Vocabulary-free Image Classification and Semantic Segmentation

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, Elisa Ricci

✦

Abstract—Large vision-language models revolutionized image classification and semantic segmentation paradigms. However, they typically assume a pre-defined set of categories, or vocabulary, at test time for composing textual prompts. This assumption is impractical in scenarios with unknown or evolving semantic context. Here, we address this issue and introduce the Vocabulary-free Image Classification (VIC) task, which aims to assign a class from an unconstrained language-induced semantic space to an input image without needing a known vocabulary. VIC is challenging due to the vastness of the semantic space, which contains millions of concepts, including fine-grained categories. To address VIC, we propose Category Search from External Databases (CaSED), a training-free method that leverages a pre-trained vision-language model and an external database. CaSED first extracts the set of candidate categories from the most semantically similar captions in the database and then assigns the image to the best-matching candidate category according to the same vision-language model. Furthermore, we demonstrate that CaSED can be applied locally to generate a coarse segmentation mask that classifies image regions, introducing the task of Vocabulary-free Semantic Segmentation. CaSED and its variants outperform other more complex vision-language models, on classification and semantic segmentation benchmarks, while using much fewer parameters. Code is available at https://github.com/altndrr/vicss.

Index Terms—Vision and Language, Classification, Segmentation.

1 INTRODUCTION

Large-scale Vision-Language models (VLMs) [1]–[3] have revolutionized the field of computer vision, connecting multimodal information in an unprecedented manner. One peculiar aspect of these models is their zero-shot transfer capabilities: for instance, CLIP [1] showed outstanding classification results in multiple datasets, even if not being explicitly trained for the task at hand. This lead to extending VLMs to other discriminative tasks, such as semantic segmentation [4]–[6] or object detection [7], [8], where their multimodal nature allows to perform such tasks in an "open-vocabulary" manner, *i.e.*, where the (finite) set of categories can be dynamically defined by the user.

In this paper, we aim to challenge the latter assumption and perform classification tasks with VLMs *without* a set

A. Conti, M. Mancini, P. Rota, and E. Ricci are with the University of Trento (Trento, Italy), Y. Wang and E. Ricci are with Fondazione Bruno Kessler (Trento, Italy). E. Fini was affiliated with the University of Trento and is now at Apple. The work was done before joining Apple. Corresponding author: A. Conti (alessandro.conti-1@unitn.it).

of target categories (*i.e.*, the vocabulary) pre-defined by the user (see Fig. 1). This has many practical advantages as this lack of priors often arises when working with autonomous agents in unconstrained environments. At the same time, it inherits various challenges as (i) the search space encompasses all possible existing semantic concepts, even very fine-grained ones that are difficult to discriminate and possibly ambiguous; (ii) we need classification models that do not rely on vocabulary-aware supervision, to avoid potential biases on the sub-part of the vocabulary within the training data. We name this task Vocabulary-free Image Classification (VIC).

Our approach exploits two core elements: multimodal representations from a contrastive VLM, *i.e.*, CLIP [1], and the information included in large scale vision-language databases (VLD), *e.g.*, PMD [9]. For classification, given an image we retrieve its closest captions in a VLD, encoding both input and captions via the CLIP encoders. We then parse these captions to obtain a set of candidate class names for the input. We encode the candidates via the text encoder of CLIP, scoring them according to their similarity with the visual input and the centroid of the retrieved captions, performing multimodal matching. We name this approach Category Search from External Databases (CaSED). On a variety of VIC benchmarks, CaSED achieves higher performance than computationally more expensive VLMs for VQA. We additionally introduce UpperCaSED, which extends CaSED with prompt ensembling to further improve the results with minimal computational overhead.

To further demonstrate the effectiveness of our proposed approach, we extended CaSED for the task of Vocabulary-free Semantic Segmentation (VSS). In contrast to the limited nature of image-level classification, often overlooking objects in the background, semantic segmentation aligns with the challenges of unconstrained environments, containing unforeseen objects that cause ambiguities in defining a fixed set of classes. In particular, when faced with the absence of predefined categories, the segmentation task becomes more complex, prompting questions about the appropriate granularity, *e.g.*, whether to segment object parts or the entire object itself. Moreover, segmentation poses new challenges for CaSED, given the tendency of vision-language models to recognize foreground objects and ignore the background, and the object-centric nature of internet-sourced captions.

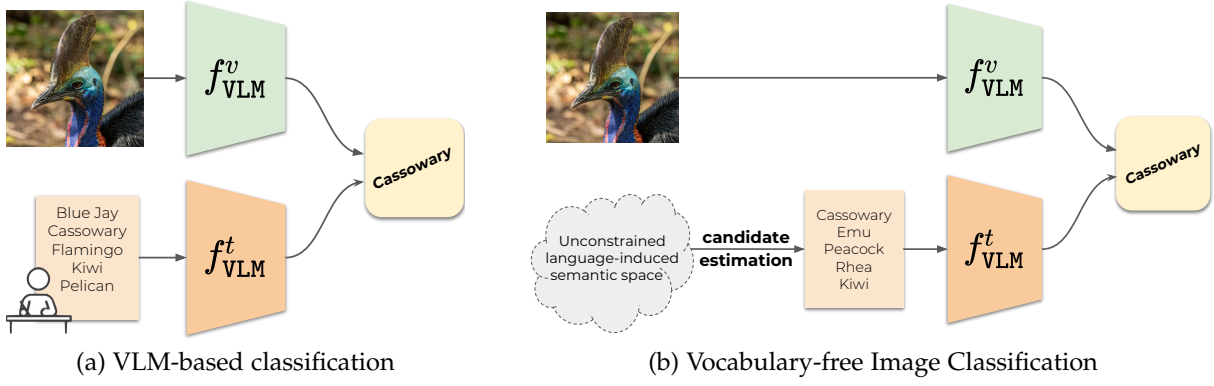(a) VLM-based classification   (b) Vocabulary-free Image Classification

Fig. 1. Vision-Language Model (VLM)-based classification (a) assumes a pre-defined set of target categories, *i.e.*, the vocabulary, while our novel task (b) lifts this assumption by directly operating on the unconstrained language-induced semantic space. f_{VLM}^v and f_{VLM}^t denote the pre-trained vision and text models of a VLM, respectively. In this work, we also extend this paradigm for the task of semantic segmentation.

In this context, we explore different strategies: the first is to use an off-the-shelf segmenter to obtain an initial set of masks. CaSED can then assign a semantic label to each mask independently. A second strategy is to do the opposite: CaSED can provide estimates of the semantic categories that can then be processed by an open-vocabulary segmentation model. Finally, we may avoid any external segmentation model and only employ a single pre-trained VLM, without additional fine-tuning. Encoding non-overlapping patches separately and classifying their content may seem a good strategy, but we found that it leads to noisy results because a single patch cannot capture the surrounding visual context. To address this issue, we first encode local information of the image by dividing it into cells of different sizes and processing them via CLIP. These multi-scale representations are then accumulated locally to obtain a more precise dense visual representation. The latter undergoes the same CaSED processing, retrieving a set of captions and candidate categories for each local representation. We then apply multimodal scoring on each cell, obtaining the final, dense semantic prediction. We name this approach DenseCaSED. Experiments show that DenseCaSED and CaSED variants outperform various semantic segmentation models in multiple benchmarks, *without* requiring any training procedure.

To summarize, the contributions of this work are:

- We present the tasks of VIC and VSS, where the goal is to classify/segment images *without* any pre-defined set of target categories, operating directly on an unconstrained semantic space. We define specific metrics for these tasks, capturing the semantic between predictions and ground-truth labels, providing a reference for future research.
- We present Category Search from External Databases, a training-free method for VIC that exploits multimodal representation of VLMs and captioning database to obtain a coarse set of candidate categories and ranks them according to their multimodal similarity with the input and retrieved captions. We also expand this strategy via prompt ensembling to further improve the performance, naming this variant UpperCaSED.
- We extend this method for VSS, presenting three variants. While two of them couple CaSED with pretrained segmentation models, the third, DenseCaSED directly exploits a VLM and multi-scale image processing to obtain

local visual representations, that are used to retrieve and score candidates at a local level, providing a dense semantic map of the input without any class priors or training.
- Our extensive evaluation on different benchmarks and with different VLM-based models, demonstrate the efficacy of CaSED and UpperCaSED for VIC, and of DenseCaSED for VSS, highlighting the potential of VLM plus retrieval as a pipeline for semantic categorization tasks with an unconstrained vocabulary.

This article extends our previous work [10] in multiple aspects. First, while [10] proposed VIC and CaSED, here we extend the latter by exploring improvements of the multimodal scoring mechanism via prompt ensembling. Moreover, we show the generality of CaSED by applying it to a more recent VLM with a slightly different pre-training objective (*i.e.*, SigLIP [11]), and including more powerful baselines (*i.e.*, LLaVa 1.5 [12]). The main contribution is, however, on the application, as we formalize the task of Vocabulary-free Semantic Segmentation, performing segmentation on an unconstrained vocabulary. In this regard, (i) we propose metrics specific for this task, capturing how well local semantic predictions match with the ground-truth maps, and (ii) we benchmark on multiple datasets (*i.e.*, PascalVOC-20 [13], PASCAL Context-59 [14], and ADE20K-150 [15]) and various competitors, even training-based open-vocabulary ones. We also show three methods to extend CaSED for VSS, either using pretrained segmentation models or fully relying on pretrained VLM and captions databases (DenseCaSED). These benchmarks, metrics and results, as well as DenseCaSED, will serve as reference for future work aiming to reduce the need of user inputs for semantic segmentation.

2 RELATED WORK

Vision-Language Models. The recent surge in models that map image-text pairs into a shared representation space has been largely driven by the availability of large-scale datasets [9], [16], [17]. These models [1], [18]–[23] employ modality-specific encoders and a contrastive objective to align the output representations of the two modalities. A prime example of this approach is CLIP [1], which has demonstrated impressive results in zero-shot classification

tasks. Further enhancements to CLIP have been proposed, including the integration of cross-modal attention [22], multi-object representation alignment [24], learning from weak-supervision [25], and leveraging unaligned data [9]. A separate stream of research has focused on enhancing vision-language pre-training for more intricate vision-language tasks, such as image captioning and visual question answering (VQA) [2], [3], [12], [26], [27]. Within this domain, BLIP [3] uses web data and generated captions to guide the pre-training of a multimodal architecture, surpassing existing VLMs in both captioning and VQA tasks, and LLaVA [12] aligns and CLIP vision encoder and LLAMA large-language model [28] to reason on visual inputs.

