Unconstrained Open Vocabulary Image Classification: Zero-Shot Transfer from Text to Image via CLIP Inversion

Philipp Allgeuer Kyra Ahrens Stefan Wermter
University of Hamburg
{philipp.allgeuer, kyra.ahrens, stefan.wermter}@uni-hamburg.de

Abstract

We introduce NOVIC, an innovative real-time uNconstrained Open Vocabulary Image Classifier that uses an autoregressive transformer to generatively output classification labels as language. Leveraging the extensive knowledge of CLIP models, NOVIC harnesses the embedding space to enable zero-shot transfer from pure text to images. Traditional CLIP models, despite their ability for open vocabulary classification, require an exhaustive prompt of potential class labels, restricting their application to images of known content or context. To address this, we propose an "object decoder" model that is trained on a large-scale 92M-target dataset of templated object noun sets and LLM-generated captions to always output the object noun in question. This effectively inverts the CLIP text encoder and allows textual object labels from essentially the entire English language to be generated directly from image-derived embedding vectors, without requiring any a priori knowledge of the potential content of an image, and without any label biases. The trained decoders are tested on a mix of manually and web-curated datasets, as well as standard image classification benchmarks, and achieve fine-grained prompt-free prediction scores of up to 87.5%, a strong result considering the model must work for any conceivable image and without any contextual clues.[1]

1. Introduction

Imagine a robot navigating through an environment, encountering a small cup, and an operator asking, *"Can you tell me what that is?"* Ideally, the robot would recognize the object with the objectives to: a) answer *cup* directly, and b) provide an answer immediately in real-time, even if there are large numbers of objects to consider at high frame rates. Contrary to the first objective, Vision-Language Models (VLM) like CLIP [53] require the robot to be prompted with a comprehensive textual list of object class candidates to perform zero-shot classification. On the other hand, equipping the robot with a high-performing multimodal Large Language Model (LLM) is computationally expensive, requires specialized remote servers for inference, and does not allow for the real-time responsiveness needed for video frame rates, thereby failing to meet the second objective. The core aim of this work is to develop models capable of unconstrained open vocabulary image classification. The models should generate *object nouns* directly as free-form text in real time, without the need for prompts or predefined label candidates.

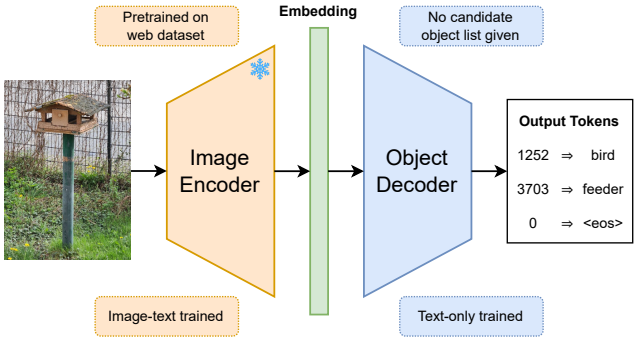

Figure 1. **Proposed open vocabulary image classifier.** The classifier processes each input image by encoding it into a CLIP embedding vector, which is then decoded into a sequence of tokens representing the object class. The object decoder is purely generative, producing free-form text without relying on a predefined list of candidate objects. This method is trained solely on textual data and is capable of zero-shot transfer to image data during inference.

Current approaches to open vocabulary learning [65] with VLMs for visual reasoning tasks (*e.g.* Tag2Text [29] and RAM [76] for image tagging) are in general limited by the need for candidate lists, which, with at most a few thousand entries, are not comprehensive enough to encompass all potential object classes. Although some parallel works propose strategies to utilize VLMs for generative tasks [22, 41], these are still ultimately restricted in diversity by the class candidates of labeled training datasets, leaving truly unconstrained prompt-free image classification unresolved. To address this gap in current research, we propose NOVIC, a novel *uNconstrained Open Vocabulary Image Classifier* (cf. Fig. 1). The

[1]The authors gratefully acknowledge support from the DFG (TRR 169 – CML) and European Commission (TRAIL).

classifier can directly label any given image with an English object noun without requiring any category prompts. Thanks to the use of synthetic large-scale text-only datasets, very efficient and scalable training of the so-called *object decoder* transformer is achieved by training purely on CLIP text embeddings. The decoder effectively learns to invert a CLIP text encoder, mapping embeddings back to object nouns via autoregressive token sequence generation. By leveraging the shared multimodal representations of pretrained VLMs, NOVIC can perform zero-shot image classification without ever having seen images during training.

To train the object decoder, we introduce an *object noun dictionary*, and use it to build several synthetically and LLM-generated caption-object datasets. These datasets are all created through automated pipelines, allowing for a high degree of scalability, especially compared to image-based training. Inspired by prior works [22, 50], noise augmentation is used on CLIP embeddings of the captions during training to bridge the large modality gap between image and text embeddings [44]. The resulting object decoder can be evaluated on classification datasets like ImageNet-1K [54], but this does not fully capture its prompt-free open vocabulary performance. To address this, we curate three open vocabulary image datasets from original and web sources, and annotate them using human and LLM knowledge.

In summary, our main contributions[2] are as follows: (i) we propose a novel open vocabulary object decoder model that is trained on text only and can zero-shot classify arbitrary images in real-time without any candidates or prompts, (ii) we develop an automated pipeline to construct a comprehensive English object noun dictionary and use a multi-set prompting scheme combined with an LLM to construct large-scale synthetic caption-object datasets, (iii) we curate three new image sets intended for evaluation of open vocabulary classification and provide human and multimodal LLM annotations, and (iv) we show that our method scales in performance with the underlying CLIP model, and is able to provide accurate yet very fine-grained classifications while achieving strong in-the-wild prediction scores up to 87.5%.

