with a score of 99.6%. We also note that the *contrastive* component is more helpful on the larger datasets: CIFAR-100 and ImageNet, which have more opportunities for mutual ambiguity between different classes, than on the smaller ones: CIFAR-10 and STL-10. Concretely, this improvement is 1.05%, on

[4] https://paperswithcode.com/sota/image-classification-on-imagenet

average. Later, in Section 3.6, we discuss these results and their implications more extensively.

Analysis

In Section 3.1, we pointed out several problems in previous methods. Here, we carefully analyze how our V-GLOSS method addresses these issues.

Label Ambiguity: Without adequate context, text models may fail to grasp the intended meaning of a polysemous word. *Crane* is a polysemous word, and ImageNet (Deng et al., 2009) has two classes that refer to different senses of the word: *construction machine* and *wading bird*, but use the same label. Thus, in *1-Template*, for example, both classes have the same description. This point highlights an important benefit of linking classes to WordNet, which resolves such ambiguity. Empirically, when compared with a ViT backbone to the *Lex Baseline*, our accuracy on CRANE (machine) and CRANE (bird) increase from 0% and 46% to 76% and 78%, respectively.

Relationship Between Performance & Context: When comparing the baselines to the other methods, we observe that accuracy generally improves as the amount of surrounding context increases. On one hand, if a sentence consists of *"my crane."* alone, the sense of *crane* is unclear. On the other, if the sentence is *"my construction crane,"* the meaning of *crane* becomes clear. We see that providing additional context helps to disambiguate words. When a description provides more useful context, models can form better representations of specific classes. By comparing V-GLOSS to the baselines (see Table 3.5), we can observe that the benefits of additional context extend to the language-vision setting. Concretely, providing visually-grounded context in the description improves performance.

Class Granularity: We consider pairs of classes that are similar enough to be mistaken, such as ALLIGATOR and CROCODILE. In WordNet, relationships between synsets are modeled through *is-a* (hyponymy-hypernymy) and *part-of* (meronymy-holonymy) relationships. For example, CROCODILIAN is a hypernym of both ALLIGATOR and CROCODILE, while only ALLIGATOR is a holonym of SNOUT, since alligators have snouts while crocodiles do not. Using our contrastive algorithm, we generate descriptions that highlight how images of a CROCODILE should depict a greener animal with a rounded snout. Empirically, using ViT, the average accuracy of V-GLOSS across these two classes jumps from 36% to 68% when contrastive glosses are used. This improvement highlights the effectiveness of our contrastive V-GLOSS variant in reducing false positives between visually similar classes.

Attention Maps: We analyze the model's attention maps to understand V-GLOSS's impact. Figure 3.4 shows the attention map for V-GLOSS (see Table 3.1 for descriptions), indicating effective utilization of visually-relevant context. Conversely, Figure 3.5 shows the attention map for the WordNet glosses (baseline), where the attention score on *bottle* is 3.5x higher, implying less distraction in V-GLOSS. These maps empirically demonstrate success in steering the model's attention toward relevant context, thus improving classification accuracy across different classes and descriptions.

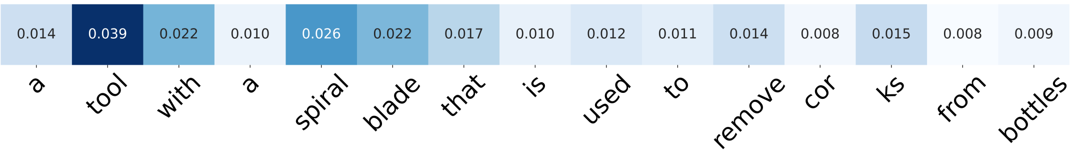


Figure 3.4: V-GLOSS Attention Map

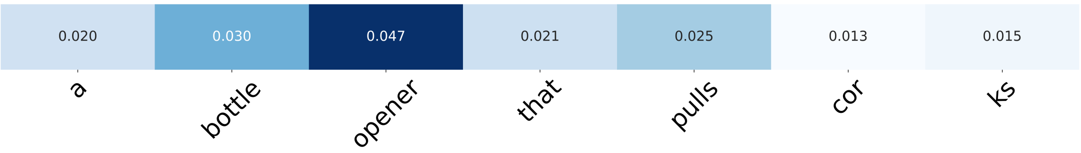


Figure 3.5: WordNet Gloss Attention Map

3.6 Discussion

When looking at our results, a pertinent question arises: Why does an SKB, such as WordNet, help us do better on tasks related to vision? In this section, we formulate two insights on how the synergy between SKBs and LLMs supports our improvements.