Here, we question a core premise of zero-shot classification with VLMs: the prior knowledge of the target classes. We present VIC, that bypasses this assumption, performing classification within an open-ended, language-induced space of semantic categories. In this setting, even advanced methods like BLIP-2 [29] struggle, whereas caption databases offer valuable priors for deducing the semantic class of an image. It is important to distinguish VIC from open-vocabulary recognition (*e.g.*, [30], [31]), as the latter still operates assuming that the list of target classes is known and accessible to the model during inference.

Retrieval augmented models. In the field of natural language processing, a number of studies have demonstrated the advantages of retrieving information from external databases to enhance the performance of large language models [32]–[34]. This approach has also found applications in computer vision, particularly in addressing the issue of class imbalance. For instance, some works have focused on long-tail recognition by learning to retrieve training samples [35] or image-text pairs from an external database [36]. Another study, [37], retrieves images from a specific dataset to learn fine-grained visual representations. More recently, the concept of retrieval-augmentation has been expanded to various types of sources for visual question answering [38], as well as to condition the generative process in diffusion models [39], and even image captioning [40].

Our work shares similarities with [36] in that we also utilize an external database. However, unlike [36] which assumes a pre-defined set of classes (and data) available for training a retrieval module, our approach does not make this assumption due to the vast semantic space of VIC. In our method, CaSED, we use retrieval to first generate a set of candidate classes, and then to perform the final class prediction. Furthermore, we assume the database to contain only captions, and not necessarily paired image-text data, thus making our approach less memory-intensive.

Semantic segmentation. The earliest works tackling semantic segmentation with an open vocabulary learns a joint embedding space between pixels and class names [41]–[44]. After the surge in image-text models such as CLIP [1], the paradigm shifted towards exploiting such web-scale pre-trained models as priors to tackle the problem [4]–[6], [31]. All prior works on the task assume a fixed pre-defined list of candidate names to address the task of semantic segmentation. Such simplification, however, is unrealistic for real-world applications, where, *e.g.*, a pre-defined list

of class names is restrictive and not exhaustive for most scenarios. Different from previous approaches, we aim to tackle a more challenging setup where this list of class names is unavailable and must be inferred from the input.

Recently, we formalized a similar task for image classification [10] and our objective is to expand such scenario to dense classification tasks. The most similar work to our proposed task of Vocabulary-free Semantic Segmentation is zero-guidance segmentation [45]. Differently from them, we formalize the Vocabulary-free Semantic Segmentation, strengthening the evaluation protocol, proposing principled metrics for the vocabulary-free scenario, and reusing traditional semantic segmentation benchmarks to assess the performance of multiple baselines methods.

3 VOCABULARY-FREE IMAGE CLASSIFICATION AND SEMANTIC SEGMENTATION

Preliminaries. Let us denote as $\mathcal{X} \subset \mathbb{R}^{N \times 3}$ the image space, where N is the number of pixels. Moreover, we can define as $\mathcal{C}$ a set of class labels. These labels are semantic entities in the much larger space of all possible semantic concepts $\mathcal{S}$, *i.e.*, $\mathcal{C} \subset \mathcal{S}$. A classifier/segmenter, is a function f mapping images/pixels to semantic labels in $\mathcal{C}$, with $f : \mathcal{X} \to \mathcal{C}$ in the case of classification and $f : \mathcal{X} \to \mathcal{C}^N$ in the case of segmentation. While f is usually trained on paired samples of images and labels, this approach is costly and does not scale with the cardinality of $\mathcal{C}$, as it may require expensive manual annotation. VLMs [1], [46] removed the need for explicit annotations, measuring similarities between image and text descriptions, *i.e.*, $f_{\text{VLM}} : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$, with $\mathcal{T}$ the textual space. In this way, we can perform classification by:

$$f(\boldsymbol{x}) = \arg \max_{\boldsymbol{c} \in C} f_{\text{VLM}}(\boldsymbol{x}, \phi(\boldsymbol{c})) \tag{1}$$

where $\phi(\boldsymbol{c})$ denotes a text concatenation, merging a static text template, or *prompt*, with a class name. Segmentation is achieved in a similar manner, computing the similarity between text and local image representations [4]. Note that this does not require any training and the model can classify/segment images into new categories defined at test time without retraining, performing zero-shot transfer. Nevertheless, this approach assumes that the set of categories $\mathcal{C}$ is given a priori by a user. Here, we describe Vocabulary-free Image Classification (VIC) to overcome this limitation in classification, and introduce its counterpart Vocabulary-free Semantic Segmentation (VSS) for segmentation.

Task definition. The goal of the vocabulary-free settings is to either classify (VIC) or segment (VSS) an image $\boldsymbol{x}$ without any prior knowledge about $\mathcal{C}$. Specifically, this means operating directly on the vast semantic space $\mathcal{S}$, which encompasses all semantic concepts. For VIC, we aim to devise a function $f : \mathcal{X} \to \mathcal{S}$ that maps an image to a semantic label within $\mathcal{S}$. Similarly, for VSS, the target function is $f : \mathcal{X} \to \mathcal{S}^N$ mapping images to semantic maps. Note that, at test time, f relies solely on the input image $\boldsymbol{x}$ and a broad repository of semantic concepts approximating $\mathcal{S}$. The task of VIC is inherently demanding due to the immense number of potential semantic classes in $\mathcal{S}$. For
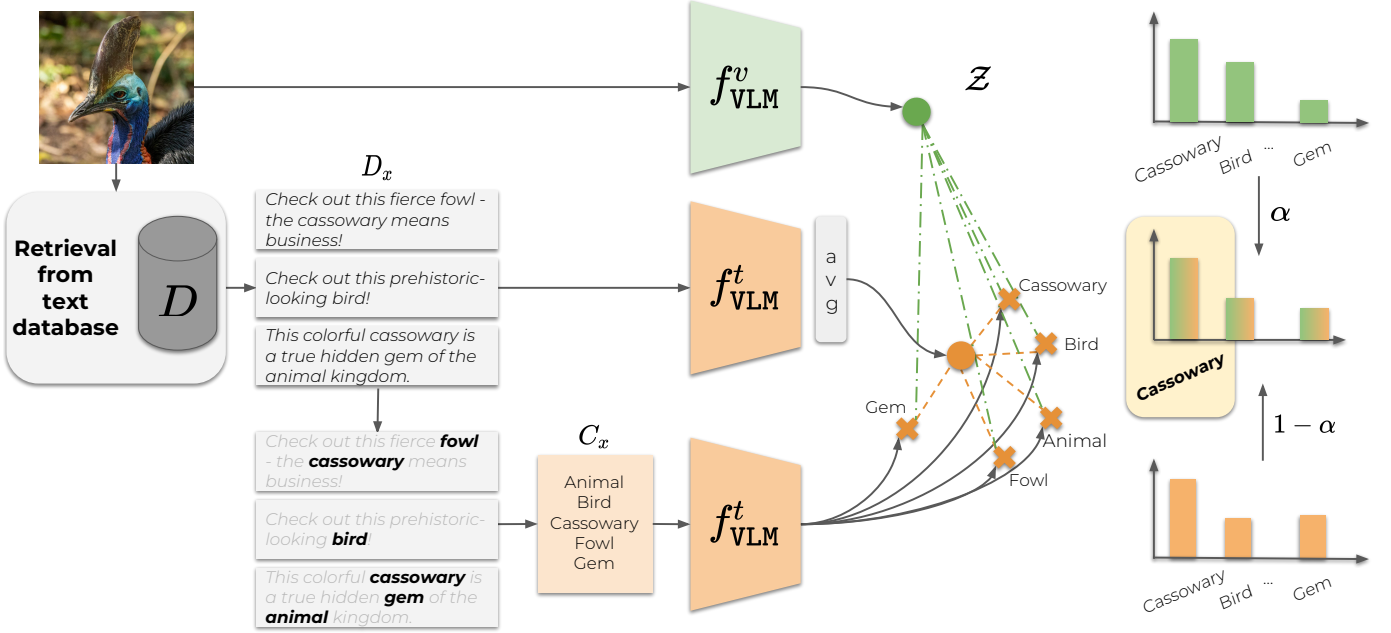
Fig. 2. **CaSED**. Given an input image, CaSED retrieves the most relevant captions from an external database filtering them to extract candidate categories. We classify image -to- text and text -to- text , using the retrieved captions centroid as the textual counterpart of the input image.

perspective, ImageNet-21k [47] is 200 times smaller than the cardinality of the semantic classes in BabelNet [48]. This vast search space presents significant challenges in differentiating nuanced concepts across diverse domains and those with a long-tailed distribution.

Challenges. Both VIC and VSS share the challenge of identifying which semantic categories in the large set $\mathcal{C}$ are present in the input image [10]. In particular, VSS is hard as $\mathcal{C}$ contains a lot of potential distractors, *i.e.*, concepts related to the ones in the image but not present. Examples are *couch* vs *sofa*, *tv* vs *monitor*, but also different animal species of plants. Thus, addressing VSS requires models with fine-grained recognition capabilities. On the other hand, a model may also segment two regions using synonyms (*e.g.*, *lawn* vs *grass*, *road* vs *highway*): these cases need to be disambiguated to obtain coherent segmentation masks. This also relates to other issues, such as the granularity of the segmentation masks (e.g., parts vs entire objects), or the object-centric focus of VLMs and internet-sourced captions. The latter, tend to ignore objects or elements in the background, making it hard to provide extremely fine-grained segmentation masks.

In the next sections, we describe how we address these challenges by (i) constraining the output space via external captions; (ii) disambiguating semantic concepts via multimodal matching, and (iii) propagating local features.

4 METHOD

In the following, we first describe Category Search from External Databases (CaSED) [10], for tackling VIC. The method leverages the power of large Vision-Language Datasets (VLDs) to find the best matching category within an unconstrained semantic space. We then describe some modifications that allow us to improve performance of the

method while also increasing its speed. Finally, we present how CaSED can be extended to semantic segmentation, either by application on top of pretrained segmentation network or by modifying how CLIP processes the input image (DenseCaSED). Fig. 2 shows an overview of CaSED.

4.1 Category Search from External Databases

CaSED is built on two modules: (i) candidate categories generation an (ii) multimodal scoring of the list of candidates. In the following, we describe each of them, and how we further improve the performance and speed in UpperCaSED.