2. Related Work

Learning a joint embedding space for multiple modalities such as images and text via contrastive pretraining produces powerful representations without the need for expensive annotation of datasets [19, 39]. VLMs such as CLIP [53] and ALIGN [30] paved the way for learning a broad range of visual concepts that can be transferred to downstream discriminative tasks, such as classification, via prompting [21]. Given an image, a set of candidate categories, and prompt templates like *"a photo of a [...]"*, CLIP can zero-shot classify the image by maximizing the similarity score between its encoded representation and all encoded candidate prompts. NOVIC, by contrast, frames the image classification problem as an unconstrained generative one, eliminating the need for category candidates and inference-time prompting. Various approaches build upon and extend the idea behind CLIP towards better pretraining [42, 59, 74] or fine-tuning [63] techniques, or add additional modalities such as audio [24, 64], video [1, 28, 46, 69, 72], or point clouds [71, 75]. Other approaches use VLMs like CLIP as feature extractors for subsequent fine-tuning on more challenging downstream tasks such as visual question-answering [14, 56, 58], captioning [6, 48], motion planning [57, 60], and navigation [33].

Open vocabulary learning is a paradigm that emerged in the context of fine-tuning pretrained VLMs for vision tasks. It aims to overcome the closed-set assumption of standard classification models by allowing them to handle novel object categories outside of their training domain. Models such as VILD [23], RegionCLIP [77], and various successor models [11, 13, 32, 51, 66, 67] enable VLMs to solve tasks such as open vocabulary object detection [73], where, given an arbitrary set of candidate object labels, the model has to detect the corresponding objects in the image. Other works extend the open vocabulary paradigm to closely related visual tasks such as semantic segmentation [25, 34, 37, 43, 45, 70] and image tagging [29, 40, 76]. Despite their merit, all the approaches mentioned above rely on candidate labels for inference, a strong assumption that we abandon in this work.

A parallel line of research aims to leverage the powerful representations provided by VLMs for generative tasks such as image captioning. As VLMs like CLIP cannot inherently produce sequences, approaches in this research area rely on additional decoders. Several methods have been proposed that either establish a mapping between a VLM and a pretrained text decoder [17, 48, 61], or train a decoder from scratch [6, 12, 31, 56] using curated captioning datasets such as MS COCO [10], CC3M [55], and Visual Genome [35].

A recent line of research on zero-shot image captioning or retrieval with VLMs proposes text-only training to eliminate the need for curating expensive image datasets. As text and images are encoded into disjoint regions of the VLM embedding space, additional mechanisms to bridge the modality gap have been proposed, like injecting noise into the text embeddings [20, 22, 50] or projecting the image embeddings [41]. These methods face two significant disadvantages—first, captions often fail to capture key image aspects or identify core objects, and second, they are biased towards the limited set of object classes contained in their training dataset. NOVIC adopts text-only training, but focuses on predicting truly open vocabulary object labels.

3. Methodology

An overview of the proposed training and inference schemes for NOVIC is shown in Fig. 2. A text-only dataset

[2] All code [2, 3] (GitHub) and datasets [4, 5] are publicly available.

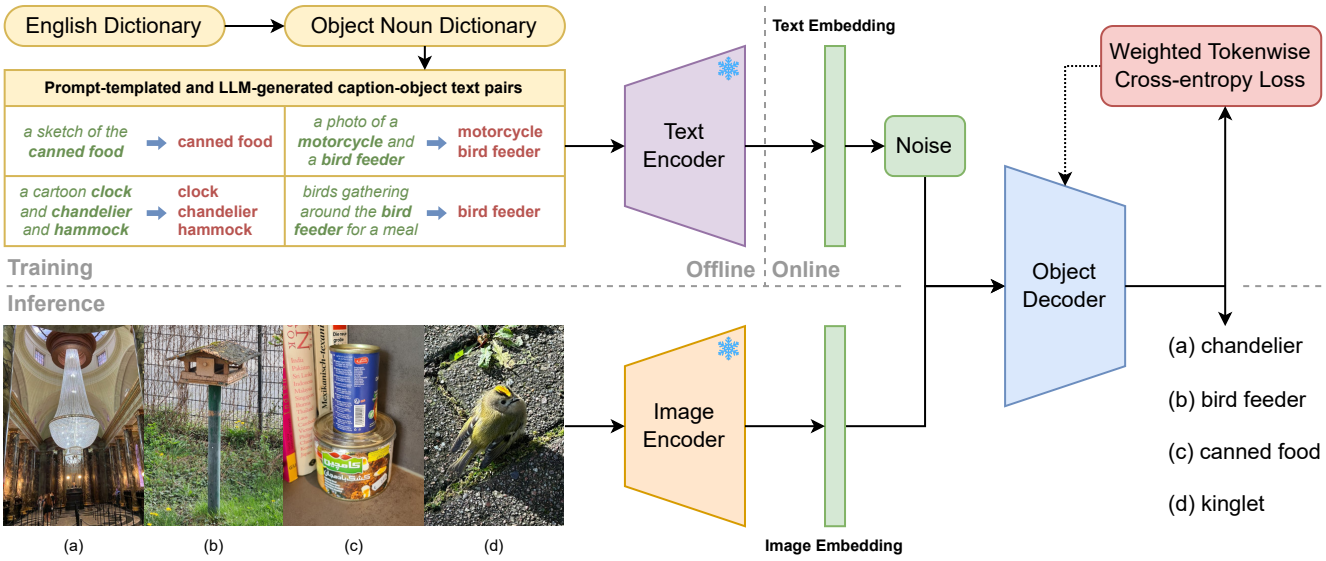

Figure 2. **Overview of the NOVIC training and inference schemes.** A dataset of caption-object text pairs is generated offline using an English dictionary, prompt templates, and LLM-based caption generation. The captions are encoded offline into text embeddings, and augmented with noise to train the object decoder. The training scheme seamlessly generalizes across the large modality gap between image and text embeddings, allowing inference of arbitrary images via the image encoder. Produced classifications can be very fine-grained.

mapping caption sentences to target object nouns is synthetically generated offline (cf. Sec. 3.2). The text embedding vectors of the captions are then precomputed using the text encoder of a frozen CLIP model. These vectors are augmented online with unit-norm-preserving noise, and used to train a decoder-only transformer, the *object decoder* (cf. Sec. 3.1), to generate the object noun corresponding to each caption. The output object nouns are generated in the form of free language using the same text tokenizer as the frozen text encoder, with no definition of candidate object nouns being provided to the decoder in any way. At inference time, the object decoder seamlessly generalizes to image embedding vectors calculated using the image encoder of the frozen CLIP model, despite CLIP models generally having very large modality gaps [44]. The performance of the decoder is measured on image classification datasets and annotated open vocabulary image datasets (cf. Sec. 3.3).

