Semantically-Prompted Language Models Improve Visual Descriptions

Anonymous ACL submission

Abstract

Language-vision models like CLIP have made significant progress in zero-shot vision tasks, such as zero-shot image classification (ZSIC). However, generating specific and expressive visual descriptions remains challenging as descriptions from current methods are ambiguous and lack granularity. To tackle these issues, we propose V-GLOSS: Visual Glosses, a novel method that consists of two key ideas. The first is Semantic Prompting, which conditions a language model on structured semantic knowledge. The second, but more pivotal, is our contrastive algorithm, which elicits fine distinctions between similar concepts. With both ideas, we demonstrate that V-GLOSS improves visual descriptions and achieves state-of-the-art results on both broad and fine-grained benchmark ZSIC datasets, such as ImageNet, STL-10, FGVC Aircraft, and Flowers 102. V-GLOSS also boosts image-generation performance. Finally, we introduce a quality-tested silver dataset with descriptions generated by V-GLOSS for all ImageNet classes.

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.

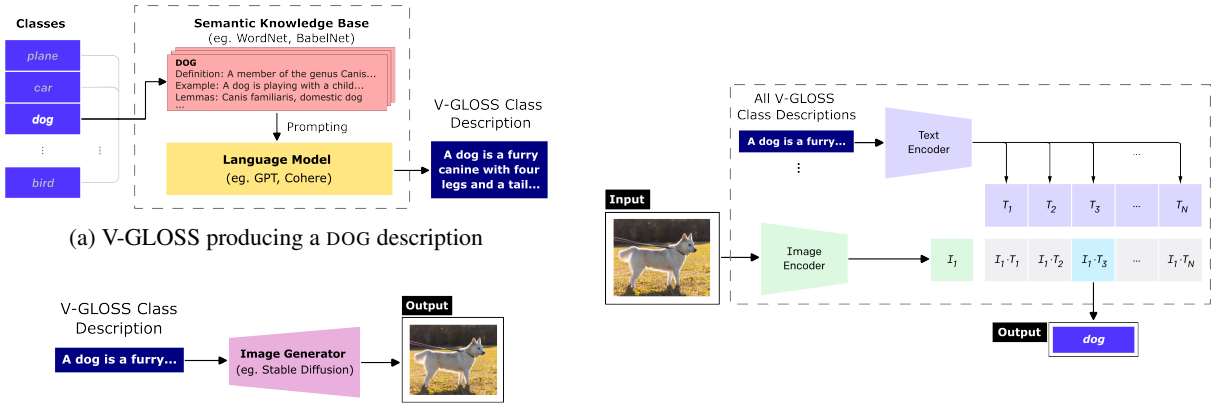
Table 1: A qualitative comparison between baseline glosses and V-GLOSS descriptions for some ImageNet classes. Our method describes what a class *looks like*, instead of what it *does* or *is*. Many more examples are shown in Table 6.

1 Introduction

Language-vision models (Radford et al., 2021; Jia et al., 2021) have made significant progress in zero-shot vision tasks. However, in agreement with Betker et al. (2023), we hypothesize that their accuracy is limited by a lack of visual concept descriptions that are both expressive and specific, that is, glosses that describe what images depicting a concept look like. In this work, we investigate this hypothesis by creating and testing a new method for producing visual descriptions with only pre-trained language models.

High-quality visual descriptions are crucial in tasks such as zero-shot image classification and text-based image retrieval. Improved descriptions facilitate the creation of more useful representations. These are essential in producing robust and adaptable methods capable of understanding novel and specific visual attributes without re-training.

Existing approaches to generating visual 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 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



(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 1: For the DOG class, we depict (a) V-GLOSS's architecture (Section 4.2.1), along with adaptations: (b) zero-shot image classification (ZSIC) (Section 5.4.1) and (c) zero-shot class-conditional image generation (ZSCIG) (Section 5.4.1)

issues limit the performance of existing models (Radford et al., 2021).

To address these challenges, we introduce V-GLOSS, a novel method that leverages language models (LMs) and semantic knowledge bases (SKBs) to generate improved visual descriptions – **v**isual **gloss**es. Table 1 shows some examples. By combining structured semantic information from SKBs such as WordNet (Miller, 1998), and BabelNet (Navigli and Ponzetto, 2012), with a contrastive algorithm to finely 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 results on benchmark datasets such as ImageNet (Deng et al., 2009), FGVC Aircraft (Maji et al., 2013), and Flowers 102 (Nilsback and Zisserman, 2008) in the zero-shot setting, and STL-10 (Coates et al., 2011) in both the zero-shot and supervised settings. Additionally, we introduce V-GLOSS Silver, a silver dataset constructed by V-GLOSS, which consists of a visual description for each ImageNet class. We show that V-GLOSS Silver is useful for zero-shot language-vision tasks such as ZSIC and ZSCIG, comparing favorably to WordNet glosses.

2 Tasks

Our main task is to generate a 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 generating 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. Visual descriptions of classes provide useful information which can facilitate both tasks, by making it easier to either recognize or generate images of each class. Therefore, by developing a novel method to improve the generation of such descriptions, we hypothesize that performance on ZSIC and ZSCIG can be improved.