Candidate category generation. We initially narrow down the vast classification space $\mathcal{S}$ to few probable candidate classes. Given an input $\boldsymbol{x}$, we use the pre-trained VLM f_{VLM} and an external image captions database D to retrieve a subset $D_{\boldsymbol{x}} \subset D$ of K closest captions to the input image as

$$D_{\boldsymbol{x}} = \underset{\boldsymbol{d} \in D}{\text{top-k}} \ f_{\text{VLM}}(\boldsymbol{x}, \boldsymbol{d}) = \underset{\boldsymbol{d} \in D}{\text{top-k}} \ \langle f_{\text{VLM}}^v(\boldsymbol{x}), f_{\text{VLM}}^t(\boldsymbol{d}) \rangle, \quad (2)$$

where $f_{\text{VLM}}^v : \mathcal{X} \rightarrow \mathcal{Z}$ and $f_{\text{VLM}}^t : \mathcal{T} \rightarrow \mathcal{Z}$ are the visual and textual encoders of the VLM, respectively, and $\mathcal{Z}$ is their shared embedding space. The operation $\langle \cdot, \cdot \rangle$ computes the cosine similarity between the two. Our approach can accommodate varying database sizes and is not dependent on the specific form of D. We then extract a finite set of candidate classes $C_{\boldsymbol{x}}$ from $D_{\boldsymbol{x}}$ using basic text parsing and filtering techniques. More details in Appendix A.

Multimodal candidate scoring. We score each candidate in the set $C_{\boldsymbol{x}}$ using both visual and textual semantic similarities via the VLM encoders to identify the best-matching class for the input image. We denote $s_{\boldsymbol{c}}^v$ as the visual score of each candidate category $\boldsymbol{c}$, computed as the similarity between

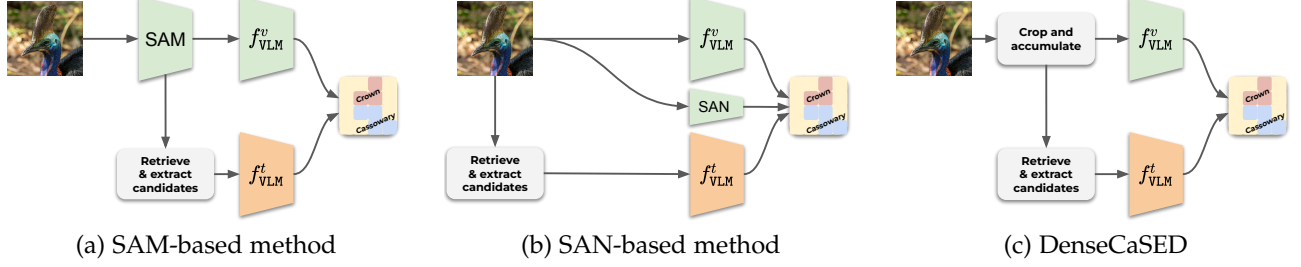(a) SAM-based method   (b) SAN-based method   (c) DenseCaSED

Fig. 3. **Extending CaSED for Semantic Segmentation**. We follow three strategies: (a) a class-agnostic segmenter (SAM) segments all objects, then CaSED labels each mask independently; (b) CaSED provides candidate categories for the image that are fed as input to an open-vocabulary segmentation model (SAN); (c) DenseCaSED, where we directly accumulate visual features from multi-scale patches, and perform CaSED locally.

the visual representation of the input image and the textual representation of the candidate name:

$$s_{\boldsymbol{c}}^{v} = \langle f_{\mathrm{VLM}}^{v}(\boldsymbol{x}), f_{\mathrm{VLM}}^{t}(\boldsymbol{c}) \rangle. \tag{3}$$

To mitigate the modality gap in the space $\mathcal{Z}$, we introduce a unimodal text-to-text scoring. Denoting the centroid $\bar{\boldsymbol{d}}_{\boldsymbol{x}}$ of the retrieved captions, the text-based matching score $s_{\boldsymbol{c}}^{t}$ is:

$$\bar{\boldsymbol{d}}_{\boldsymbol{x}} = \frac{1}{K} \sum_{\boldsymbol{d} \in D_{\boldsymbol{x}}} f_{\mathrm{VLM}}^{t}(\boldsymbol{d}), \tag{4}$$

$$s_{\boldsymbol{c}}^{t} = \langle \bar{\boldsymbol{d}}_{\boldsymbol{x}}, f_{\mathrm{VLM}}^{t}(\boldsymbol{c}) \rangle. \tag{5}$$

The use of the caption centroids as anchor for classification is motivated by the analyses in [10], showing how they are well correlated with the semantic of the visual input. We report this analysis in Appendix A. We obtain the final score $s_{\boldsymbol{c}}$ for each candidate $\boldsymbol{c}$ by merging the two scores, as:

$$s_{\boldsymbol{c}} = \alpha \, \sigma(s_{\boldsymbol{c}}^{v}) + (1 - \alpha) \, \sigma(s_{\boldsymbol{c}}^{t}) \tag{6}$$

where $\sigma(\cdot)$ is the softmax operation on the two scores of each candidate class, and α is a hyperparameter. The output category is $f_{\mathrm{CaSED}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{c} \in C_{\boldsymbol{x}}} s_{\boldsymbol{c}}$. Our approach, CaSED, is *training-free*, uses a pre-trained and frozen VLM, and is flexible for various architectures and databases.

4.2 UpperCaSED

To improve the performance of CaSED, we propose a simple yet effective modification, introducing prompt ensembling [1] after the caption extraction and filtering. Prompt ensembling applies a predefined list of templates to the class names to enhance their contextual information. For instance, the class name "dog" could be expanded to "a photo of a dog". By applying a set of templates rather than a single one and computing their average representation, the resulting features better capture the semantic meaning of the word, leading to consistent performance gains [1].

We apply prompt ensembling to the candidates generated from the retrieved captions. The number of templates used is variable, depending on the specific dataset[1]. The average representation of these ensembled prompts is then used as anchors to compute the image-to-text and text-to-text scores as described above. Formally, for a candidate

1. For each classification dataset, we use the templates defined in CLIP [1], spanning from 1 for, *e.g.*, Flowers-102 [49], Food-101 [50], and Oxford Pets, to 48 for UCF101 [51]

$\boldsymbol{c}$ and a set of templates T, the ensembled representation $f_{\mathrm{ens}}^{t}(\boldsymbol{c})$ is computed as:

$$f_{\mathrm{ens}}^{t}(\boldsymbol{c}) = \frac{1}{|T|} \sum_{t \in T} f_{\mathrm{VLM}}^{t}(t(\boldsymbol{c})), \tag{7}$$

where $t(\boldsymbol{c})$ means applying template t to candidate $\boldsymbol{c}$. $f_{\mathrm{ens}}^{t}(\boldsymbol{c})$ is then used for image-to-text and text-to-text scoring.

5 CaSED FOR VOCABULARY-FREE SEMANTIC SEGMENTATION

To extend CaSED for the task of semantic segmentation, we can follow three strategies. The first is to exploit an available class-agnostic segmentation model [52], extract segmentation masks and then assigning them a label via CaSED, independently. The second does the opposite: the initial set of candidates generated by CaSED is the input to an open-vocabulary segmentation network [5]. While these strategies lead to good results but they require the additional computational cost of using another segmentation network. We thus propose a third approach which generates local visual representations directly from patches of the input image. In the following, we describe the three strategies.

5.1 Coupling CaSED with Segmentation Networks

Assigning semantics to class-agnostic masks with CaSED. Let us assume to have a class-agnostic segmentation network f_{SEG} that, given as input an image $\boldsymbol{x}$ maps it to a set of k segmentation masks $M = \{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_k\}$. Note that these masks have no semantic attached, and the number of masks k may be input dependent [52]. From the masks, we extract a set of image regions $R = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_k\}$, *e.g.*, by cropping around the relative mask. We can then attach a semantic to each region using CaSED, propagating the prediction to the pixels of the relative mask. Given a pixel $x \in \boldsymbol{x}$ that assigned to mask $\boldsymbol{m}_i \in M$ by f_{SEG}, its semantic label is simply $f_{\mathrm{CaSED}}(\boldsymbol{r}_i)$. The rationale behind this approach is that the semantic of a pixel is the one assigned by CaSED to an image extracted from the mask the pixels belong to.

Candidates generation with CaSED for Open-Vocabulary Segmentation. The main limitation of the previous approach is that extracting meaningful image regions from masks may require solutions that exploit VLM priors on object localization (*e.g.*, circle drawing [53]). To sidestep this

problem, we can invert the pipeline by first obtaining candidate class-names using CaSED and then segmenting the image using an open-vocabulary segmentation model, *e.g.*, [54]. Specifically, an open-vocabulary segmentation model takes as input an image and a set of possible labels $\mathcal{Y} \subseteq \mathcal{S}$ and maps them to a segmentation mask, assigning pixels to elements of $\mathcal{Y}$, *i.e.*, $f_{\text{OV-SEG}} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Y}^N$. As we will show experimentally in Sec. 6, if we do not take into account the challenges of VSS and set $\mathcal{Y} = \mathcal{S}$, we obtain poor segmentation results. This is mainly due to the extremely large cardinality of $\mathcal{S}$, requiring to distinguish local, fine-grained differences of potentially similar semantic concepts.

To overcome these challenges, we follow the rationale behind CaSED, restricting the search space by estimating a set of candidate classes. Thus, given an image $\boldsymbol{x}$, we define a set of candidates $C_{\boldsymbol{x}}$ by filtering a set of captions $D_{\boldsymbol{x}}$, with the latter obtained as in Eq. (2). We then feed the set $C_{\boldsymbol{x}}$ as input to the open-vocabulary segmentation network and obtain the relative segmentation mask as $f_{\text{OV-SEG}}(\boldsymbol{x}, C_{\boldsymbol{x}})$.

5.2 DenseCaSED

The previous approaches rely on the presence of an additional (pretrained) semantic segmentation model, and the assumption that the module is not biased toward particular input distributions. Here we propose a different strategy, directly exploiting the available VLM.

The approach applies CaSED to local image representations, as done with the class-agnostic strategy previously described. As we do not have access to masks, we need to define local image regions that we can feed as input to the VLM. To obtain such representations, We divide the image in multiple grids, where each grid has n^2. For simplicity, we choose n to be powers of 2, *i.e.*, $n \in \{1, 2, 3\}$. We also replicate the grids by shifting them vertically and/or horizontally by a stride equal to half the size of a grid cell. This creates a hierarchy of patches where neighboring patches are likely to belong to the same super patch and therefore their representation (loosely) depends on each other. This helps in embedding contextual information in the aggregated pixel-level representation.