3.1. Object Decoder Model

The architecture of the object decoder model is shown in Fig. 3. The model is built around an autoregressive decoder-only language transformer [62] that has been configured to use learned token embeddings, weight tying [52], learned positional embeddings, pre-layer normalization [68], GELU activation functions [26], beam search, custom weight initialization (cf. App. A), a contractive inner feed-forward dimension, and no biases in the linear and layer normalization layers, including the final token logits linear layer. Dropout is applied after the positional embedding and within the transformer layers. The input embedding vector is supplied to the transformer in the form of P prefix token vectors, *i.e.* P dynamically calculated vectors of dimension H (hidden dimension). The prefix tokens are computed from the input embedding using a learned linear projection that has PH outputs and no biases, as the immediately following positional embedding makes biases redundant. The tokenized target object nouns do not require a start token, as the first output token is predicted at the sequence location of the last prefix token. Appropriate sequence masking is also used to ensure that only sequence locations up to the first output `<eos>` contribute to the loss. The internal attention mask of the transformer is causal, with the exception that the P prefix tokens can attend to each other without restriction, allowing them to share information freely.

3.2. Training Dataset for Text-Only Supervision

The object decoder needs to be trained on a dataset of associative caption-object text pairs (cf. Fig. 2). One could attempt to generate such a dataset from a large-scale image-text web dataset ranging into the billions of samples, but it is not a simple task to accurately and sufficiently efficiently construct the required ground truth object noun annotations, especially in such a way that a fair coverage of the English language is achieved. We instead *invert* the chosen CLIP text encoder by synthetically generating captions that explicitly cover the entire English language of object nouns, and use this to systematically sample in the CLIP embedding space essentially all possible object concepts, and learn to map them to their corresponding textual representations.

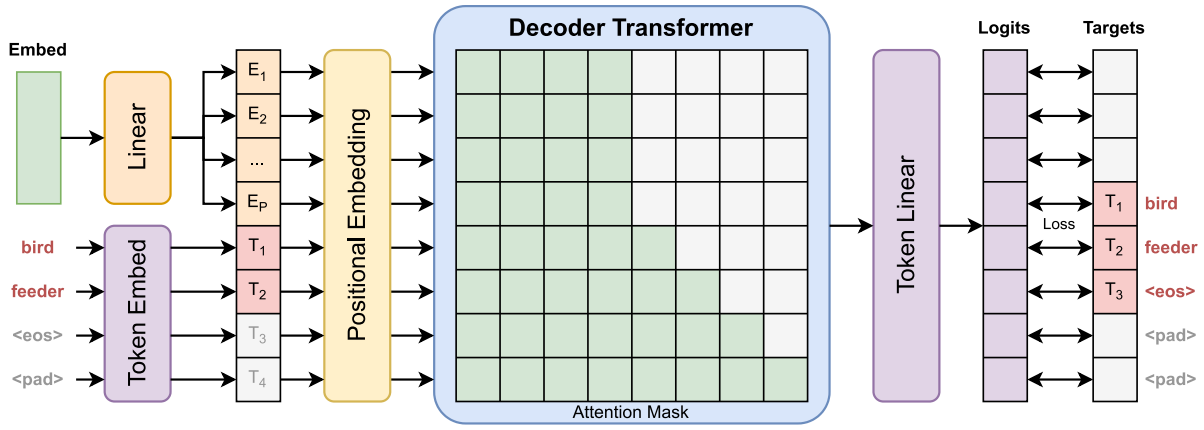

Figure 3. **Architecture of the object decoder.** The decoder-only transformer is given a sequence of linearly projected embedding vector tokens (full attention) followed by autoregressive object noun tokens (causal attention). The output sequence is converted to token logits using a linear layer that is weight-tied to the token embeddings. Cross-entropy loss with left-shifted target tokens is used during training.

Object noun dictionary. An exhaustive dictionary of simple and compound object nouns, referred to as the *object noun dictionary*, was created based on the WordNet dictionary [47], GNU Collaborative International Dictionary of English [9], and the categories used by various image datasets (cf. App. B). Noun entries were cleaned, collected and merged as appropriate to account for word duplication, plurality, inflection and alternate spellings, and filtered down to only include nouns that correspond to suitable object classifications (*e.g.* no abstract nouns). This resulted in 42 919 unique case-normalized object nouns corresponding to 96 526 noun variants. Each object noun was then assigned a log-scale frequency value based on its observed frequency in the English language as per the n-gram frequencies in the Google Web Trillion Word Corpus [8, 49]. Object nouns that are categories in a curated, comprehensive collection of image datasets were given appropriate frequency boosts. The log-scale frequencies were scaled and rounded so that very rare nouns saturate at a value of 1, and very frequent nouns have values $\geq$12. Aside from being used to adjust relative sampling frequencies during training, the resulting frequencies allowed the construction of object noun dictionary subsets by filtering out nouns less or equal to a particular frequency threshold (FT), *e.g.* FT2.[3]

Prompt templates. A direct and extremely cheap way of generating caption-object text pairs is by using exactly the same prompt templates as is usually used for zero-shot CLIP classification. This maps captions like *"a blurry photo of a bottle"* to the object noun *"bottle"*. Adjusting the ImageNet-1K and CIFAR prompts used by [53] to account for pluralization and allow for correct indefinite article use leads to 85 unique prompt templates for singular forms of object nouns, and 68 for plural forms. Combined with the full object noun dictionary, and including repetitions based on the log-scale frequencies of the nouns leads to 13.4M samples in the dataset. The mean observed object noun frequency is 3.5, but the true mean number of repetitions is much smaller at only 1.5, due to the use of the available alternate spellings during prompt templating.