3 Related Work

Language Models The advent of transformer-based language models has 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 by

their number of parameters and quantity of training data, they exhibit emergent abilities such as few-shot and zero-shot learning (Wei et al., 2022).

Language-Vision Models Significant strides have been made in the field of language-vision models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These models apply contrastive pre-training approaches on large image-text datasets, leading to improved representation learning for both text and images and enhanced performance on several multi-modal tasks (Mokady et al., 2021; Song et al., 2022). Further advancements have been achieved by scaling up pre-training and incorporating auxiliary training objectives (Pham et al., 2021; Yu et al., 2022).

Producing Descriptions & Prompting The generation of descriptions and prompting has been explored in various studies. Radford et al. (2021) introduced the template ensemble (TE) method, which uses a custom set of class labels and a fixed set of templates. Each label is inserted into these templates, and the completed templates for each class are aggregated into a single representation of the class. The CuPL method (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) fine-tuned GPT models (Radford et al., 2018, 2019) to rephrase image-generation prompts, resulting in improved images. (Zhou et al., 2022) learned soft prompts that improve performance, but are intractable to humans. In this work, we prompt language models with semantic knowledge to generate visual descriptions.

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 information the SKB contains. We then introduce of 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 class descriptions produced using V-GLOSS.

4.1 Mapping Classes to 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. 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 most frequent, but the RAY in the dataset refers to the sea creature. We show mis-mapped CIFAR-100 classes in Table 7 of the appendix. If no suitable synset is found in WordNet, as frequently happens for FGVC Aircraft and Flowers 102, we fall back to BabelNet or our manually-produced definitions.

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

Figure 2: Class descriptions for PLATYPUS generated by two different methods that use LMs. Input prompts, output descriptions, and **plugged values** are shown.

4.2 V-GLOSS

We discuss the two variants of V-GLOSS below, *normal* and *contrastive*. In both, for each class, we produce multiple descriptions resulting in an ensemble.

4.2.1 *Normal* V-GLOSS

We generate normal descriptions via in-context learning with an LM, beginning by providing the LM 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.

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

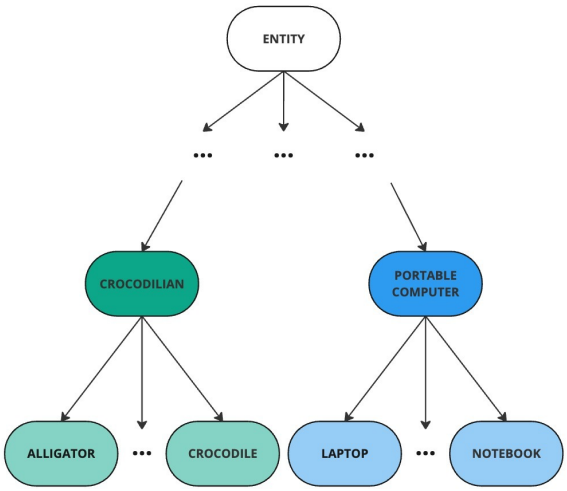


Figure 3: A sample of an SKB hypernym hierarchy. For *contrastive* prompting, we only distinguish classes that are semantically similar to the target class, like ALLIGATOR to CROCODILE.

We selected these to expose the model to ambiguous class labels (*bat*), a natural object (*eagle*), and an artificial object (*television*). For each class, we obtain the hypernyms, hyponyms, usage examples, synonyms, and gloss of the sense to which the class is mapped, and provide this to the LM. Figure 2b shows a session with the LM, beginning with the example of *eagle*, with output generated for the class *platypus*. Table 1 compares our descriptions to baseline glosses.

4.2.2 *Contrastive* V-GLOSS

During development, we observed that many errors were caused by false positives involving visually similar classes. For example, the classes CROCODILE for ALLIGATOR refer to similar-looking animals, and are often confused with one another. Moreover, ImageNet contains 120 distinct dog species, and the fine-grained datasets contain only airplanes (FGVC Aircraft) or flowers (Flowers 102). The contrastive variant of V-GLOSS is designed to address these issues by using semantic similarity 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 LM prompt, e.g. we generate a description for an ALLIGATOR *but not* a CROCODILE, using the same prompt structure as for normal V-GLOSS.

We create a similarity matrix M as follows:

$$M_{i,j} = Sim(S[i], S[j]) \quad (1)$$

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

$Sim(s_1, s_2)$ 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 tree (Figure 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 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 1.

In Algorithm 1, 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 indicates the maximum number of classes to generate contrastive descriptions for. k is the number of distinct descriptions to generate for a class pair. LM_c takes in the $target$ class, a neighbor class, and k, then prompts the LM to generate k descriptions that distinguish the $target$ and neighbor classes. In summary, for each class, Algorithm 1 identifies the classes most similar to it, excluding itself, and generates descriptions that distinguish them. Table 2 compares the normal and contrastive descriptions for ALLIGATOR and CROCODILE; note that distinguishing features of the two classes are included in the LM's output. Table 3 shows examples of classes with high false positive rates, and the classes they are contrasted with.