Formally, let us denote the visual representation of a pixel i in $\boldsymbol{x}$ as $\boldsymbol{l}_i$. Moreover, let us denote as $\{\boldsymbol{g}_i^1, \cdots, \boldsymbol{g}_i^N\}$ all patches that contain i. The local representation $\boldsymbol{l}_{i,j}$ is then

$$\boldsymbol{l}_i = \frac{1}{N} \sum_{q=1}^{N} f_{\text{VLM}}^v(\boldsymbol{g}_i^q). \qquad (8)$$

Note that we divide the aggregated value for each pixel by the number of times it was forwarded within a cell. As these local representations are already encoded using a VLM, we can retrieve the most relevant set of captions to a cell from a VLD D using the cosine similarity of the embeddings, as:

$$D_{\boldsymbol{l}_i} = \underset{\boldsymbol{d} \in D}{\text{top-k}} \langle \boldsymbol{l}_i, f_{\text{VLM}}^t(\boldsymbol{d}) \rangle, \qquad (9)$$

For each cell in $\boldsymbol{g} \in G_g$, we follow the pipeline of CaSED, by (i) filtering $D_{\hat{\boldsymbol{l}}}$ to obtain the corresponding set of candidates $C_{\boldsymbol{g}}$; (ii) computing the visual score as in Eq. (3) but using $\hat{\boldsymbol{l}}$

as visual representation; (iii) compute the textual scoring as in Eq. (5) with $D_{\hat{\boldsymbol{l}}}$ as set of captions; (iv) merging the two scores to compute the final multimodal one, as in Eq. (6). The CaSED predictions on accumulated local visual features are then propagated to the whole cell, producing the final segmentation mask. We name this approach DenseCaSED.

Note that this approach is not only training-free but does not use any segmentation network, relying only on a contrastive-based VLM. Moreover, accumulating local cells representations across scales allows to model the context in which a cell appears while enforcing a consistent vocabulary across neighboring cells. This latter aspect is important in VSS, as modeling cells in isolation may lead to inconsistent choices of labels in the large search space (e.g. "sofa" vs "couch") leading to lower segmentation results.

6 EXPERIMENTS

6.1 Classification

Datasets. As in previous studies [55], [56], we use ten datasets that span both broad and detailed classification in various domains. These datasets include Caltech-101 (C101) [57], DTD [58], EuroSAT (ESAT) [59], FGVC-Aircraft (Airc.) [60], Flowers-102 (Flwr) [49], Food-101 (Food) [50], Oxford Pets (Pets), Stanford Cars (Cars) [61], SUN397 (SUN) [62], and UCF101 (UCF) [51]. For tuning hyperparameters, we use the ImageNet dataset [47].

Evaluation metrics. The unrestricted nature of the semantic space in VIC requires unique evaluation metrics. In [10], we propose two primary measures: *semantic relevance*, which assesses the similarity between the predicted and actual labels, and *image grouping*, which evaluates the quality of image clustering based on the predicted labels. For semantic relevance, we consider two aspects: i) *Semantic Similarity*, which measures the similarity between the predicted and actual labels in a semantic space, and ii) *Semantic Intersection over Union (IoU)*, which calculates the overlap of words between the prediction and the ground truth. Formally, given an input $\boldsymbol{x}$ with ground-truth label $\boldsymbol{y}$ and prediction $\hat{\boldsymbol{c}} = f(\boldsymbol{x})$, the *Semantic Similarity* is computed as $\langle g(\hat{\boldsymbol{c}}), g(\boldsymbol{y}) \rangle$, where $g : \mathcal{T} \rightarrow \mathcal{Y}$ is a function that maps text to an embedding space $\mathcal{Y}$. To accommodate free-form text, we employ Sentence-BERT [63] as g. For *Semantic IoU*, given a predicted label $\boldsymbol{c}$ (considered as a set of words), we compute the Semantic IoU as $|\boldsymbol{c} \cap \boldsymbol{y}| / |\boldsymbol{c} \cup \boldsymbol{y}|$, where $\boldsymbol{y}$ is the set of words in the ground-truth label. To evaluate image grouping, we use the traditional *Cluster Accuracy* metric inspired by protocols for deep visual clustering [64]–[66]. This involves clustering images based on their predicted labels and then assigning each cluster to a ground-truth label with a many-to-one match, where a predicted cluster is assigned to the most common ground-truth label.

Baselines. We consider a diverse set of baselines, categorized into three groups. The first uses CLIP with extensive vocabularies, such as WordNet [67], which contains approximately 117k names, and the English Words dataset [68], featuring around 234k names. As a theoretical upper limit, we evaluate CLIP with an ideal vocabulary, specifically

	Method	Cluster Accuracy (%) ↑										
		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	**Avg.**
CLIP	WordNet	34.0	20.1	16.7	16.7	58.3	40.9	52.0	29.4	18.6	39.5	32.6
CLIP	English Words	29.1	19.6	22.1	15.9	64.0	30.9	44.4	24.2	19.3	34.5	30.4
Caption	Closest Caption	12.8	8.9	16.7	13.3	28.5	13.1	15.0	8.6	20.0	17.8	15.5
Caption	BLIP-2 (ViT-L)	26.5	11.7	23.3	5.4	23.6	12.4	11.6	19.5	14.8	25.7	17.4
Caption	BLIP-2 (ViT-g)	37.4	13.0	25.2	10.0	29.5	19.9	15.5	21.5	27.9	32.7	23.3
Caption	LLaVA 1.5 (7B)	41.1	11.1	19.7	10.4	13.4	11.1	12.8	14.0	12.1	29.5	17.5
VQA	BLIP-2 (ViT-L)	60.4	20.4	21.4	8.1	36.7	21.3	14.0	32.6	28.8	44.3	28.8
VQA	BLIP-2 (ViT-g)	62.2	23.8	22.0	15.9	57.8	33.4	23.4	36.4	**57.2**	**55.4**	38.7
VQA	LLaVA 1.5 (7B)	**76.2**	**30.6**	**38.9**	3.0	5.8	22.7	7.7	27.5	2.6	48.0	26.3
	CaSED	51.5	29.1	23.8	22.8	68.7	58.8	60.4	37.4	31.3	47.7	43.1
	UpperCaSED	51.3	29.3	21.0	**24.4**	**70.4**	**61.2**	**60.9**	**37.7**	38.5	46.6	**44.1**
	CLIP upper bound	87.6	52.9	47.4	31.8	78.0	89.9	88.0	65.3	76.5	72.5	69.0

TABLE 1
Cluster Accuracy on the ten datasets. Green is our method, gray shows the upper bound.

	Method	Semantic IoU (%) ↑										
		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	**Avg.**
CLIP	WordNet	15.0	3.0	1.3	0.5	31.3	7.8	14.7	9.0	4.8	3.8	9.1
CLIP	English Words	8.0	2.0	0.0	1.1	16.4	2.0	17.2	8.1	2.7	1.8	5.9
Caption	Closest Caption	4.5	0.8	1.3	1.9	5.9	3.1	3.0	2.3	11.4	1.0	3.5
Caption	BLIP-2 (ViT-L)	13.4	1.4	4.8	0.0	7.5	4.7	1.7	4.7	11.6	1.1	5.1
Caption	BLIP-2 (ViT-g)	16.8	1.8	4.1	0.1	13.9	7.9	2.9	5.7	24.7	1.9	8.0
Caption	LLaVA 1.5 (7B)	13.8	0.87	5.13	0.0	2.9	1.7	0.3	3.5	4.6	1.4	3.4
VQA	BLIP-2 (ViT-L)	36.1	1.8	7.0	0.1	21.5	3.7	5.7	11.5	18.9	2.5	10.9
VQA	BLIP-2 (ViT-g)	**41.5**	2.4	**7.5**	2.0	**38.0**	8.6	10.2	13.8	**33.2**	2.8	16.0
VQA	LLaVA 1.5 (7B)	41.4	0.4	6.1	0.0	5.0	3.0	0.7	6.6	1.3	2.0	6.6
	CaSED	35.4	**5.1**	2.3	**4.8**	33.1	**19.4**	**35.1**	17.2	16.2	**8.4**	17.7
	UpperCaSED	37.8	4.6	5.2	4.4	35.2	18.9	34.9	**17.8**	16.0	7.8	**18.3**
	CLIP upper bound	86.0	52.2	51.5	28.6	75.7	89.9	88.0	66.6	84.5	71.3	69.4

TABLE 2
Semantic IoU on the ten datasets. Green is our method, gray shows the upper bound.

the ground-truth names from the target dataset (CLIP upper bound). While we primarily present results for CLIP with the ViT-L architecture [69] due to space constraints, additional results utilizing other architectures are in Appendix A. The second group of baselines includes captioning methods, directly describing the semantic content of images. We explore two approaches: one that retrieves captions from a database and another that generates captions using a pre-trained image captioning model. For caption retrieval, we use the same VLD of CaSED. For caption generation, we use BLIP-2 [29], a VLM known for its exceptional performance across various tasks, including image captioning, to generate image descriptions. The third group use a Visual Question Answering (VQA) model to directly infer the class in the image. We use BLIP-2 [29], and extend the baselines of [10] with LLaVA 1.5 (7B) [12], a larger VLM.

Implementation Details. We conduct our experiments on NVIDIA A6000 GPUs, using mixed-precision for efficiency. We use a subset of PMD [9] as our database, which includes five of its largest datasets: Conceptual Captions (CC3M) [70], Conceptual Captions 12M (CC12M) [71], Wikipedia Image Text (WIT) [72], Redcaps [73], and the portion of YFCC100M* [74] curated for PMD. For retrieval, we embed the database with the CLIP text encoder f_{VLM}^t, using a fast indexing technique, i.e., FAISS [75]. The hyperparameter

α in Eq. (6) and the number of retrieved captions K are fixed to $\alpha = 0.7$ and $K = 10$ via selection on ImageNet.

Quantitative results. The performance of CaSED is superior to all baseline models across all metrics, as shown in Tab. 1, Tab. 3, and Tab. 2. Notably, CaSED outperforms BLIP-2 (VQA) over ViT-g by $+4.4\%$ in Cluster Accuracy and $+1.7\%$ on Semantic IoU, while using significantly fewer parameters. The performance gap is even wider when compared to the same visual backbone, with CaSED surpassing BLIP-2 on ViT-L (VQA) by $+14.3\%$ on Cluster Accuracy, $+5.7$ in Semantic Similarity, and $+6.8\%$ on Semantic IoU. In addition, the CLIP with retrieval a large pre-defined vocabulary does not yield effective results for VIC, likely due to the challenge of classifying over a vast search space with difficult-to-model class boundaries. This is evidenced by CLIP with English Words and with WordNet underperforming across all metrics. Captioning models, despite their ability to capture image semantics in challenging settings, exhibit high variability across images of the same category, resulting in poor performance on Clustering and Semantic IoU. LLaVA 1.5 (VQA), while generally underperforming compared to BLIP-2 (VQA) and CaSED, does achieve the best results for cluster accuracy on Caltech-101, DTD, and EuroSAT. However, its performance on semantic IoU and semantic similarity is not significantly better than other approaches,