Multiset prediction. Image embeddings can capture significantly more complicated dynamics than just the main entity present. If there is a photo of a lion with a zookeeper, for instance, the image embedding will likely contain some features corresponding to *lion*, *person*, *fence*, and potentially more abstract concepts like *zoo* as well. To help also capture these dynamics in the text embeddings seen during training, multi-target samples are introduced. The complete object noun dictionary, including explicit repetition of alternate spellings based on their individual log-scale frequencies of mean 1.5, is randomly shuffled m times into m separate lists. These lists are 'zipped' to form m-tuples of object nouns that are then used to generate prompt-templated captions exactly as before. The m object nouns per caption are linked by 'and' (cf. Fig. 2), and all m nouns use a target weight of $\frac{1}{m}$. This produces a dataset of 13.4M further samples for each value of m, referred to as the $\mathcal{M}_m$ *multiset* in each case (*e.g.* $\mathcal{M}_3$ for $m = 3$). In total, a combined dataset of $\mathcal{M}_1$[4] and the two multisets $\mathcal{M}_2$ and $\mathcal{M}_3$ contains 40.2M samples with 80.4M targets. We qualitatively observed that multiset training specifically helped the open vocabulary classification of images that prominently contain multiple different elements or objects. On an intuitive level, the use of multi-target samples also prevents the object decoder from being able to just 'rote-memorize' which local regions of the text embedding space correspond to which pure object con-

[3] FT0, FT2, FT6, and FT9 have 42 919, 11 897, 5 899, and 2 919 unique object nouns, respectively. FT0 is the complete object noun dictionary.

[4] $\mathcal{M}_1$ is equivalent to the prompt-templated dataset as initially described.

cepts, and instead forces the decoder to learn *which* values in *which* parts of the embedding vector are actually associated with *which* object concept. Such knowledge is essential for proper generalization to image embeddings.

Caption data. In order to supply the object decoder with even more diverse synthetic training data to generalize from, captions were generated for each noun in the object noun dictionary using an LLM (cf. App. C). This approach leverages the implicit contextual knowledge of LLMs to place object nouns in *relevant* contexts—incorporating appropriate adjectives, actions, scenes, and secondary objects—in a style resembling real image captions.[5] The minimum number of unique generated captions for each noun variant was set proportionally to the corresponding log-scale frequency (up to 100) in order to match the relative word frequencies found in the English language. OpenAI GPT-3.5 Turbo was selected as an ideal balance of cost and sufficient performance, and was used to generate a total of 1.8M captions. Ensuring frequency-appropriate use of alternate spellings, this resulted in a total of 2.9M caption samples (2.0M unique) that were then oversampled to around 30% the size of any multiset they were merged with. For the combined dataset $\mathcal{M}_{1-3}$ this resulted in 51.7M samples with 91.9M targets.

Noise augmentation. In order to allow more robust and generalizable features of the CLIP embedding space to be learned, data augmentation in the form of noise is used during training. Due to the large modality gap between text and image embeddings [44], noise is in fact a crucial component of the training scheme. Noise allows the object decoder to learn to be invariant to even very large changes in the non-object-related parts of the text embedding vectors, which helps the generalization to image embeddings, as these are very sensitive to secondary objects, backgrounds, and even subtle visual variations. Gu et al. [22] used elementwise Gaussian noise with vector renormalization to train a captioning model on text embeddings only, but did not entirely accurately explain its effect. The CLIP embedding space is deceptively and unimaginably vast[6], so the random noise statistically *cannot* be significantly filling the gaps between samples, or causing a meaningful overlap with image embedding vectors. The main benefit of noise is that the model can learn to focus on which parts (*i.e*. subspaces) of the embedding vector really encode which individual object concepts, instead of focusing on which region of the embedding space a vector is in as a whole. We observe that the amount of noise required for best training results is actually quite large, not small or minor [22], and its effect is to completely scatter the cone [44] of text embeddings to a thin hyperannulus on the order of 66–80° away, with no single sample remaining significantly closer than that to any original text embedding (cf. App. D). By comparison, the very large modality gap between the image and text embedding cones can also be measured, and is around 70–85° for *matching* image-text pairs (depending on the CLIP model), and only 8–12° more than this for completely uncorrelated image-text pairs. As indicated previously though, statistically there is negligible region overlap between the scattered hyperannulus and the image embeddings cone due to the vast nature of the embedding space. The large amount of noise required can instead be seen as a trade-off between the disruptive 'blurring' effect of noise, and the benefit of the object decoder having seen many embedding vectors during training that are as far removed from text embeddings as image embeddings are. In this work, we use an 85% to 15% mix of elementwise Gaussian noise and uniform angle noise. The latter was developed to beneficially cover a greater range of noise dynamics and levels, by rotating input text embedding vectors uniformly in random directions by angles in the uniform 45–75° range.

3.3. Evaluation Datasets for Zero-Shot Recognition

One way to evaluate how much NOVIC has learned is by examining its zero-shot performance on standard image classification datasets like ImageNet-1K and Food-101. Context is everything for image classification datasets however. Knowing which categories are available in a dataset serves as a *strong* prior for supervised models. For instance, if a dataset contains only one bird category, such as *kinglet*, a supervised model only needs to determine whether an image is a bird to confidently classify it as a kinglet. For prompt-free open vocabulary models like NOVIC however, the task is more challenging. When presented with an image that happens to come from Food-101, it is not clear that the model should suddenly focus on the specific dish shown instead of potentially more prominent or general objects in the image. Consequently, to meaningfully evaluate NOVIC on image classification datasets, we restrict the generated beam search text tokens in every forward pass of the object decoder to those consistent with the defined category names, and score whether the correct final category is identified.