5 Evaluation

In this section we present our evaluation of V-GLOSS, alongside comparable methods. We describe our datasets, evaluation metrics, baselines,

Algorithm 1 Generate Contrastive Descriptions: We generate contrastive descriptions to help distinguish the most similar classes.

Require: M: Equation 1 result
Require: λ, N, k: Hyperparameters
Require: S: All classes in dataset, $\mathcal{D}$
Require: LM_c: LM prompted contrastively

```
G ← empty |S|-list for class descriptions
for i ← 0 to |S| − 1 do
    target ← S[i]
    S* ← top N classes : λ ≤ M_{i,*} ≤ 1
    for s* in S* do
        samples ← LM_c(target, s*, k)
        G[i].insert(samples)
return G
```

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: False positives and their counts vs. classes selected by the contrastive algorithm (see Equation 1 and Algorithm 1). Hits and misses are shown.

previous methods, and experiments. To ensure robustness, we report the mean over five random seeds in Tables 4 and 5.

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. STL-10 (Coates et al., 2011) comprises 100,000 test samples and is designed for unsupervised learning. FGVC Aircraft (Maji et al., 2013) contains 10,200 images across 102 aircraft model variants, with 100 images per variant. Flowers 102 (Nilsback and Zisserman, 2008) features 102 flower categories common in the UK, with 40 to 258 images per category. For CIFAR-10, CIFAR-100, STL-10, FGVC Aircraft, and Flowers 102, which are not pre-mapped to WordNet, we employ the two-step process detailed in Section 4.1 to map each class to a synset.

Experiment 1 (Section 5.4) involves ImageNet alone and covers both the ZSCIG and ZSIC tasks. In contrast, Experiment 2 (Section 5.5), 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.

We selected these datasets to evaluate the following properties of V-GLOSS:

1. **Performance on common 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. **Proficiency in fine-grained conceptual distinctions.** Although some datasets (ImageNet and CIFAR) cover diverse domains, we importantly consider fine-grained datasets like FGVC Aircraft and Flowers 102. This enables testing our method's ability to distinguish very similar classes (e.g., distinguishing between closely related species or types).

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 for comparison.

5.3 Baseline & Previous Methods

In this section, we describe the methods to which we compare V-GLOSS. 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 for each description.

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."*

The next approach we consider is **Template Ensemble** (Radford et al., 2021), which generates an ensemble of descriptions for a class by inserting the class label into each of a set of 80 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.

CuPL (Pratt et al., 2022) also generates an ensemble of descriptions for each class. The descriptions are generated by prompting a 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.

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 may produce more than one gloss for each class.

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

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

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

ZSCIG For ZSCIG (see Figure 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.

5.4.2 Results

The results of Experiment 1 are shown in Table 4. Based on our extrinsic evaluation in 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: our visual descriptions yield better results on ZSIC and ZSCIG.

5.4.3 Analysis

V-GLOSS Silver descriptions are considerably more detailed, more expressive, and better grounded than their WordNet counterparts (see Figure 1). Specifically, we observe that V-GLOSS descriptions make greater use of descriptive words and phrases, e.g. *spiral, brown, green, thick, small*, etc.

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 5.4), except that we generate an ensemble of descriptions per class, as opposed to only one description. We

Method	Model	Accuracy (%) on Datasets						# LM Parameters
		ImageNet	CIFAR 100	CIFAR 10	STL 10	FGVC Aircraft	Flowers 102	
1-Template Baseline	ViT	72.4	77.3	95.2	99.5	31.7	77.6	0
	RN50	68.7	57.7	81.0	98.4	27.4	71.6	
Template Ensemble	ViT	76.2	77.9	96.2	99.4	32.9	78.5	0
	RN50	73.2	61.3	86.8	98.3	29.7	74.3	
CuPL	ViT	76.7	**78.6**	95.8	-	36.1	79.7	175B
V-GLOSS *Normal-Only*	ViT	77.3	77.5	95.6	99.4	33.2	79.2	6.1B
	RN50	73.3	63.5	86.8	98.3	30.8	75.1	
V-GLOSS (*Normal* + *Contrastive*)	ViT	**78.5**	78.2	**97.0**	**99.6**	**38.7**	**81.3**	6.1B
	RN50	74.5	64.6	87.8	98.8	35.2	77.3	

Table 5: Top-1 accuracy on ZSIC. ViT-L14-336 and RN50x64 are Transformer- and ResNet-based CLIP variants. See Table 8 for more model variants.

also experiment with two image encoder variants: ViT (Dosovitskiy et al., 2020) and RN50 (He et al., 2016). For all baselines and methods (Section 5.3, Section 4.2.1), we follow the same evaluation procedure after generating class descriptions.

5.5.1 Technical Details

We generate class descriptions using the 6.1B-parameter Cohere LM[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 InstructGPT (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 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 1). 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.