Insight #1: SKBs represent concepts precisely When LLMs are prompted with better information, they produce better output (Borgeaud et al., 2022). WordNet provides a precise representation of a class and its relationship to other classes, leaving minimal room for ambiguity. Afterward, we can prompt an LLM with this precise information to produce unambiguous and high-quality class descriptions.

Insight #2: Pre-trained LLMs hold significant world knowledge By virtue of their pre-training over an extensive corpus of text, LLMs gain non-trivial knowledge about the world, that the language describes. This proves useful in our task because we can extract visual descriptions for all kinds of concepts.

Insight #3: Semantic similarity is a useful proxy for visual similarity WordNet models lexical semantics as a graph (see Figure 3.3), with synsets as nodes and *is-a* relationships as directed edges. The distance between different nodes reflects the level of semantic similarity and is by extension an indicator of the level of visual similarity between synsets. ALLIGATOR and CROCODILE are semantically similar because they are both kinds of CROCODILIAN, but they are visually similar as well (see Table 3.2). Semantic similarity informs what classes we distinguish with our contrastive descriptions and why they work (see Table 3.3). This is because semantic and visual similarity are usually correlated (Brust and Denzler, 2018).

3.7 Conclusion

This work focuses on the task of generating visual concept descriptions for use in two downstream tasks: zero-shot image classification and zero-shot class-conditional image generation. We employ a novel technique that combines semantic knowledge bases (SKBs) and transformer-based, pre-trained large language models (LLMs) to produce high-quality visual descriptions. In addition to providing significant empirical improvements, we gain useful insights into the behavior of both SKBs and LLMs. First, we learn that the lexico-semantic information tied to concepts from SKBs is enough to condition an LLM to generate visually-grounded text. Second, we learn that LLMs, after being pre-trained solely on text data, possess latent knowledge about the visual properties of concepts which can be accessed and leveraged using our novel V-GLOSS method. And by extension, we learn that we can ground concepts more adequately to vision with V-GLOSS. In short: visually-grounded concept descriptions improve the accuracy of zero-shot image classification and generation models. This exemplifies the strong relationship between language and vision modalities and opens up the possibility of LLMs being used, without fine-tuning, for more multi-modal tasks in the future.

Chapter 4

Conclusion

This thesis evaluates the hypothesis that visually-grounded descriptions, which emphasize the visual attributes of a concept, can enhance performance on language-vision tasks. Our extensive experiments confirm our hypothesis, demonstrating that visually-grounded descriptions, produced using pre-trained, transformer-based large language models and semantic knowledge bases (SKBs), can significantly improve performance on these tasks. We have shown advancements in two specific tasks: visual word sense disambiguation (V-WSD), which we focus on in Chapter 2, as well as visual concept description (VCD), which is applied to two downstream tasks: zero-shot image classification (ZSIC), and zero-shot class-conditional image generation (ZSCIG), both of which are addressed in Chapter 3. Our results provide compelling evidence to support our hypothesis.

In Chapter 2, we proposed an image-scoring algorithm for the V-WSD task. This algorithm calculates the sum of the similarity between (1) the context and candidate images, and (2) the context and potential senses of the focus word. Most importantly, we demonstrated that enhancing the original context with more visually grounded information, as our hypothesis suggests, significantly improves performance compared to using the original context directly. Our simple-but-effective method significantly outperforms the baselines and ranks 13th out of 57 final contestants, demonstrating the utility of our approach. Moreover, our method is applicable to various languages, with only minor performance degradation after translating the context to English. These results robustly support our hypothesis.

In Chapter 3, we tackled the upstream task of VCD before addressing its related downstream tasks: ZSIC and ZSCIG. We demonstrated that we could generate visually-grounded descriptions using specific lexico-semantic information from WordNet to prompt a pre-trained, transformer-based large language model. This approach extends on Chapter 2 by eliminating ambiguity issues due to polysemy, such as the word “crane” which could refer to a bird or a machine. It also allows us to clearly differentiate similar concepts, such as “alligator” versus “crocodile”, preventing our system from confusing them. Our results showed significant improvements compared to the baselines and previous methods, further supporting our primary hypothesis.