As NOVIC is trained without providing explicit categories, we wish to evaluate it also on datasets that do not specify categories. No suitable such dataset is known however, as such a dataset would need to have annotations for every possible correct object noun in the English language for every visible entity in every part of each image. To address this, three open vocabulary image datasets were curated as part of this work (cf. App. E), and individually annotated by both human and multimodal LLM annotators (cf. App. F) for the object nouns that were predicted by the trained models. The annotations specify whether each clas-

[5] CLIP embedding spaces are *trained* on real image captions.

[6] Even if we assumed each element of a 768-dim embedding vector only contains a *single* bit of information, this would still lead to 10^{231} different embeddings, against which even a trillion samples is essentially zero.

sification is *correct*, *close*, or *incorrect*, and for the human annotations, whether it relates to a *primary* or *secondary* element of the image. The three new datasets are (i) World: 272 images of which the grand majority are originally sourced (have never been on the internet) from 10 countries by 12 people, with an active focus on covering as wide and varied concepts as possible, including unusual, deceptive and/or indirect representations of objects, (ii) Wiki: 1 000 Wikipedia lead images sampled from a scraped pool of 18K, (iii) Val3K: 3 000 images from the ImageNet-1K validation set, sampled uniformly across the classes (LLM annotations only). The suffixes -H and -L are used to clarify whether human or LLM annotations are in use, *e.g.* Wiki-H. A total of 17.4K human and 112K LLM (OpenAI GPT-4o) annotations were made across the three datasets. This was at the limit of feasibility in terms of number of samples, person-hours and cost.

4. Experiments

For experiments, the SigLIP ViT B/16 model [74] is used as the base CLIP model, due to its favorable balance of speed and performance, especially when compared to the original CLIP models [53]. Training is performed for 18 epochs with the AdamW optimizer, cosine learning rate decay, weight decay on all multidimensional weights, and a gradient-accumulated batch size of $B = 8192$. Hyperparameter details can be found in App. G. The object decoder model is quite small at only 12.2M parameters, and takes less than 3 days to train on a *single* RTX A6000. For FT2, this drops to 11.4M parameters and only 1.5 days. Overall, this is *extremely* efficient, considering that up to a billion samples are being seen by a single GPU. As a scale comparison, using our method on the very same GPU to train an object decoder on all 1.28M ImageNet-1K samples for 18 epochs takes a mere 19 minutes, and results in an impressive top-1 accuracy of 88.21%. The object decoder is also very resource efficient at inference time. Requiring at most 5 GB GPU memory, the total pipeline—including CLIP model, autoregression, and beam search—takes on average 26 ms per image for single images, and 7 ms per image when using batching ($B = 256$), making it very real-time capable.

Open vocabulary scoring. Model predictions on the three introduced open vocabulary image datasets are scored by assigning $\{1, 0.8, 0.5, 0.4, 0\}$ points for {*correct primary*, *correct secondary*, *close primary*, *close secondary*, *incorrect*} predictions respectively, and expressing the sum as a percent of the maximum possible score. For verification purposes, the prediction scores achieved by many models were statistically compared between human and LLM annotations on the same dataset and found to be suitably correlated (cf. App. F), thus allowing model performance insights to be gained even when manual annotation is just too prohibitive. It is expected that LLM prediction scores are lower than human ones due to some residual LLM mistakes, a generally narrow acceptance of classifications, and inherently less frequent use of the *close* annotation by the LLM.

Dataset	CLIP	FT9	FT6	FT2	FT0
ImageNet-1K [54]	*75.91*	46.76†	**69.50**	68.11	60.56
Tiny ImageNet [38]	*59.77*	54.53†	**57.36**	55.83	53.78
Imagenette [16]	*99.59*	**99.44**	99.27	99.32	99.23
Imagewoof [16]	*93.28*	**93.07**	92.07	92.20	90.15
ImageNet-A [27]	*45.05*	35.33†	**42.43**	41.42	38.23
ImageNet-R [27]	*90.24*	74.46†	**88.01**	86.62	83.43
Food-101 [7]	*91.55*	**85.25**	83.24	82.73	68.80
CIFAR-10 [36]	*92.33*	**91.08**	90.53	90.77	90.92
CIFAR-100 [36]	*72.19*	**70.65**	69.93	69.43	67.80

Table 1. **Performance on image classification benchmarks (% mean of 3).** †FT9 has disadvantaged scores on some datasets due to the seen object nouns not covering all the classes (cf. App. J).

Annotations	FT9	FT6	FT2	FT0
ImageNet-1K	46.52	69.47	67.92	60.72
Open (correct)	**70.33**	**75.04**	**72.90**	**66.19**
Open (correct + close)	*71.90*	*76.50*	*74.35*	*67.86*

Table 2. **Performance on Val3K dataset (% mean of 3),** showing that category-constrained scores are a significant underestimation of the true open vocabulary performance of NOVIC.

4.1. Zero-Shot Classification Performance

Image classification benchmarks. Tab. 1 shows the zero-shot performance of NOVIC on various image classification benchmarks. The performance of the underlying CLIP model is also shown as a measure of how completely the vast knowledge of the CLIP model has been captured by the object decoder and converted into generative textual understanding. The decoder relies on the organization of the embedding space provided by the CLIP model, and so naturally is not expected to outperform it in this category-prompted setting. Overall, NOVIC shows competitive top-1 accuracies, with some 'distillation' losses being observed for tendentially finer-grained classification datasets and lower frequency thresholds. Classification benchmarks, however, do not tell the whole story. The act of restricting the object decoder to a fixed vocabulary *forces* it to choose object nouns it does not even consider to be best. In Tab. 2, we evaluate the top-1 accuracy on Val3K using the original ImageNet-1K annotations, and compare this to the percent of open vocabulary classifications on the very same images that are *correct* as per Val3K-L. Prediction scores that include points for *close* predictions are also shown. The results demonstrate that the object decoder's knowledge is more accurately assessed when used for unconstrained open vocabulary classification.