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

5.5.2 Results

The results from Experiment 2, as shown in Table 5, primarily underscore the significant efficiency and accuracy gains of V-GLOSS (*Normal* + *Contrastive*) over CuPL. Notably, V-GLOSS achieves these gains with substantially fewer parameters, emphasizing its efficiency.

Key findings include: (1) V-GLOSS demonstrates an average accuracy improvement of 6.0% over the baseline (6.1% for ViT and 5.8% for RN50). (2) Compared to the Template Ensemble, V-GLOSS shows an average improvement of 1.8%. (3) Against its own variant, V-GLOSS *Normal-Only*, the *Normal* + *Contrastive* version improves accuracy by an average of 1.2%.

The standout observation, however, is the comparison between CuPL and V-GLOSS. Despite having 28.7 times fewer parameters than CuPL (6.1B compared to CuPL's 175B), V-GLOSS exhibits notable improvements. This is especially evident in fine-grained datasets (FGVC Aircraft and Flowers 102) with an average improvement of 5.4%, compared to the 2.4% average improvement on general datasets (ImageNet, CIFAR, STL). Furthermore, on larger datasets like CIFAR-100 and ImageNet, V-GLOSS shows an average improvement of 4.9%, highlighting its robustness in complex classification scenarios with a more resource-efficient model.

These results not only demonstrate the effectiveness of the contrastive algorithm within V-GLOSS but also highlight the potential of using more efficient models in complex image classification tasks.

For a detailed discussion on these implications, see Section 6.

5.5.3 Analysis

In Section 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.

Performance-Context Relationship: 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 5), we can observe that the benefits of additional context extend to the vision-language 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.

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 LMs supports our improvements.

Insight #1: SKBs represent concepts precisely When LMs 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 LM with this precise information to produce unambiguous and high-quality class descriptions.

Insight #2: Semantic similarity is a useful proxy for visual similarity WordNet models lexical semantics as a tree (see Figure 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 2). Semantic similarity informs what classes we distinguish with our contrastive descriptions and why they work (see Table 3). This is because semantic and visual similarity are highly correlated.

7 Conclusion

This study concentrates on generating visual class descriptions for ZSIC and ZSCIG tasks. We utilize a unique method that merges Semantic Knowledge Bases (SKBs) and Language Models (LMs) to create high-quality descriptions. Our findings reveal that the semantic information from SKBs can condition an LM to generate improved descriptions with better accuracy and expressiveness. We also show that our contrastive algorithm improves fine-grained discrimination between similar concepts. Our findings underscore the strong relationship between language and vision, suggesting potential for LMs in future multi-modal tasks.

Limitations

The dataset must be mapped to an SKB. As described earlier, mapping the dataset to WordNet, although a one-time step, is not fully automatic. In future work, we look to fully automate this step, possibly by selecting a synset based on the similarity between sample class images and potential senses of the class label.

We are limited in terms of language, dataset class count, and our SKB's size. First, our English-focused stance may prove a limiting factor in our method being applied to ZSIC or ZSCIG tasks based in other languages. Some classes are strongly related to non-English languages.

Second, our largest evaluation dataset, ImageNet (Deng et al., 2009), has 1,000 classes, representing just 0.64% coverage of WordNet. We look forward to evaluating our methods on a larger ImageNet set: ImageNet-21k, which would cover 14.06% of WordNet.

Third, although our method can be applied to BabelNet (Navigli and Ponzetto, 2012), which has over 1.5 billion synsets, we focus on WordNet, which has 155,287. We look to explore alternative SKBs such as BabelNet, or non-English wordnets, both of which offer the benefit of being multilingual.

Ethics Statement

In normal use, we discover no direct ethical issues with our method. Note, however, that we may inherit ethical problems from the components used by our method. Both CLIP (Agarwal et al., 2021) and LMs (Liang et al., 2021) have independently been shown to exhibit some level of bias. Also, semantic resources such as WordNet (Miller, 1995) tend to focus on formalized concepts. This poses a problem if our method's use concerns people on the fringes of society.

We noted earlier that our method is mostly English-focused. This could be a source of bias if our method is applied in a multilingual context. We ask that people do not apply our method to real-world problems where multilingual knowledge is required. There is also the issue of semantic resources for low-resource languages not being extensive enough (Magueresse et al., 2020).

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*.

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*.

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.

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.

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.

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*.

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

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.

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

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

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*.

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. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729.

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*.

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.

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.

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*.

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.

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*.

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*.

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

A Appendices

The appendices contain Table 6 which compares WordNet glosses to V-GLOSS descriptions for the first 100 classes in ImageNet. Next, we show examples of the cases where our most frequent sense heuristic for mapping a class to WordNet failed in Table 7. Finally, in Table 8, we show a more detailed variant of Table 5 which compares multiple variants of the CLIP backbone. The authors of CuPL also combined their method with Template Ensemble. The resulting method, CuPL + Template Ensemble, combines the class descriptions from both methods and leads to marginally better performance.