Our work makes three core contributions. Firstly, we demonstrate that lexico-semantic knowledge can be used to prompt a language model to produce rich, visual descriptions, thereby improving language-to-vision grounding. Secondly, we introduce contrastive prompting, a method that encourages a language model to generate descriptions that highlight the most visually distinctive differences between two concepts. Lastly, we create a new silver dataset of visual descriptions for all 1,000 classes in ImageNet. The validity of all our contributions is experimentally verified.

4.1 Limitations and Future Work

Our work’s primary limitation is the use of visual descriptions instead of actual images, as some visual concepts are challenging to express in language. In the future, we aim to develop methods that understand vision natively. We also acknowledge other limitations and propose potential solutions for future work.

Our next limitation is that the dataset must be mapped to an SKB. The process of mapping the dataset to WordNet, while a one-time step, is not fully automatic. We aim to automate this process in future work, potentially by selecting a synset based on the similarity between sample class images and potential senses of the class label.

Our final limitation is that we are limited in terms of language, dataset class count, and the size of the SKB. Our focus on English may limit the applicability of our method to ZSIC or ZSCIG tasks in other languages. Some classes are strongly associated with non-English languages. Our largest evaluation dataset, ImageNet (Deng et al., 2009), only covers 0.64% of WordNet with its 1,000 classes. We plan to evaluate our methods on a larger ImageNet set, ImageNet-21k, which would cover 14.06% of WordNet. While our method can be applied to BabelNet (Navigli and Ponzetto, 2012), which has over 1.5 billion synsets, we have focused on WordNet, which has 155,287. We plan to explore alternative SKBs such as BabelNet or non-English wordnets, both of which are multilingual.

4.2 Final Remarks

This thesis embarked on a journey to explore the potential of lexico-semantic knowledge combined with language models in aiding grounding. Our hypothesis has been conclusively confirmed, as the studies in Chapters 2 and 3 effectively demonstrate the value of visual descriptions for grounding concepts in the visual domain. We hope to engage in future work that more intimately explores the close relationship between language and vision.

References

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. ConSeC: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, pages 1–5.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688.*

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745.*

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 1006–1017, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Clemens-Alexander Brust and Joachim Denzler. 2018. Not just a matter of semantics: the relationship between visual similarity and semantic similarity. *arXiv preprint arXiv:1811.07120*.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Eugene Charniak. 1996. *Statistical language learning*. MIT press.

Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.

Slawomir Dadas. 2023. OPI at SemEval-2023 task 1: Image-text embeddings and multimodal information retrieval for visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 155–162, Toronto, Canada. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*.

Spandana Gella, Frank Keller, and Mirella Lapata. 2019. Disambiguating visual verbs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):311–322.

Andrea Gesmundo. 2022. A continual development methodology for large-scale multitask dynamic ml systems. *arXiv preprint arXiv:2209.07326*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Daniel Jurafsky and James H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3 edition. Draft Version, Stanford University; University of Colorado at Boulder.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 547–554, Sydney, Australia. Association for Computational Linguistics.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Michael Ogezi, Bradley Hauer, and Grzegorz Kondrak. 2023a. Visually-grounded descriptions improve zero-shot image classification. *arXiv preprint arXiv:2306.06077*.

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, and Grzegorz Kondrak. 2023b. Ualberta at semeval-2023 task 1: Context augmentation and translation for multilingual visual word sense disambiguation.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Rahul Patil, Pinal Patel, Charin Patel, and Mangal Verma. 2023. Rahul patil at SemEval-2023 task 1: V-WSD: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1271–1275, Toronto, Canada. Association for Computational Linguistics.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. 2021. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*.

Sarah Pratt, Rosanne Liu, and Ali Farhadi. 2022. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*.