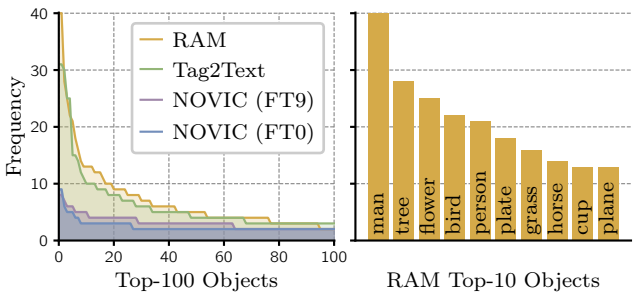

Figure 4. **Diversity comparison of predicted object nouns.** Plots showing the sorted top frequency counts of the object nouns predicted for Wiki-H. Even when trained on FT9, NOVIC shows much greater diversity and has a less peaked frequency distribution. **Right:** The top-10 nouns for RAM are generic and overused.

Model	Seen Objects	Used Objects	RAM Vocab	Primary Score	Prediction Score	Overall Score
Tag2Text [29]	3 429	397	89.0	69.15	**81.35**	62.61
RAM [76]	4 585	348	100.0	65.05	80.73	59.98
RAM (FT2)	11 897	391	83.4	62.30	77.62	60.58
RAM (FT0)	**42 919**	399	81.7	61.70	76.94	60.25
NOVIC (FT9)	2 919	693	64.0	68.55	72.35	69.08
NOVIC (FT6)	5 899	794	49.3	69.90	72.70	71.07
NOVIC (FT2)	11 897	846	39.7	68.65	72.53	71.88
NOVIC (FT0)	**42 919**	852	37.5	68.70	72.34	71.31
NOVIC† (FT2)	11 897	893	36.3	77.05	80.13	**79.02**
NOVIC† (FT0)	**42 919**	**934**	**28.7**	**77.10**	79.18	78.21

Table 3. **Open vocabulary Wiki-H results in comparison to baselines (best of 3).** Tag2Text and RAM achieve artificially high scores by frequently outputting uninformative generic tags like *man*, *tree*, *grass*, *plate*, and not for instance what is *on* the grass or plate, or what *type* of tree it is, or what the man is *using*. Scores are in %. *RAM Vocab* is the percent of samples whose predictions are in the core RAM vocabulary. *Primary Score* is the *Prediction Score* when rewarding only primary predictions. *Overall Score* is the *Prediction Score* with $\times 0.5$ applied to the scores of coarse predictions. See App. K. †Using DFN-5B H/14-378 CLIP model.

Comparison with related work. As discussed in Sec. 2, no directly equivalent state-of-the-art model is known that performs real-time prompt-free open vocabulary image classification. The most suitable models for comparison are image tagging models like Tag2Text [29] and the Recognize Anything Model (RAM) [76], as they are designed to be able to provide text labels for images of arbitrary content. We evaluate the largest model variants by taking the highest scoring prediction for each image, and compare this to NOVIC in Tab. 3 and App. K. While the tagging models at first glance numerically seem to obtain a reasonable *Prediction Score*, the comparatively low *Primary Scores* indicate that the models, unlike NOVIC, make a large portion of their score from *secondary* or *close* predictions, as opposed to *correct primary* ones (a deficiency clearly illustrated in Tab. 7). Also, when accounting for prediction specificity, the resulting *Overall Score* clearly demonstrates the superiority of NOVIC by up to 19% absolute. A closer look reveals that the low scores of the tagging models are due to a severe lack of diversity and specificity (cf. Fig. 4 and qualitative results in App. H). The produced tags are tendentially coarse and repetitive, with around a third of all images receiving one of only 20 generic top labels, often missing the main point of images. Even when provided with the full object noun dictionary, the total diversity of labels stays below 400, with over 80% of predictions still being made from the small core RAM vocabulary. This is in contrast to our model, which has a much flatter prediction distribution, diversities up to 934 (cf. Tab. 3), and a remarkably high prediction score despite generating *fine-grained* classifications. The tags produced by Tag2Text and RAM also tend to be uninformative—when applied to the images of the Imagewoof [16] validation set for example, both models simply output *dog* 88% of the time, and never a single dog breed. Our FT0 model, on the other hand, predicts over 90 different dog breeds, with much of the weight covering variations of the actual ground truth classes. When RAM is evaluated on ImageNet-1K by only providing it the 1000 class names to choose from, it only scores 38.3% (25.4% for non-core vocabulary), which is *far* less than the 75.54% of the ViT-L CLIP model it was trained on, and *far* less than the 60–70% nominally scored by NOVIC.

4.2. Ablation Studies

Tab. 4 shows the effect on decoder performance of varying the dataset configuration, training noise, and underlying CLIP model. FT2 was chosen as the ablation baseline, as many ablation runs were required, and FT2 strikes a balance between overall feasibility of training times, and ensuring that a diverse and expressive range of object nouns are seen during training. Further ablations are presented in App. I. Overall, as expected, the human annotations exhibit generally higher scores and are slightly more consistent, due to human checking and more prevalent use of the *close* category. The decoder ImageNet-1K scores are also observed to be well-correlated with general open prediction performance.