We also briefly analyze V-GLOSS attention maps to better understand its impact on performance. Figure 4 shows the attention map for V-GLOSS (see Table 1 for descriptions), indicating effective utilization of visually-relevant context. Conversely, Figure 5 shows the attention map for the WordNet glosses, where the attention score on *bottle* is 3.5x higher, showing less distraction in V-GLOSS. These maps demonstrate success in steering the model's attention toward relevant context, thus improving classification accuracy across different classes and descriptions. We speculate that our descriptions also reduce distraction in images, but leave this to future work.

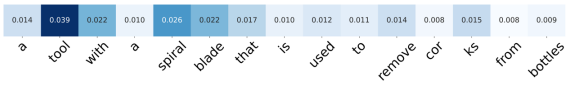


Figure 4: Attention map for V-GLOSS description

Figure 5: Attention map for WordNet gloss

A.1 WordNet vs. V-GLOSS

Table 6: Comparison of WordNet glosses and V-GLOSS descriptions for the first 100 ImageNet classes

Class	WordNet Gloss	V-GLOSS
tench	freshwater dace-like game fish of Europe and western Asia noted for ability to survive outside water	Freshwater fish of the family Cyprinidae, found in Europe and western Asia, with a long, slender body, a large mouth, and a forked tail.
goldfish	small golden or orange-red freshwater fishes of Eurasia used as pond or aquarium fishes	Small freshwater fishes of the family Cyprinidae, native to East Asia, introduced to many other parts of the world.
great white shark	large aggressive shark widespread in warm seas; known to attack humans	Large, white, shark-like fish with a large mouth and sharp teeth.
tiger shark	large dangerous warm-water shark with striped or spotted body	Large warm-water shark with a dark brown or black body, white underbelly, and white or yellowish-white dorsal fin.
hammerhead shark	medium-sized live-bearing shark with eyes at either end of a flattened hammer-shaped head; worldwide in warm waters; can be dangerous	Large shark with a flattened head and eyes at either end of the head.
electric ray	any sluggish bottom-dwelling ray of the order Torpediniformes having a rounded body and electric organs on each side of the head capable of emitting strong electric discharges	Sluggish bottom-dwelling rays of the order Torpediniformes with electric organs on each side of the head.
stingray	large venomous ray with large barbed spines near the base of a thin whiplike tail capable of inflicting severe wounds	Large, flat, wingless ray with a long, whip-like tail and sharp, barbed spines near the base.
rooster	adult male bird	Male bird larger than a hen with a crest of feathers on its head.
hen	adult female bird	Bird that lays eggs.
ostrich	fast-running African flightless bird with two-toed feet; largest living bird	Large flightless bird with a long neck and legs, large head, and long, pointed, featherless, yellow-brown or black beak.
brambling	Eurasian finch	Small brown bird with a black head and a white patch on its chest.
goldfinch	small European finch having a crimson face and yellow-and-black wings	Small bird with a bright yellow face and black wings.
house finch	small finch originally of the western United States and Mexico	Small bird with a bright red head, black back, and white breast.
junco	small North American finch seen chiefly in winter	Small brownish-gray bird with a white belly and a dark cap, found in open areas.
indigo bunting	small deep blue North American bunting	Small deep blue North American bunting.
American robin	large American thrush having a rust-red breast and abdomen	Small brown bird with a red breast and a black head and tail.
bulbul	nightingale spoken of in Persian poetry	Small bird with a long tail and a long, pointed beak.
jay	crested largely blue bird	Crested, largely blue bird with a crest on its head and a long tail.
magpie	long-tailed black-and-white crow that utters a raucous chattering call	Black-and-white crow with a long tail, often seen in groups.
chickadee	any of various small grey-and-black songbirds of North America	Small grey-and-black songbird of North America.
American dipper	small stocky diving bird without webbed feet; frequents fast-flowing streams and feeds along the bottom	Small bird with a black head, white breast and back, and white belly; short, thick, black bill and a black tail with white tips.
kite (bird of prey)	any of several small graceful hawks of the family Accipitridae having long pointed wings and feeding on insects and small animals	Large bird with a long pointed tail and a forked tail, used to catch insects and small animals.
bald eagle	a large eagle of North America that has a white head and dark wings and body	Large bird of prey with a white head and dark wings and body.
vulture	any of various large diurnal birds of prey having naked heads and weak claws and feeding chiefly on carrion	Large bird of prey with a bald head, hooked beak, and bare neck.
great grey owl	large dish-faced owl of northern North America and western Eurasia	Large owl with a round head, large eyes, short tail, white face, and gray body with a white patch on the back of the neck.

Continued on the next page...