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. 2021. Learning transferable visual models from natural language supervision.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Amir Ahmadi Sajjad Ayoubi, Navid Kanaani. 2022. Clipfa: Connecting farsi text and images. https://github.com/SajjjadAyobi/CLIPfa.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Dean R Spitzer. 1975. What is a concept? *Educational Technology*, 15(7):36–39.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Sebastiano Vascon, Sinem Aslan, Gianluca Bigaglia, Lorenzo Giudice, and Marcello Pelillo. 2021. Transductive visual verb sense disambiguation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3049–3058.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 6000–6010, USA. Curran Associates Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6229–6240, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Qihao Yang, Yong Li, Xuelin Wang, Shunhao Li, and Tianyong Hao. 2023. TAM of SCNU at SemEval-2023 task 1: FCLL: A fine-grained contrastive language-image learning model for cross-language visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 506–511, Toronto, Canada. Association for Computational Linguistics.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xudong Zhang, Tiange Zhen, Jing Zhang, Yujin Wang, and Song Liu. 2023. SRCB at SemEval-2023 task 1: Prompt based and cross-modal retrieval enhanced visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 439–446, Toronto, Canada. Association for Computational Linguistics.

Appendix A: Text-Conditioned Image Segmentation

A.1 Success Mode

In the successful case of this system, we see that we are able to properly segment the object based on the text provided. See Figure A.4 for details.

A.2 Failure Mode

In the failure case of this system, we see that we cannot confidently segment the object based on the text provided. See Figure A.8 for details.

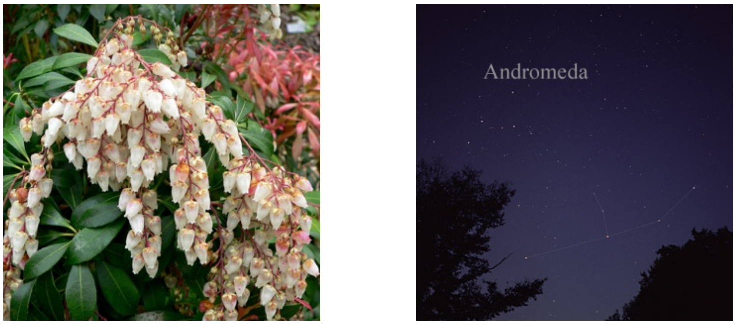


Figure A.1: Original images from the dataset are depicted on the left: ANDROMEDA and on the right: ANDROMEDA.