Training dataset. The effect of the object noun dictionary frequency threshold (cf. Sec. 3.2) on decoder classification performance is shown in Tab. 4. The prediction scores grow as the training dataset size is increased, until a limit is reached between FT2 and FT0 where the exploding diversity of seen object nouns saturates the amount of order the CLIP model can provide in the embedding space. In Tab. 4, we systematically analyze the contribution of each component of the training set. Starting with $\mathcal{M}_1$ and incrementally building the nominal training set (*i.e.* $\mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3 +$ LLM captions), we observe consistent gains. Immediate losses are seen if the LLM captions are removed, or if the lower-order $\mathcal{M}_1$ and $\mathcal{M}_2$ multisets are removed. Further tests with

Configuration	World-H	Wiki-H	Wiki-L	IN1K
FT9	76.13	71.20	63.57	46.76
FT6	78.42	**72.33**	62.62	**69.50**
FT2*	**78.92**	72.03	**63.97**	68.11
FT0	73.98	71.48	61.27	60.56
$\mathcal{M}_1$ dataset	70.11	-	50.05	63.88
+ LLM captions	72.16	-	55.37	66.06
+ $\mathcal{M}_2$ dataset	78.37	-	61.60	67.97
+ $\mathcal{M}_3$ dataset*	**78.92**	**72.03**	**63.97**	**68.11**
$\mathcal{M}_1$ + $\mathcal{M}_2$ + $\mathcal{M}_3$	77.56	71.60	62.08	65.36
$\mathcal{M}_3$ + LLM captions	77.88	71.28	62.93	61.12
Gauss 1.00	19.44	-	15.62	13.36
Gauss 1.75	70.37	-	52.07	49.23
Gauss 2.50	76.26	-	62.10	67.94
Gauss 3.25	77.57	70.95	62.38	66.46
+ 15% Uniform 45–75*	**78.92**	**72.03**	**63.97**	**68.11**
Uniform 45–75	76.46	-	63.27	64.75
DataComp-1B B/16 [18]	76.85	-	63.55	64.17
SigLIP B/16* [74]	78.92	72.03	63.97	68.11
DataComp-1B L/14 [18]	84.27	-	70.35	70.05
SigLIP SO/14 [74]	**87.49**	-	70.28	74.85
DFN-5B H/14-378 [15]	87.08	-	**74.88**	**75.82**
→ FT0	86.41	-	74.48	75.68

Table 4. **Dataset, training noise, and CLIP model ablations (% mean of 3).** IN1K is the ImageNet-1K classification performance of the trained decoders, and the asterisks correspond to the nominal ablation baseline. See App. I for more ablation results. Some Wiki-H scores are unavailable due to the high cost of manual annotation.

variations in training noise configuration also validate that very large amounts of Gaussian noise are *required* in order to produce models that can generalize across the modality gap,[7] and that instead using uniform angle noise on 15% of the samples produces the best seen configuration.

CLIP model. The proposed NOVIC training scheme scales well with the performance of the underlying CLIP model. Tab. 4 demonstrates that progressively stronger CLIP models yield progressively stronger object decoders. Remarkably however, the training times remain almost unchanged, as text encoders only scale slowly with CLIP model strength and the datasets of text embedding vectors are generated offline and only once anyway (cf. Fig. 2). It can be observed from Tab. 4 that the performance of the FT0 object decoder that was trained on the strong DFN-5B CLIP model is only marginally below the corresponding FT2 performance. This is much closer than for the nominal CLIP model, and can be attributed to the improved order and separation of concepts that the CLIP model provides in the embedding space. Based on the World-H results, it is actually expected that the DFN-5B FT0 model can provide very fine-grained *correct primary* classifications for up to 84.2% of pictures it is given. Having trained on the full unfiltered object noun dictionary, the model is frequently able to make correct classifications as detailed as *red velvet cake*, *brazil nut tree*, *electron microscope*, and *egretta caerulea*.

Language generalization. During training, the object decoder develops an implicit understanding of language and its relationship to object concepts, demonstrating qualitative evidence that it can occasionally use this knowledge to generalize beyond the dataset. For instance, the unseen noun *New Zealand willow* was predicted for a scenic photo of a willow tree that was indeed taken in New Zealand. The object decoder also inherits CLIP's slight tendency to be able to 'read', as evidenced for example by a generic blue plushie being correctly classified by multiple decoders as *genus staphylococcus*, but not anymore if the small visible text *Staphylokokke* is removed, falling back to *stuffed toy*. In another case, a sign reading *Aral* is classified as *ara* (a macaw), indicating that the decoder can associate certain features in the embedding space with certain spellings. As another form of language generalization, the object decoder can easily be trained to output high-performing predictions in any chosen language—even low resource languages for which insufficient image-text data exists to train a CLIP model. By simply applying dictionary or machine translation to the target object nouns in the $\mathcal{M}_m$ multisets, a predominantly English CLIP model can be used to directly train a strong object decoder in any translatable language.

5. Conclusion

In this paper, we address the problem of unconstrained real-time open vocabulary image classification, and highlight the limitations of current state-of-the-art methods in accomplishing this task. To overcome these limitations, we propose an autoregressive object decoder trained on a large-scale text-only dataset. The decoder generates object nouns directly as language on the basis of CLIP embeddings, and crucially, does not require any class candidates to be provided. To create the text-only training dataset, we introduce a comprehensive object noun dictionary and employ multi-set prompt-templating in combination with LLM-generated captions and a pivotal noise augmentation strategy to effectively bridge the modality gap from text to images. Our approach demonstrates promising results on both established and newly proposed in-the-wild classification datasets. We lift the common limitation of zero-shot image recognition models regarding prompts and class candidates, ultimately allowing agents to immediately recognize and label *any* object in dynamic environments. Future work can include providing an input to the decoder that controls the level of semantic granularity used in the predicted output nouns.