Table 6 is continued from previous page

Class	WordNet Gloss	V-GLOSS
fire salamander	a kind of European salamander	Small amphibian with a long tail and a long, thin body covered in black and yellow spots.
smooth newt	small semiaquatic salamander	Small semiaquatic salamander with a long tail and a long, pointed snout.
newt	a newt in its terrestrial stage of development	Small amphibian with a long tail, a long, thin body and a short head.
spotted salamander	glossy black North American salamander with yellow spots	Glossy black amphibian with yellow spots.
axolotl	larval salamander of mountain lakes of Mexico that usually lives without metamorphosing	Salamander living in mountain lakes of Mexico, usually found in muddy water.
American bullfrog	largest North American frog; highly aquatic with a deep-pitched voice	Large amphibian with a greenish-brown back and dark brown or black belly; large head with bulging eyes and a long, pointed snout.
tree frog	any of various Old World arboreal frogs distinguished from true frogs by adhesive suckers on the toes	Small frog with a long, thin body, long, thin tail, and long, thin tongue.
tailed frog	western North American frog with a taillike copulatory organ	Small frog with a long, thin tail used for balance and jumping.
loggerhead sea turtle	very large carnivorous sea turtle; wide-ranging in warm open seas	Large, slow-moving, carnivorous sea turtle with a hard shell and long, pointed head.
leatherback sea turtle	wide-ranging marine turtle with flexible leathery carapace; largest living turtle	Large marine turtle with a leathery shell and long, pointed snout.
mud turtle	bottom-dwelling freshwater turtle inhabiting muddy rivers of North America and Central America	Turtle living in muddy rivers and lakes in North America and Central America.
terrapin	any of various edible North American web-footed turtles living in fresh or brackish water	Large, flat-bodied, freshwater turtle with a diamond-shaped shell and long tail.
box turtle	chiefly terrestrial turtle of North America; shell can be closed tightly	Large, slow-moving, terrestrial turtle with a hard shell and long tail.
banded gecko	any of several geckos with dark bands across the body and differing from typical geckos in having movable eyelids; of United States southwest and Florida Gulf Coast	Small lizard with dark bands across its body and a movable eyelid.
green iguana	large herbivorous tropical American arboreal lizards with a spiny crest along the back; used as human food in Central America and South America	Large, bright green lizard with a spiny crest along the back and long tail.
Carolina anole	small arboreal tropical American insectivorous lizards with the ability to change skin color	Small arboreal lizard with a long tail and color-changing skin.
desert grassland whiptail lizard	any of numerous very agile and alert New World lizards	Small lizard with a long tail, usually black and white or brown and white.
agama	small terrestrial lizard of warm regions of the Old World	Small lizards with long tails, long legs, and a long, pointed snout.
frilled-necked lizard	large arboreal insectivorous Australian lizard with a ruff of skin around the neck	Large arboreal insectivorous Australian lizard with a ruff of skin around the neck.
alligator lizard	slim short-limbed lizard having a distinctive fold on each side that permits expansion; of western North America	Slim, short-limbed lizard with a distinctive fold on each side permitting expansion; of western North America.
Gila monster	large orange and black lizard of southwestern United States; not dangerous unless molested	Large, orange and black lizard with a long tail and forked tongue.
European green lizard	a common Eurasian lizard about a foot long	Small reptile with a long tail, pointed snout, and row of spikes along its back.
chameleon	a chameleon found in Africa	Small lizard with a long tail, long neck, and long, thin body covered with many small, sharp scales.
Komodo dragon	the largest lizard in the world (10 feet); found on Indonesian islands	Large lizard with a thick, scaly body, long tail, and large head with sharp teeth.
Nile crocodile	a dangerous crocodile widely distributed in Africa	Large crocodile with a broad, flat snout, long tail, and long, pointed snout.
American alligator	large alligator of the southeastern United States	Large reptile with a long snout, broad head, and long tail.
triceratops	huge ceratopsian dinosaur having three horns and the neck heavily armored with a very solid frill	Large herbivorous dinosaur with three horns and a frill on its neck.

Continued on the next page...