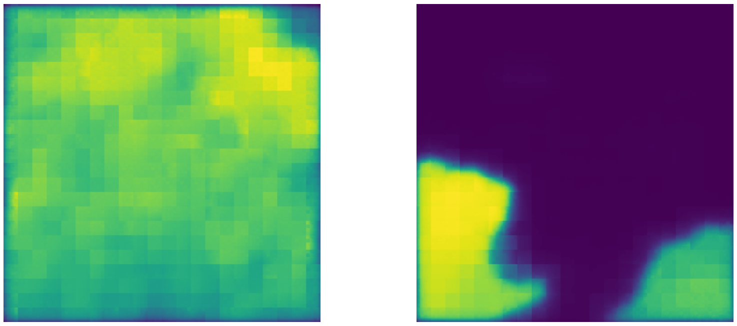


Figure A.2: Conditioned on the full ***"andromeda tree"***

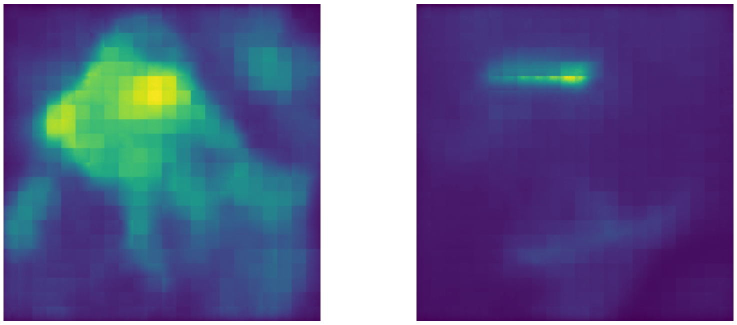


Figure A.3: Conditioned on *"andromeda"*

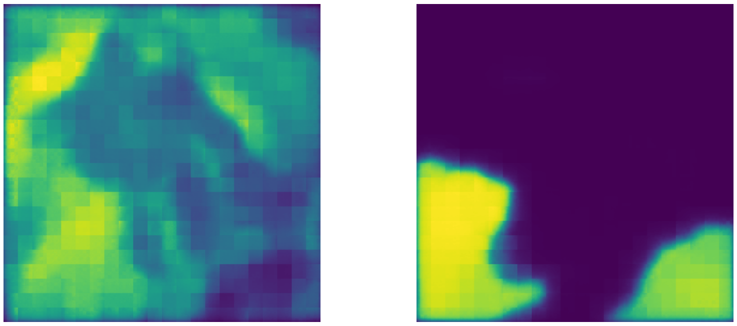


Figure A.4: Conditioned on *"tree"*

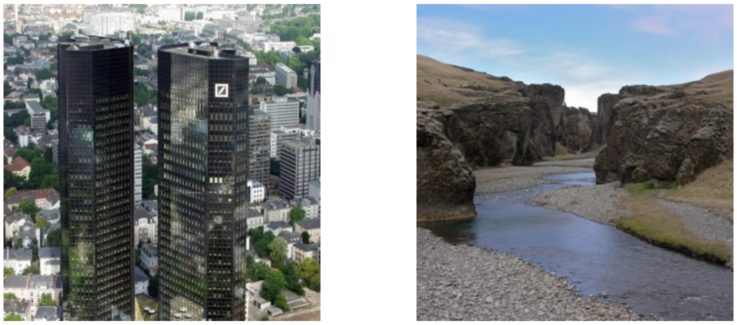


Figure A.5: Original images from the dataset are depicted on the left: BANK (finance) and on the right: BANK (river).

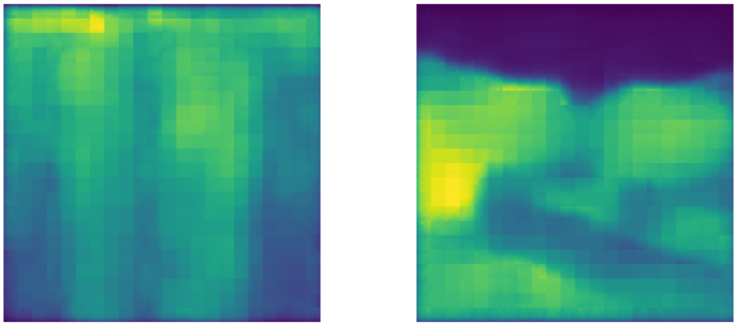


Figure A.6: Conditioned on the full ***"bank erosion"***

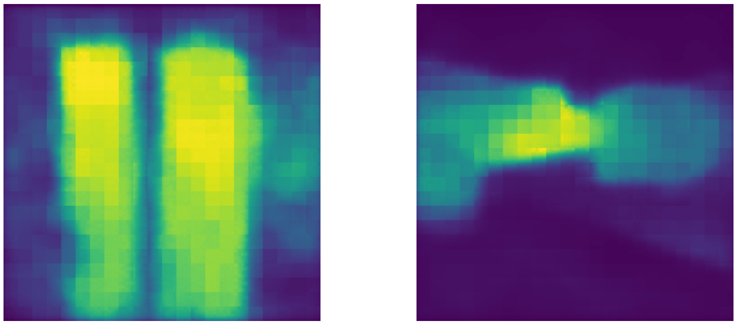


Figure A.7: Conditioned on *"bank"*

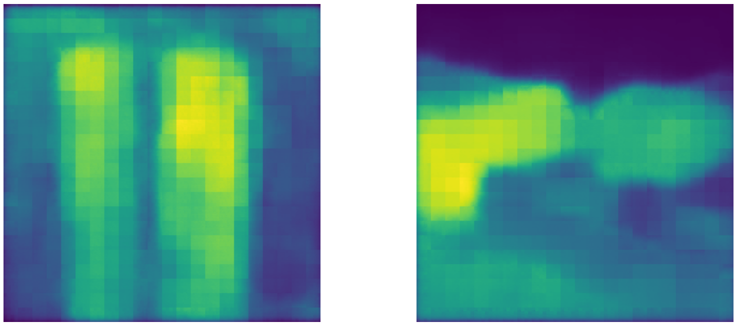


Figure A.8: Conditioned on *"erosion"*