[7] *Gauss 3.25* in the table refers to elementwise Gaussian noise with a standard deviation so high that the expected norm of the noise vector that is added to the unit-norm embedding is 3.25, *i.e.* many times its magnitude.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022. 2

[2] Philipp Allgeuer. Object Noun Dictionary generation source code. https://github.com/pallgeuer/object_noun_dictionary. 2

[3] Philipp Allgeuer and Kyra Ahrens. NOVIC: Unconstrained open vocabulary image classification. https://github.com/pallgeuer/novic and https://paperswithcode.com/paper/unconstrained-open-vocabulary-image. 2

[4] Philipp Allgeuer and Kyra Ahrens. NOVIC: Unconstrained open vocabulary image classification. https://www.inf.uni-hamburg.de/en/inst/ab/wtm/research/corpora.html#novic and https://paperswithcode.com/dataset/novic-caption-object-data. 2

[5] Philipp Allgeuer and Kyra Ahrens. Open vocabulary image classification datasets. https://www.inf.uni-hamburg.de/en/inst/ab/wtm/research/corpora.html#ovic-datasets and https://paperswithcode.com/dataset/ovic-datasets. 2

[6] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4662–4670, 2022. 2

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 6

[8] Thorsten Brants and Alex Franz. Google web trillion word corpus: Web 1T 5-gram version 1. https://catalog.ldc.upenn.edu/LDC2006T13. 4, 15

[9] Patrick Cassidy et al. GNU Collaborative International Dictionary of English. https://gcide.gnu.org.ua. 4, 15

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server, 2015. arXiv:1504.00325. 2

[11] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024. 2

[12] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets. *International Journal of Computer Vision*, 2023. 2

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning To Prompt for Open-Vocabulary Object Detection With Vision-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14084–14093, 2022. 2

[14] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. 2

[15] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *12th International Conference on Learning Representations*, 2024. 8

[16] fast.ai. Imagenette GitHub. https://github.com/fastai/imagenette. 6, 7

[17] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable Decoding with Visual Entities for Zero-Shot Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3136–3146, 2023. 2

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, et al. DataComp: In search of the next generation of multimodal datasets. In *37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 8

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 2

[20] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only Training of Zero-shot Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13225–13234, 2024. 2

[21] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models, 2023. arXiv:2307.12980. 2

[22] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I Can't Believe There's No Images! Learning Visual Tasks Using only Language Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2672–2683, 2023. 1, 2, 5

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *International Conference on Learning Representations*, 2022. 2

[24] Andrey Guzhov, Federico Raue, Jorn Hees, and Andreas Dengel. Audioclip: Extending Clip to Image, Text and Audio. In *ICASSP 2022 - 2022 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pages 976–980, Singapore, Singapore, 2022. IEEE. 2

[25] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. CLIP-S4: Language-Guided Self-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11207–11216, 2023. 2

[26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[27] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[28] Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a Unified Video-Language Alignment and Fusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14856–14866, 2023. 2

[29] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2Text: Guiding Vision-Language Model via Image Tagging. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 7, 23

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 2

[31] Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. Noise-Aware Learning from Web-Crawled Image-Text Data for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2942–2952, 2023. 2

[32] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, Verify, Correct: A Simple Few Shot Object Detection Method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14237–14247, 2022. 2

[33] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but Effective: CLIP Embeddings for Embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14829–14838, 2022. 2

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2

[36] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, University of Toronto, 2009. 6

[37] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 2

[38] Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. *CS 231N, Stanford University*, 2015. 6

[39] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal Foundation Models: From Specialists to General-Purpose Assistants, 2023. arXiv:2309.10020. 2

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 2

[41] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[42] Xianhang Li, Zeyu Wang, and Cihang Xie. CLIPA-v2: Scaling CLIP Training with 81.1% Zero-shot ImageNet Accuracy within a $10,000 Budget; An Extra $4,000 Unlocks 81.8% Accuracy, 2023. arXiv:2306.15658. 2

[43] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-Vocabulary Semantic Segmentation With Mask-Adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 2

[44] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 17612–17625. Curran Associates, Inc., 2022. 2, 3, 5

[45] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-World Semantic Segmentation via Contrasting and Clustering Vision-Language Embedding. In *Computer Vision – ECCV 2022*, pages 275–292, Cham, 2022. Springer Nature Switzerland. 2

[46] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[47] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 4, 15

[48] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning, 2021. arXiv:2111.09734. 2

[49] Peter Norvig. Natural language corpus data: Beautiful data. http://norvig.com/ngrams. 4, 15

[50] David Nukrai, Ron Mokady, and Amir Globerson. Text-Only Training for Image Captioning using Noise-Injected CLIP. In

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4055–4063, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 2

[51] Chau Pham, Truong Vu, and Khoi Nguyen. LP-OVOD: Open-Vocabulary Object Detection by Linear Probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 779–788, 2024. 2

[52] Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2016. 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 4, 6

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 6

[55] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 2

[56] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How Much Can CLIP Benefit Vision-and-Language Tasks? In *International Conference on Learning Representations*, 2022. 2

[57] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and Where Pathways for Robotic Manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 894–906. PMLR, 2022. 2

[58] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[59] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale, 2023. arXiv:2303.15389. 2

[60] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing Human Motion Generation to CLIP Space. In *Computer Vision – ECCV 2022*, volume 13682, pages 358–374. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 2

[61] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17918–17928, 2022. 2

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[63] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust Fine-Tuning of Zero-Shot Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, 2022. 2

[64] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning Robust Audio Representations from Clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567, Singapore, Singapore, 2022. IEEE. 2

[65] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024. 1

[66] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. CLIM: Contrastive Language-Image Mosaic for Region Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6117–6125, 2024. 2

[67] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for Open-Vocabulary Detection With Region Prompting and Anchor Pre-Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7031–7040, 2023. 2

[68] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 3, 13

[69] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[70] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 2

[71] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, 2023. 2

[72] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, et al. Florence: A New Foundation Model for Computer Vision, 2021. arXiv:2111.11432. 2

[73] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021. 2

[74] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 2, 6, 8

[75] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point Cloud Understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8552–8562, 2022. 2

[76] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize Anything: A Strong Image Tagging Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1724–1732, 2024. 1, 2, 7, 23

[77] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-Based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022. 2