Table 6 is continued from previous page

Class	WordNet Gloss	V-GLOSS
worm snake	small reddish wormlike snake of eastern United States	Small reddish wormlike snake of eastern United States.
ring-necked snake	any of numerous small nonvenomous North American snakes with a yellow or orange ring around the neck	Small nonvenomous snake with a yellow or orange ring around the neck.
eastern hog-nosed snake	harmless North American snake with upturned nose; may spread its head and neck or play dead when disturbed	Harmless North American snake with upturned nose; may spread its head and neck or play dead when disturbed.
smooth green snake	either of two North American chiefly insectivorous snakes that are green in color	Slender, smooth-scaled snake with a green or yellowish-green coloration.
kingsnake	any of numerous nonvenomous North American constrictors; feed on other snakes and small mammals	Large, nonvenomous snake with a pattern of alternating light and dark bands on its body.
garter snake	any of numerous nonvenomous longitudinally-striped viviparous North American and Central American snakes	Slender, smooth-scaled, nonvenomous snake with a long tail and a pattern of alternating light and dark bands.
water snake	any of various mostly harmless snakes that live in or near water	Slender, elongated, usually nonvenomous snake with a flattened head and a long tail often held above the head.
vine snake	slender arboreal snake found from southern Arizona to Bolivia	Slender arboreal snake found from southern Arizona to Bolivia.
night snake	nocturnal prowler of western United States and Mexico	Nocturnal colubrid snake that is black with a white belly and a white stripe down its back.
boa constrictor	very large boa of tropical America and West Indies	Large snake with a thick body and large head covered with scales, forked tongue.
African rock python	very large python of tropical and southern Africa	Large, heavy, non-venomous snake, grey or brown with a black head and yellowish belly.
Indian cobra	a cobra of tropical Africa and Asia	Large venomous snake with a hooded head and forked tongue.
green mamba	green phase of the black mamba	Large venomous snake with a green body and black head and tail.
sea snake	any of numerous venomous aquatic viviparous snakes having a fin-like tail; of warm littoral seas; feed on fish which they immobilize with quick-acting venom	Venomous snake living in the sea with a long, thin body, large head, and forked tongue.
Saharan horned viper	highly venomous viper of northern Africa and southwestern Asia having a horny spine above each eye	Venomous snake with a horny spine above each eye.
eastern diamondback rattlesnake	large deadly rattlesnake with diamond-shaped markings	Large, venomous snake with a diamond-shaped pattern on its back.
sidewinder rattlesnake	small pale-colored desert rattlesnake of southwestern United States; body moves in an s-shaped curve	Small pale-colored desert rattlesnake of southwestern United States; triangular head, long tail, and rattle on its tail.
trilobite	an extinct arthropod that was abundant in Paleozoic times; had an exoskeleton divided into three parts	Fossilized arthropod from the Paleozoic era with a hard exoskeleton divided into three parts.
harvestman	spiderlike arachnid with a small rounded body and very long thin legs	Spiderlike arachnid with a small rounded body and very long thin legs.
scorpion	arachnid of warm dry regions having a long segmented tail ending in a venomous stinger	Venomous arachnid with a segmented tail ending in a stinger.
yellow garden spider	a widely distributed North American garden spider	Large spider with a black body and a yellow abdomen.
barn spider	an orange and tan spider with darkly banded legs that spins an orb web daily	Large, hairy spider with a dark brown body and a white, orange and black pattern on its abdomen.
European garden spider	a spider common in European gardens	Small spider with a long, thin body and a large, round abdomen.
southern black widow	venomous New World spider; the female is black with an hourglass-shaped red mark on the underside of the abdomen	Spider with a black body and a red hourglass-shaped mark on the underside of the abdomen.
tarantula	large hairy tropical spider with fangs that can inflict painful but not highly venomous bites	Large hairy tropical spider with fangs that can inflict painful but not highly venomous bites.

Continued on the next page...

Table 6 is continued from previous page

Class	WordNet Gloss	V-GLOSS
wolf spider	ground spider that hunts its prey instead of using a web	Large, hairy spider with a long, thin body, large head, two large eyes, and a pair of fangs.
tick	any of two families of small parasitic arachnids with barbed proboscis; feed on blood of warm-blooded animals	Small parasitic arachnid that feeds on blood.
centipede	chiefly nocturnal predacious arthropod having a flattened body of 15 to 173 segments each with a pair of legs, the foremost pair being modified as prehensors	Small, segmented, wormlike arthropod with a pair of long, segmented legs and a pair of short, segmented antennae.
black grouse	grouse of which the male is bluish-black	Grouse of which the male is bluish-black.
ptarmigan	large Arctic and subarctic grouse with feathered feet and usually white winter plumage	Large grouse with a white head and neck, brown body, and white tail.
ruffed grouse	valued as a game bird in eastern United States and Canada	Medium-sized game bird with a black body, white breast, and a ruff of feathers around the neck.
prairie grouse	brown mottled North American grouse of western prairies	Large brown mottled North American grouse of western prairies.
peafowl	male peafowl; having a crested head and very large fanlike tail marked with iridescent eyes or spots	Large, colorful, iridescent bird with a fan-shaped tail and a crest on its head.
quail	small gallinaceous game birds	Small game bird with a plump body, short tail, long, pointed bill, and short, rounded tail.
partridge	small Old World gallinaceous game birds	Small bird with a brown body, white breast, and black head and neck.
african grey parrot	commonly domesticated grey parrot with red-and-black tail and white face; native to equatorial Africa	Medium-sized parrots with a grey body, red-and-black tail, and white face.
macaw	long-tailed brilliantly colored parrot of Central America and South America; among the largest and showiest of parrots	Large brightly colored parrot with a long tail and long beak.
sulphur-crested cockatoo	white cockatoo with a yellow erectile crest	Large white cockatoo with a yellow erectile crest.
lorikeet	any of various small lories	Small brightly colored parrot-like bird with a long tail and curved beak.
coucal	Old World ground-living cuckoo having a long dagger-like hind claw	Large bird with a long dagger-like hind claw.
bee eater	colorful chiefly tropical Old World bird having a strong graceful flight; feeds on especially bees	Colorful Old World bird with a strong graceful flight that feeds on bees.
hornbill	bird of tropical Africa and Asia having a very large bill surmounted by a bony protuberance; related to kingfishers	Large tropical bird with a large bill and long tail.
hummingbird	tiny American bird having brilliant iridescent plumage and long slender bills; wings are specialized for vibrating flight	Small bird with a long slender bill and iridescent feathers.
jacamar	tropical American insectivorous bird having a long sharp bill and iridescent green or bronze plumage	Small, colorful birds with long bills and iridescent feathers.
toucan	brilliantly colored arboreal fruit-eating bird of tropical America having a very large thin-walled beak	Large colorful bird with a long beak and crest on its head.
duck	adult male of a wild or domestic duck	Male duck.
red-breasted merganser	widely distributed merganser of America and Europe	Large duck with a red breast and black head and neck.
goose	web-footed long-necked typically gregarious migratory aquatic birds usually larger and less aquatic than ducks	Large bird with a long neck, short tail, usually white with black or brown markings.