	Method	Semantic Similarity (x100) ↑										
		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	**Avg.**
CLIP	WordNet	48.6	32.7	24.4	18.9	55.9	49.6	53.7	44.9	28.8	44.2	40.2
	English Words	39.3	31.6	19.1	18.6	43.4	38.0	44.2	36.0	19.9	34.7	32.5
Caption	Closest Caption	42.1	23.9	23.4	29.2	40.0	46.9	40.2	39.8	49.2	40.3	37.5
	BLIP-2 (ViT-L)	57.8	31.4	39.9	24.4	36.1	44.6	29.0	45.3	46.4	38.0	39.3
	BLIP-2 (ViT-g)	63.0	33.1	36.2	24.3	45.2	51.6	31.6	48.3	61.0	44.6	43.9
	LLaVA 1.5 (7B)	56.8	29.4	40.5	21.3	31.1	36.9	24.8	42.5	38.1	37.9	35.9
VQA	BLIP-2 (ViT-L)	70.5	34.9	29.7	29.1	48.8	42.0	40.0	50.6	52.4	48.6	44.7
	BLIP-2 (ViT-g)	**73.5**	36.5	31.4	**30.8**	**59.9**	52.1	43.9	**53.3**	**65.1**	**55.1**	50.1
	LLaVA 1.5 (7B)	72.6	36.7	**44.1**	29.4	41.8	41.1	36.0	41.9	35.3	46.6	42.6
	CaSED	65.7	40.0	32.0	30.3	55.5	64.5	62.5	52.5	47.4	54.1	50.4
	UpperCaSED	66.3	**40.3**	34.9	27.0	56.1	**65.0**	**62.9**	53.0	46.5	53.8	**50.6**
	CLIP upper bound	90.8	69.8	67.7	66.7	83.4	93.7	91.8	80.5	92.3	83.3	82.0

TABLE 3

Semantic Similarity (x100) on the ten datasets. Values multiplied by x100 for readability. Green highlights our method and gray the upper bound.

and on average, it achieves unsatisfactory results across the ten datasets. These results highlight the efficacy of CaSED in all metrics. Finally, UpperCaSED consistently improves performance across all datasets, with an average gain of $+1.0\%$, $+0.6$, and $+0.2$ on the three metrics w.r.t. CaSED.

6.2 Semantic segmentation

Datasets. We experiment with three datasets: Pascal VOC [13] (VOC-20), PASCAL Context-59 [14] (CTX-59), and ADE20K-150 [15] (ADE-150). As opposed to [54], [76], [77], we do not use COCO Stuff [78], as it is used for training and all the considered baselines and methods are training-free.

Evaluation metrics. To address the openness VSS, we extend two popular metrics for open-vocabulary semantic segmentation, namely Jaccard Index (JI) and Recall (R). We will refer to the metrics as Hard JI (HJI) and Hard Recall (HR). For their soft variants, we replace the binary "hard" values (*i.e.*, zero and one) with the semantic similarity between predicted and ground-truth word. The Soft Jaccard Index (SJI), directly accounts for this similarity at pixel-level. Formally, for a pixel p with label y and prediction $\hat{c} = f(p)$, the textual similarity is computed as $\langle g(\hat{c}), g(y) \rangle$, where the function $g : \mathcal{T} \to \mathcal{Y}$ maps text to an embedding space $\mathcal{Y}$. As in [10], we use Sentence-BERT [63] as g. Similarly, Soft Recall (SR) expands the recall metric with the semantic proximity of the predicted and ground-truth classes in the image.

Furthermore, we introduce two variants of the JI, *i.e.*, Nearest Jaccard Index (NJI) and Overlap Jaccard Index (OJI) to account for cases where proposed words may not perfectly align with the annotations due to linguistic ambiguities or to specificity of the proposed segmentation masks, *i.e.*, part vs whole cases (*e.g.*, predicting "head" and "shirt" on two parts of a "person"). We propose to map predicted names to ground-truth ones, evaluating the traditional JI on the projected predicted mask. More formally, given the predicted segmentation mask C and its ground-truth mask Y, we extract the lists of predicted names $L^C \in C$ and the ground-truth names $L^Y \in Y$. We then create a mapping $M : \mathcal{L}^C \to \mathcal{L}^Y$, so that each predicted word is mapped to one ground-truth word. The criteria behind this mapping is the main difference between NJI and OJI. In the former, we use textual similarity between predictions and the list of annotated words. In the latter, we directly evaluate the co-occurrence of predictions and annotations in the pixel space.

Baselines. We consider two groups of baseline methods for comparison. The first exploits SAM [52] to first extract regions and then propose region-based candidate names via retrieval or generation with, *e.g.*, English Words or BLIP-2. The second uses a open-vocabulary semantic segmentation model in the absence of the pre-defined list of class names for the dataset, requiring the ad-hoc generation with traditional vocabulary-free methods. In this context, we employ SAN [54] as it offers a minor variation from the CLIP architecture, incorporating only an auxiliary network to address the task on the pre-trained backbone. For both groups, we report results with the same set of baselines of VIC, without the captioning baselines. We also report results when CaSED is used for vocabulary generation. For all main experiments, we use CLIP with the ViT-L/14 backbone.

Implementation details. We conduct all our experiments following the same setup as for classification. DenseCaSED introduces a single hyperparameter, *i.e.*, the grid sizes to crop the input image. We empirically set $N = \{2, 4, 8\}$ to have a final dense pixel map of 16×16, for a total of 256 sub-regions for each input image. For computational efficiency, we use FastSAM [79] in place of SAM for segmentation.

Quantitative results. We report results on Tab. 4, Tab. 5, and Tab. 6. Our experiments consistently demonstrate that English Words and WordNet are sub-optimal for the task, independently of whether a method first segment the image into regions of interest and then classify (*i.e.*, SAM-based methods), or first generates candidate names and then segment (*i.e.*, SAN-based methods). With SAM, the best approach exploits LLaVA 1.5 (7B) followed by BLIP-2 (both ViT-L and ViT-g), generally achieving comparable results across the six metrics and three datasets, with, *e.g.*, 12.7 and 12.3 on NJI, and 50.8 and 47.3 on SR, respectively. The strongest method consists of using SAN with CaSED to generate a custom list of candidates for each image and then segment with the open-vocabulary segmentation model. When compared with more naive approaches, *i.e.*,

Method		VOC-20					
		HJI	NJI	OJI	SJI	HR	SR
		Jaccard Index				Recall	
SAM	English words	4.2	12.1	48.9	11.1	5.4	30.4
	WordNet	4.9	14.5	48.2	12.5	6.1	35.9
	BLIP-2 (ViT-L)	15.8	19.2	31.7	17.2	33.4	60.6
	BLIP-2 (ViT-g)	15.1	17.5	27.2	16.4	33.1	61.7
	LLaVA 1.5 (7B)	18.3	18.9	29.3	17.5	**41.6**	**63.6**
SAN	English words	5.8	18.6	48.4	13.8	7.2	38.1
	WordNet	7.2	19.3	46.4	15.7	8.7	45.3
	SAM + CaSED	13.7	17.5	44.6	15.5	20.1	48.0
	SAN + CaSED	**26.9**	**30.2**	**53.2**	**20.8**	34.2	61.8
	DenseCaSED	20.5	20.3	32.2	18.5	31.4	61.8

TABLE 4
Semantic segmentation on PascalVOC-20. Green are our methods.

Method		ADE-150					
		HJI	NJI	OJI	SJI	HR	SR
		Jaccard Index				Recall	
SAM	English words	2.2	4.5	**31.2**	6.3	3.3	27.4
	WordNet	2.3	6.3	30.9	6.4	3.5	27.45
	BLIP-2 (ViT-L)	5.0	7.7	24.7	7.7	10.5	36.5
	BLIP-2 (ViT-g)	4.9	8.2	23.9	7.7	10.9	37.8
	LLaVA 1.5 (7B)	6.2	7.1	24.2	7.8	14.4	40.4
SAN	English words	2.2	4.5	19.2	7.2	3.6	30.5
	WordNet	2.8	5.1	19.1	7.5	4.8	32.9
	SAM + CaSED	6.1	7.7	29.4	7.4	10.0	35.2
	SAN + CaSED	7.2	7.5	20.8	8.7	11.3	40.9
	DenseCaSED	**8.6**	**9.1**	24.1	**8.8**	**16.8**	**43.6**

TABLE 6
Semantic segmentation on ADE20K-150. Green are our methods.

Method		CTX-59					
		HJI	NJI	OJI	SJI	HR	SR
		Jaccard Index				Recall	
SAM	English words	1.0	6.4	**40.2**	8.9	2.5	27.0
	WordNet	1.0	8.7	39.6	9.6	2.5	28.9
	BLIP-2 (ViT-L)	7.0	10.5	28.4	12.1	16.1	42.7
	BLIP-2 (ViT-g)	6.5	10.3	27.4	12.0	15.5	44.4
	LLaVA 1.5 (7B)	9.1	12.0	29.7	12.9	**21.4**	48.3
SAN	English words	1.2	8.0	34.9	10.3	2.7	30.7
	WordNet	2.0	9.4	33.0	11.2	3.6	33.6
	SAM + CaSED	7.5	11.0	38.2	11.5	11.2	39.4
	SAN + CaSED	**15.5**	**16.2**	38.1	**14.7**	20.8	46.9
	DenseCaSED	13.4	13.1	32.6	13.9	20.8	**48.6**

TABLE 5
Segmentation on PASCAL Context-59. Green highlights our methods.

English Words or WordNet for retrieval, SAN with CaSED improves by, *e.g.*, about +12.9 on HJI and +17.0 on HR. SAN with CaSED and the best SAM-based approaches generally performs comparably on recall metrics, whether hard or soft. Notably, DenseCaSED, despite not including any component pre-trained for semantic segmentation, achieves and sometimes surpasses all the SAM-based approaches, even achieving the highest scores on hard and soft recall, *i.e.*, +0.9% and +1.4% against SAN with CaSED. Due to the coarse nature of the segmentation mask, however, DenseCaSED falls behind against methods trained for open-vocabulary semantic segmentation, while keeping an edge against SAM-based methods. Specifically, it achieves −2.3 and −1.0 on HJI and SJI against SAN with CaSED, but improves on average by +4.4 and +1.8 against SAM with LLaVA (7B) or BLIP-2.

6.3 Ablation studies

In this section, we study various aspects of our approach. First, we validate the impact of the retrieval database. Then, we demonstrate how prompt ensemble improves the performance of retrieval-based baselines compared to image-text comparison. Lastly, we analyze how the backbone and training objective impact the performance of CaSED.

Retrieval database. We examine the effects of sourcing captions from different databases with various noise levels and sizes (*e.g.*, from 0.07M captions of Ade20K to 413.8M of LAION-400M). For each database, we report cluster accuracy, semantic similarity and semantic IoU on ImageNet, alongside dataset statistics, *i.e.*, percentage of nouns, adjectives, verbs, and number of concepts. Last, we also report the semantic similarity of the closest caption (C.C. S-Sim.) for each ImageNet sample. As indicated in Tab. 7, there is a general trend of improved results with increasing database size. However, the quality of the captions plays also role a role, as seen from CC12M and Redcaps, achieving results that are either on par with or slightly superior to our database, combining CC3M, WIT, Redcaps, CC12M, and a subset of YFCC100M. Interestingly, the average performance generally correlates with the C.C. S-Sim., hinting is as a possible criterion for database selection in VIC and VSS.