A.2 Mis-mappings from Most Frequent Sense Heuristic

Class	Wrong Sense	Correct Sense
Beaver	the soft brown fur of the beaver	large semiaquatic rodent with webbed hind feet and a broad flat tail; construct complex dams and underwater lodges
Castle	a large and stately mansion	interchanging the positions of the king and a rook
Cloud	any collection of particles (e.g., smoke or dust) or gases that is visible	a visible mass of water or ice particles suspended at a considerable altitude
Flatfish	sweet lean whitish flesh of any of numerous thin-bodied fish; usually served as thin fillets	any of several families of fishes having flattened bodies that swim along the sea floor on one side of the body with both eyes on the upper side
Leopard	the pelt of a leopard	large feline of African and Asian forests usually having a tawny coat with black spots
Lobster	flesh of a lobster	any of several edible marine crustaceans of the families Homaridae and Nephropsidae and Palinuridae
Otter	the fur of an otter	freshwater carnivorous mammal having webbed and clawed feet and dark brown fur
Raccoon	the fur of the North American racoon	an omnivorous nocturnal mammal native to North America and Central America
Ray	a column of light (as from a beacon)	cartilaginous fishes having horizontally flattened bodies and enlarged winglike pectoral fins with gills on the underside; most swim by moving the pectoral fins
Seal	fastener consisting of a resinous composition that is plastic when warm; used for sealing documents and parcels and letters	any of numerous marine mammals that come on shore to breed; chiefly of cold regions
Shrew	a scolding nagging bad-tempered woman	small mouselike mammal with a long snout; related to moles
Skunk	a person who is deemed to be despicable or contemptible	American musteline mammal typically ejecting an intensely malodorous fluid when startled; in some classifications put in a separate subfamily Mephitinae
Table	a set of data arranged in rows and columns	a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs
Television	broadcasting visual images of stationary or moving objects; ; - Ernie Kovacs	an electronic device that receives television signals and displays them on a screen
Tiger	a fierce or audacious person	large feline of forests in most of Asia having a tawny coat with black stripes; endangered
Turtle	a sweater or jersey with a high close-fitting collar	any of various aquatic and land reptiles having a bony shell and flipper-like limbs for swimming

Table 7: CIFAR-100 classes where the most frequent sense heuristic failed

A.3 A More Detailed Comparison of Methods Over CLIP variants

Method	Model	Datasets				# LM Parameters
		ImageNet	CIFAR-100	CIFAR-10	STL-10	
Lex Baseline	ViT-B-32	55.7	60.5	87.4	96.3	0
	ViT-L-14	67.7	72.2	91.4	97.7	
	ViT-L-14-336	69.1	71.9	91.5	98.2	
	RN50	51.6	34.1	69.7	91.9	
	RN50x64	65.9	52.6	81.1	96.4	
1-Template Baseline	ViT-B-32	59.4	64.5	88.3	97.3	0
	ViT-L-14	71.1	77.3	95.2	99.5	
	ViT-L-14-336	72.4	76.6	94.8	99.5	
	RN50	55.6	42.1	70.3	94.4	
	RN50x64	68.7	57.7	81.0	98.4	
Template Ensemble	ViT-B-32	63.2	65.1	91.3	97.2	0
	ViT-L-14	75.3	77.9	96.2	99.3	
	ViT-L-14-336	76.2	77.5	95.7	99.4	
	RN50	59.6	41.6	75.6	94.3	
	RN50x64	73.2	61.3	86.8	98.3	
CuPL + Template Ensemble	ViT-B-32	64.6	-	-	-	175B
	ViT-L-14	76.6	-	-	-	
	ViT-L-14-336	77.6	-	-	-	
	RN50	61.3	-	-	-	
	RN50x64	75.1	-	-	-	
V-GLOSS *Normal-Only*	ViT-B-32	63.2	65.1	91.2	97.3	6.1B
	ViT-L-14	75.3	76.5	95.9	99.5	
	ViT-L-14-336	77.3	77.5	95.6	99.4	
	RN50	57.9	45.6	76.7	94.3	
	RN50x64	73.3	63.5	86.8	98.3	
V-GLOSS (*Normal* + *Contrastive*)	ViT-B-32	65.7	66.3	92.1	97.7	6.1B
	ViT-L-14	77.6	**78.2**	**97.0**	**99.6**	
	ViT-L-14-336	**78.5**	78.0	96.0	**99.6**	
	RN50	62.8	45.8	76.8	95.0	
	RN50x64	74.5	64.6	87.8	98.8	

Table 8: Top-1 accuracy on ZSIC across five CLIP variants.