Prompt ensemble. We report the performance of prompt ensembling on all the retrieval-based baselines, *i.e.*, English Words, WordNet, and CaSEDin Tab. 8. The results show consistent improvements across all baselines when using multiple templates. This outcome is consistent across the three metrics, except for one configuration, *i.e.*, cluster accuracy for English Words. On average, prompt ensemble improves performance by +0.3% on cluster accuracy, +1.2 on semantic similarity, and +0.9 on semantic IoU.

Scaling CaSED backbone. We test CaSED considering different backbones. Specifically, we employ ResNet-50, ViT-L/14, and ViT-L/16. The first two backbones are pre-trained on 400M image-text samples by OpenAI, while the latter is trained with a sigmoid loss [11] on WebLI [81]. For this analyses, we use the smaller CC12M as retrieval database instead of PMD, as they achieve comparable performance in Tab. 7. We report the results in Tab. 9 Notably, our approach shows positive changes when equipped with an higher number of parameters and better training losses, hinting that larger VLMs can lead to higher VIC performance.

7 CONCLUSIONS

In this work, we formalise two tasks, Vocabulary-free Image Classification and Vocabulary-free Semantic Segmentation, both operating on an unconstrained semantic space, tackling classification at image- and pixel-level without a pre-defined set of classes. We propose a method to tackle the first,

Database	Size	CA	S-Sim.	S-IoU	Nouns (%)	Adjs (%)	Verbs (%)	Concepts	C.C. S-Sim.
Ade20K	0.07M	11.0	31.8	1.7	82.7	10.6	6.7	3.1K	19.0
COCO	0.9M	13.8	35.6	2.6	86.1	9.5	4.4	14.6K	22.6
SBU Captions	1.0M	20.1	40.8	5.7	98.0	1.0	0.9	29.3K	26.9
OpenImages	1.4M	16.8	37.8	4.2	85.4	10.0	4.5	21.4K	24.8
Loc. Narr.	1.9M	15.7	37.1	4.1	86.9	9.8	3.2	30.9K	24.9
CC3M	2.8M	37.9	53.9	16.2	57.9	23.6	18.5	28.4K	35.2
WIT	4.8M	32.9	47.9	14.5	99.8	0.05	0.07	163.7K	30.0
Visual Genome	5.4M	14.6	35.2	3.8	75.1	15.6	9.2	21.1K	26.7
Redcaps	7.9M	41.1	54.7	19.6	98.6	0.7	0.6	50.5K	37.5
CC12M	10.3M	**43.8**	**57.4**	**21.2**	96.6	1.8	1.5	36.7K	**38.7**
YFCC100M*	29.9M	39.1	53.5	18.9	99.7	0.2	0.1	179.6K	<u>37.8</u>
Ours	54.8M	<u>41.5</u>	<u>55.7</u>	<u>20.4</u>	99.8	0.05	0.05	334.1K	**38.7**
LAION-400M	413.8M	37.7	52.7	18.7	99.4	0.5	0.1	1.68M	37.3

TABLE 7

ImageNet accuracy of CaSED for different databases. Blue are the top-5 ones of PMD, used for Ours . *C.C. S-Sim.* is the semantic similarity of the closest caption to each image in the dataset. **Bold** is best, while <u>underline</u> second best. YFCC100M* denotes the same subset used in PMD.

Vocabulary	Prompt ensemble	CA	S-Sim.	S-IoU
WordNet		32.6	40.2	9.1
	✓	33.3	41.8	10.1
English Words		30.4	32.5	5.9
	✓	29.6	34.3	7.0
Ours		43.1	50.4	17.7
	✓	44.1	50.6	18.3

TABLE 8

Ablation on prompt ensembling. Results are averaged on the ten classification datasets. Green is our configuration.

Backbone	Size (M)	CA	S-Sim.	S-IoU
ResNet-50	102	32.7	50.1	15.4
ViT-L/14	427	43.5	56.9	21.2
ViT-L/16 (SigLIP-384)	652	**47.8**	**59.4**	**22.5**

TABLE 9

Ablation on CaSED backbone. Results are collected on ImageNet, using CC12M as the text database. Sizes are taken from [80].

CaSED, confirming its efficacy across a wide range of benchmark datasets. In addition, we present an upgraded version of CaSED for classification and a modification enabling dense image classification. We term the two variations UpperCaSED and DenseCaSED. We confirm the validity of our approaches across a wide range of benchmark datasets and evaluation metrics. For both the tasks, we present a suite of metrics designed for the openness of classification and semantic segmentation for vocabulary-free settings.

Limitations and future works. The efficacy of the CaSED family of methods is strongly influenced by the selection of the retrieval database, with potential challenges in retrieving concepts that are not adequately represented therein. For instance, if the application domain encompasses fine-grained concepts, a generic database may not be appropriate. Similarly, CaSED may mirror potential biases in the dataset. Nevertheless, we can flexibly address these issues by progressively incorporating new concepts into the database (including domain-specific ones) from textual corpora, without the need for retraining. Enhancements in bias mitigation and data quality control could alleviate this problem. Future studies could investigate strategies for automatically selecting or expanding a database based on

test samples.

Both CaSED and DenseCaSED (or variants) do not maintain a record of their output history, which could lead to inconsistent predictions. This can lead in VIC to assign slightly different labels to images of the same semantic concept (for example, *cassowary* versus *Casuarius*). In VSS this may even lead to inconsistencies at the image-level, with distant pixels of the same category receiving different labels (*e.g.*, *ground* vs *floor*). Implementing a memory feature in CaSED that stores the predicted labels could resolve this issue for classification, while for segmentation predictions could benefit from post-processing techniques merging regions with semantically and visually similar features. Lastly, CaSED and its variants do not handle class granularity. For instance, in classification a *cassowary* could also be predicted as a *bird*. In segmentation, this may require understanding the level of detail of the masks (*e.g.*, parts vs whole). Future research could clarify such instances by explicitly incorporating user needs into the VIC and VSS family of models. Finally, while our DenseCaSED does not use a segmentation network and does not require training, it performs multiple forwards of the same input image to generate a dense feature map. This is not computationally efficient at inference time and provides only coarse spatial information. The next steps for DenseCaSED could develop tailored, even training-based, paradigms to produce finer dense feature maps faster.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv*, 2022.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[4] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023.

[5] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.

[6] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," *NeurIPS*, 2023.

[7] C. Ma, Y. Jiang, X. Wen, Z. Yuan, and X. Qi, "Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection," *NeurIPS*, 2023.

[8] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "F-vlm: Open-vocabulary object detection upon frozen vision and language models," *ICLR*, 2023.

[9] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022.

[10] A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang, and E. Ricci, "Vocabulary-free image classification," *NeurIPS*, 2023.

[11] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *ICCV*, 2023.

[12] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv*, 2023.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, 2010.

[14] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.

[15] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *IJCV*, 2019.

[16] C. Schuhmann, R. Kaczmarczyk, A. Komatsuzaki, A. Katta, R. Vencu, R. Beaumont, J. Jitsev, T. Coombes, and C. Mullis, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," in *NeurIPS Workshop Datacentric AI*, 2021.

[17] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.

[18] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *ECCV*, 2016.

[19] L. Gomez, Y. Patel, M. Rusinol, D. Karatzas, and C. Jawahar, "Self-supervised learning of visual features through embedding images into text topic spaces," in *CVPR*, 2017.

[20] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *CVPR*, 2021.

[21] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *NeurIPS*, 2021.

[23] E. Fini, P. Astolfi, A. Romero-Soriano, J. Verbeek, and M. Drozdzal, "Improved baselines for vision-language pre-training," 2023.

[24] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, 2022.

[25] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," in *ICLR*, 2022.

[26] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *CVPR*, 2022.

[27] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *NeurIPS*, 2022.

[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv*, 2023.

[29] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv*, 2023.

[30] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *CVPR*, 2021.

[31] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *ECCV*, 2022.

[32] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *ICML*, 2020.

[33] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *NeurIPS*, 2020.

[34] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improving language models by retrieving from trillions of tokens," in *ICML*, 2022.

[35] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019.

[36] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *CVPR*, 2022.

[37] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, and H. Jégou, "Grafit: Learning fine-grained image representations with coarse labels," in *ICCV*, 2021.

[38] Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. A. Ross, and A. Fathi, "Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory," in *CVPR*, 2023.

[39] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," *NeurIPS*, 2022.

[40] R. Ramos, D. Elliott, and B. Martins, "Retrieval-augmented image captioning," in *EACL*, 2023.

[41] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *CVPR*, 2019.

[42] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *NeurIPS*, 2019.

[43] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[44] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, "Open vocabulary scene parsing," in *ICCV*, 2017.

[45] P. Rewatbowornwong, N. Chatthee, E. Chuangsuwanich, and S. Suwajanakorn, "Zero-guidance segmentation using zero segment labels," *ICCV*, 2023.

[46] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[48] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *ACL*, 2010.

[49] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

[50] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV 2014*, 2014.

[51] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv*, 2012.

[52] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *ICCV*, 2023.

[53] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," *ICCV*, 2023.

[54] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *CVPR*, 2023.

[55] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," *arXiv*, 2022.

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, 2022.

[57] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *CVPR Workshop*, 2004.

[58] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.

[59] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[60] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv*, 2013.

[61] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV Workshops*, 2013.

[62] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.

[63] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *EMNLP-IJCNLP*, 2019.

[64] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in *ECCV*, 2020.

[65] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *ICCV*, 2019.

[66] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," in *ICLR*, 2019.

[67] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, 1995.

[68] FreeBSD, "Web2 dictionary (revision 326913)," 2023, accessed May 17, 2023. https://svnweb.freebsd.org/base/head/share/dict/web2?view=markup&pathrev=326913.

[69] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[70] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

[71] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021.

[72] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Research and Development in Information Retrieval*, 2021.

[73] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv*, 2021.

[74] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, 2016.

[75] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *Transactions on Big Data*, 2019.

[76] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023.

[77] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *CVPR*, 2023.

[78] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018.

[79] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," 2023.

[80] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," 2021.

[81] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv*, 2022.

[82] Y. Cui, L. Zhao, F. Liang, Y. Li, and J. Shao, "Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision," *arXiv*, 2022.
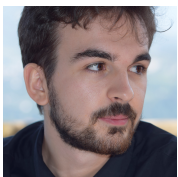
Enrico Fini is a Research Scientist at Apple in Zurich. He completed his Ph.D. at the University of Trento in the Multimedia and Human Understanding Group (MHUG). During his Ph.D., Enrico interned at FAIR (Meta AI), Amazon, SAP AI Research, and Inria Grenoble. His research interests include self-supervised learning, vision-language pre-training, continual learning, and open-set recognition.

Massimiliano Mancini is an assistant professor at the University of Trento, in the Multimedia and Human Understanding Group. He completed his Ph.D. at the Sapienza University of Rome, and was a postdoc at the Cluster of Excellence ML, University of Tübingen. He was a member of the ELLIS Ph.D. program, the TeV lab at Fondazione Bruno Kessler, and the VANDAL lab of the Italian Institute of Technology. His research interests include transfer learning and compositionality.

Paolo Rota is an assistant professor at the Center for Mind and Brain (CIMeC) in the University of Trento. He received his Ph.D. from the same university and has worked as a postdoctoral Marie Curie fellow at TU Wien and as a postdoc at the Istituto Italiano di Tecnologia in Genoa. He also worked as an ML researcher at the ProM Facility in Rovereto. He has been working as an assistant professor at the University of Trento since 2019 and started his tenure track in 2022. His research interests are focused on image and video classification using Vision and Language.
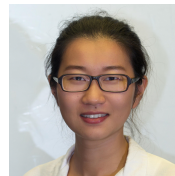
Yiming Wang is a researcher in the Deep Visual Learning (DVL) unit at Fondazione Bruno Kessler (FBK). Yiming obtained her PhD in Electric Engineering from Queen Mary University of London (UK) in 2018. She is a member of ELLIS and an associate editor of International Journal of Social Robotics. She has expertise in robotic perception and scene understanding. Her recent research focuses on training-free methods addressing open-world recognition.

Alessandro Conti is a PhD student at the University of Trento, in the Multimedia and Human Understanding Group (MHUG). His research interests include multimodal learning, self- and weakly-supervised learning, and domain adaptation.

Elisa Ricci is an associate professor with the University of Trento and a head of research unit with Fondazione Bruno Kessler. She received the Honorable mention award at ICCV 2021 and the Best Paper Award at ACM MM 2021. Her research interests are mainly in the areas of computer vision and deep learning. She is an ELLIS fellow.

APPENDIX

A. SEMANTIC SPACE REPRESENTATION

As the main challenge of VIC and VSS, how to represent the large semantic space plays a fundamental role in the method design. We can either model the multimodal semantic space directly with a VLM equipped with an autoregressive language decoder [3] or via image-text retrieval from VLDs. Consequently, we can approach vocabulary-free tasks either via VQA-enabled VLMs by querying for the candidate class given the input image, or by retrieving and processing data from an external VLD to obtain the candidate class.

To investigate the two potential strategies, we perform a preliminary experimental analysis to understand how well the output of a method semantically captures the image category, or in other words, to assess the alignment of class boundaries in the visual and textual representations in VIC. Specifically, we compare the semantic accuracy of querying VQA VLMs and of retrieving from VLDs w.r.t. the ground-truth class labels. We consider the output of a method as correct if its closest textual embedding among the target classes of the dataset corresponds to the ground-truth class of the test sample[2]. We exploit the text encoder of CLIP (ViT-L) [1] to obtain textual embeddings.

Regarding experimented methods, we select BLIP-2 [29] to represent VQA-enabled VLMs for its state-of-the-art performance in VQA benchmarks, while we use a subset of PMD [9] as the VLD. In particular, we compare the following methods: i) *BLIP-2 VQA*, which directly queries BLIP-2 for the image category; ii) *BLIP-2 Captioning*, which queries BLIP-2 for the image caption; iii) *Closest Caption*, which is the closest caption to the image, as retrieved from the database; iv) *Caption Centroid*, which averages the textual embeddings of the 10 most similar captions to the input image. As we use CLIP embeddings, if visual and textual representations perfectly align, the performance would be the same as zero-shot CLIP with given target classes. We thus report zero-shot CLIP to serve as the upper bound for retrieval accuracy.

We experiment on a variety of test datasets for both coarse- and fine-grained classification (see details in Sec. 6), and report the results in Fig. 4. The average textual embedding of the retrieved captions (*i.e.* Caption Centroid) achieves the best semantic accuracy for 9 datasets out of 10, consistently surpassing methods based on BLIP-2. On average, the accuracy achieved by Caption Centroid is 60.47%, which is $+17.36\%$ higher than the one achieved by BLIP-2 Captioning (43.11%). Moreover, Captions Centroid achieves results much closer to the CLIP upper bound (67.17%) than the other approaches. Notably, such VLD-based retrieval is also computationally more efficient, faster (approximately 2 seconds for a batch size of 64 on a single A6000 GPU), and requires fewer parameters (approximately 10 times less) than BLIP-2 (see Tab. 12 in Appendix A).

The results of this preliminary study clearly suggest that representing the large semantic space with VLDs can

---

2. Note that this metric is not the standard accuracy in image classification as we use distances in the embedding space to ground predictions from the unconstrained semantic space to the set of classes in a specific dataset.

produce (via retrieval) semantically more relevant content to the input image, in comparison to querying VQA-enabled VLMs, while being computationally efficient. Based on this conclusion, we develop an approach, Category Search from External Databases (CaSED), that searches for the semantic class from the large semantic space represented in the captions of VLDs.

B. DETAILS ON BLIP-2 AND LLaVA PROMPTING

In the main paper, we compare CaSED and its variants by replacing the candidate generation module and prediction directly with a VLM with VQA capabilities, testing BLIP-2 [29] and LLaVA 1.5 (7B) [12]. Here we provide details on how we prompt these models.

In line with the BLIP-2 [29] demo, for captioning, we used the prompt "Question: what's in the image? Answer:". For VQA, we used "Question: what's the name of the object in the image? Answer: a". For LLaVA 1.5, we use the same prompts of BLIP-2, but append an "Omit any superfluous text." to avoid introductory or closing statements.

C. CANDIDATES FILTERING

With the closest captions retrieved from the external database given the input image (please refer to Sec. 4 of the main manuscript for further details), we post-process them to filter out a set of candidate category names. We first create a set of all words that are contained in the captions. Then we apply sequentially three different groups of operations to the set to (i) remove noisy candidates, (ii) standardize their format, and (iii) filter them.

With the first group of operations, we remove all the irrelevant textual contents, such as tokens (*i.e.*, "⟨ PERSON ⟩"), URLs, or file extensions. Note that, for the file extensions, we remove the extension but retain the file name as it might contain candidate class names. We also remove all the words that are shorter than three characters and split compound words by underscores or dashes. Finally, we remove all those terms containing symbols or numbers and meta words that are irrelevant to the classification task, such as "image", "photo", or "thumbnail". As shown in Table 10, when compared to having no operation for candidate filtering (first row), this set of operations removes inappropriate content and increases the accuracy of clusters by $+6.4\%$ and improves the semantic IoU by $+0.3$. However, we can observe a drop in semantic similarity by -1.5. This might be due to the removal of unnatural words that could still describe well the content of the image, *i.e.*, underline- or dash-separated words, or URLs since they are longer w.r.t. natural words.

The second group of operations standardize the candidate names by aligning words that refer to the same semantic class to a standard format, reducing class redundancy. For example, "cassowary" and "Cassowary" will be considered as a single class instead of two. To this end, we perform two operations—lowercase conversion and singular form conversion. With such standardizing conversions,
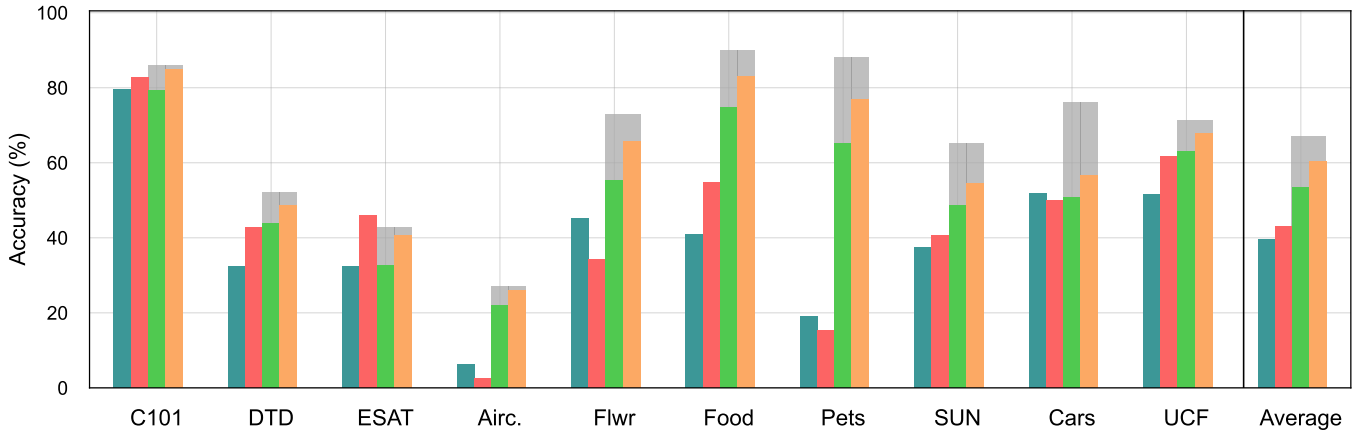
Fig. 4. Results of our preliminary study, showing the top-1 accuracy when matching semantic descriptions to ground-truth class names in ten different datasets. We compare BLIP-2 (VQA) and BLIP-2 (Captioning) with Closest Caption and Captions Centroid , *i.e.*, the average representation of the retrieved captions. We additionally highlight the Upper bound for zero-shot CLIP. Representing the large semantic space as VLDs and retrieving captions from it produces semantically more similar outputs to ground-truth labels w.r.t. querying outputs from VQA-enabled VLMs, while requiring 10 times fewer parameters compared to the latter.

we observe a sizeable boost in terms of performances when compared to the results obtained by applying only removal-related operations. As shown in Table 10, we achieve higher results across all three metrics, leading to a relative improvement of $+7.2\%$, $+0.7$, and $+1.2$ in terms of cluster accuracy, semantic similarity, and semantic IoU, respectively.

The last group of operations considers two forms of filtering, where the first aims to filter out entire categories of words via Part-Of-Speech (POS) tag and the second aims to filter out rare and noisy contents based on the word occurrences. We select these two operations since common dataset class names do not contain terms that carry no semantics, *e.g.*, articles and pronouns, and since [82] showed that CLIP performs better when exposed to a smaller amount of unique tokens. The POS tagging[3] categorizes words into groups, such as adjectives, articles, nouns, or verbs, enabling us to filter all the terms that are not semantically meaningful for a classification task. Regarding the occurrence filtering, we first count how often a word appears in the retrieved captions and then we remove words that appear only once to make the candidate list less noisy. We can see from Table 10 that the inclusion of this final set of operations scores the best among all three metrics when compared to the results obtained when only the previous two groups of operations are applied.

Number of captions vs number of selected candidates. To complement the previous analysis, in Tab. 11, we report the number of unique candidates extracted by the candidate filtering procedure, averaged over the ten datasets, and with an increasing number of retrieved captions, *i.e.*, 1, 2, 5, 10, and 20. In the table, we show both the number of candidates extracted and the number of selected words. As the number of retrieved captions increases, the unique number of candidate words also increases, *i.e.*, from 3849 with 1 caption to 28781 with 20. The number of selected words, however, stabilizes around 800 as soon as we retrieve more

3. We use the NLP library flair (https://github.com/flairNLP/flair).

than 1 caption. Having more captions reduces the noises in the selected words, something that might be present when relying on a single caption.

D. COMPUTATIONAL COST

We analyze the computational efficiency of CaSED versus BLIP-2 and LLaVA 1.5 performing VQA and captioning, and report their respective number of parameters and inference time in Tab. 12. Notably, the methods using external databases are consistently faster than BLIP-2 and LLaVA. For instance, CaSED achieves a speed-up of 2x with respect to both the largest captioning model and the largest VQA model, while also achieving better performance. In a similar fashion, our method is approximately 3.0x faster than LLaVA 1.5 for VQA and 3.5x for captioning. Overall, the fastest method is Closest Caption, which exploits the external database to retrieve a single caption and does not consider any candidate extraction pipeline. Conversely, our method retrieves the ten most similar captions and post-processes them to extract class name, resulting in a increase in inference time of approximately 2 times. Compared with the CLIP upper bound, our method considers multiple additional steps, each adding extra inference time. First, our method retrieves candidates from the external database to then extract the class names. Second, we have to forward the class names through the text encoder for each sample, while CLIP can forward them once and cache the features for later reuse.

When using an external database, note that an increase in database size implies a minimal variation in retrieval time. This is demonstrated by the computational cost required by retrieving from the large LAION-400M [16] database. As the results show, inference time is comparable between CaSED and CaSED (LAION-400M) despite the latter being approximately 8 times larger.

Operations			CA	S-Sim.	S-IoU
Remove	Standardize	Filter			
			27.7	48.8	15.0
✓			34.1	47.3	15.3
✓	✓		41.3	48.0	16.5
✓	✓	✓	**43.1**	**50.4**	**17.7**

TABLE 10

Ablation on the candidate filtering operations. Metrics are averaged across the ten datasets.

Num. captions	CA	S-Sim.	S-IoU	Candidates	Selected
1	30.9	43.2	12.4	3849	1047
2	35.0	46.4	14.6	6867	854
5	42.3	**50.6**	17.5	14548	775
10	**43.1**	50.4	**17.7**	22475	794
20	42.9	49.3	17.1	28781	802

TABLE 11

Ablation on the number retrieved captions. Green represents our selected configuration. We expand the results of Fig. 4 in the main manuscript to show the number of unique words extracted from captions (*i.e.*, "candidates") and the number of words selected by CaSED for classification (i.e., "selected"). Results are averaged across the ten datasets.

E. BACKBONE ARCHITECTURE

To answer the natural question about whether the outcome of our model depends on the backbone architecture, we further extend our main results with a CLIP model with a ResNet50 architecture and with a ViT-L/16 architecture pre-trained with the sigmoid loss [11]. We report this additional ablation in Tab. 13, Tab. 14, and Tab. 15 for the cluster accuracy, the semantic IoU, and the semantic similarity, respectively. We can see that the performance with the CLIP ResNet50 is lower across all the metrics compared to CLIP ViT-L/14. This is expected since ResNet50 is a a smaller architecture, thus with a reduced capacity for semantic representation learning as compared to ViT-L/14. Nevertheless, our method with ResNet50 is still competitive against BLIP-2 models while using 40x fewer parameters (note that our ViT-L implementation uses 10x fewer parameters). A similar trend occurs for the SigLIP family of models, as their performance generally surpasses CLIP with the ViT-L/14 backbone, while only having approximately $1.5\times$ more parameters. Notably, performance improves consistently, with an average improvement of $+3.8\%$ for cluster accuracy, $+1.6$ for semantic IoU, and $+2.1$ for semantic similarity. Differently from the smaller models, we do not experience a consistent improvement of UpperCaSED against CaSED with the SigLIP pre-trained backbone, where we notice a negative fluctuation on the three metrics when applying prompt ensembling.

F. QUALITATIVE RESULTS

Classification. Last, we report some qualitative results of CaSED applied on three different datasets, namely Caltech-101 (Fig. 5), Food101 (Fig. 6), and SUN397 (Fig. 7), where the first is coarse, and the last two are fine-grained, focusing on food plates and places respectively. For each, we present a batch of five images, where the first three represent success cases and the last two show interesting failure cases. Each sample shows the image we input to our method with the top-5 candidate classes.

From the results, we can see that for many success cases, our method not only generates the correct class name and selects it as the best matching label, but it also provides valid alternatives for classification. For example, the third image in Fig 5 or the second image in Fig 6, where CaSED provides the names "dessert" for the cheesecake and the label "bird" for the ibis. This phenomenon also happens in failure cases, where *e.g.* the last sample in Fig 6 provides both the name "pizza" and the name "margherita" for the dish, despite selecting the wrong name from the set.

Another interesting observation is that our method provides names for different objects in the same scene. For instance, the third and fourth samples in Fig 6 contain labels for both "guacamole" and "tortillas" for the first, and for "mozzarella", "insalata", and "balsamic" for the second. A further detail on the latter case is the ability of CLIP to reason in multiple languages since "insalata" translates to "salad" from Italian to English.

Regarding failure cases, it is interesting to note that the candidate names and the predicted label often describe well the input image despite being wrong w.r.t the dataset label. For instance, the two failure cases in Fig. 7 select "stadium" and "dumpsite" when the ground-truth class names are "football" and "garbage site". In addition, for the first case, the exact name "football" is still available among the best candidate names, but our method considers "stadium" as a better fit. Another example is the last failure case in Fig 5, where the model assigns the name "nokia" to a Nokia 3310 cellphone, while the ground-truth class name is "cellphone". Also in this case, the ground-truth label is present in the candidate list but our method considers "nokia" a more fitting class.

Finally, we notice the discovery of correlations between terms in the reasoning of our model. In the provided examples, it happens multiple times that the candidate class names do not describe objects in the scene but rather a correlated concept to the image. For instance, the third example in Fig 5 shows a Dalmatian, and among the candidate names there is "cruella", which is the name of the villain of the movie "The Hundred and One Dalmatians". Another instance of this appears in the first example of Fig 6, where the model correctly associates the "bibimbap" dish to its place of origin, Korea, with the candidate name "korean".

Segmentation. In a similar fashion, we report some qualitative results of CaSED for semantic segmentation, taking the best models in each group: of SAM with LLaVA 1.5 (7B), SAN with WordNet as vocabulary, and our localizer-free model for VSS, DenseCaSED. We report some output examples in Fig. 8, Fig. 9, and Fig. 10. From a segmentation perspective, it is notable how SAM achieves the best segmentation masks, despite missing some more local objects, *e.g.*, it misses all the objects except the cat in the first example of Fig. 9, or the horse in the first example of Fig. 10. Moreover, it sometimes misclassifies one object for another

	Method	Num. Params.	Inference time (ms) ↓
Caption	Closest Caption	0.43B	1390 ± 10
	BLIP-2 (ViT-L)	3.46B	5710 ± 153
	BLIP-2 (ViT-g)	4.37B	6870 ± 177
	LLaVA 1.5 (7B)	7.49B	8260 ± 20
VQA	BLIP-2 (ViT-L)	3.46B	5670 ± 135
	BLIP-2 (ViT-g)	4.37B	6650 ± 117
	LLaVA 1.5 (7B)	7.49B	7040 ± 28
	CaSED	0.43B	2630 ± 14
	CaSED (LAION-400M)	0.43B	2640 ± 11
	CLIP upper bound	0.43B	645 ± 77

TABLE 12

Computational cost of different methods. Green is our method, gray shows the upper bound. Inference time is reported on batches of size 64, as the average over multiple runs.
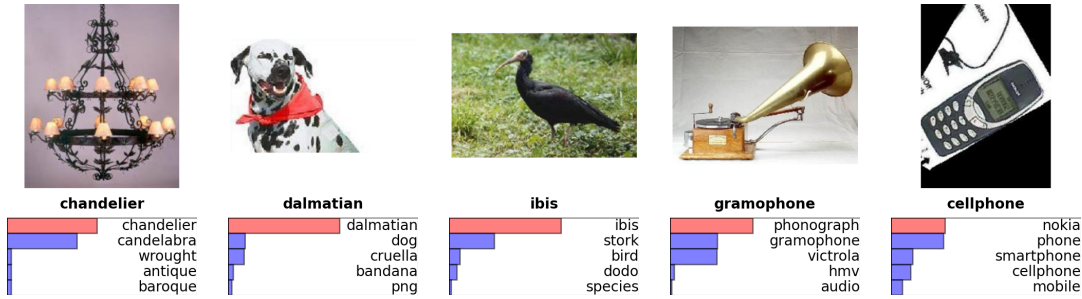
Fig. 5. Qualitative results of CaSED on Caltech-101. The first three samples represent success cases, the last two shows failure cases.
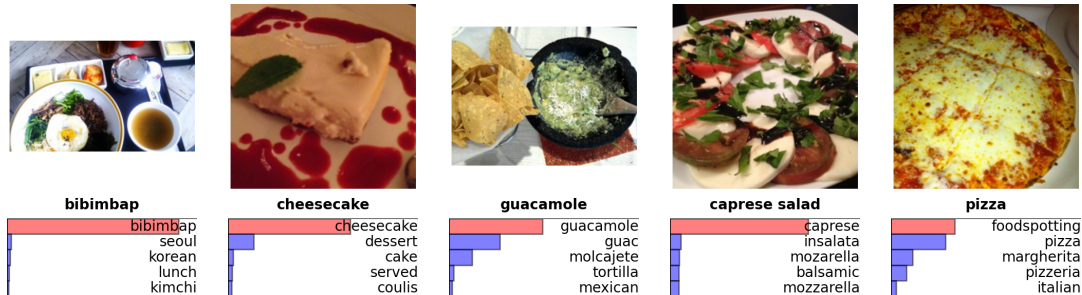

Fig. 6. Qualitative results of CaSED on Food101. The first three samples represent success cases, the last two shows failure cases.

(*e.g.*, it considers the table as a "car" in the second example of Fig. 10, or the aisle in the airplane as "staircase" in the second example of Fig. 8, or assigns the label "tree" to the sky and the car in Fig. 9). In some other cases, the most correct label for a region is assigned to a near segmentation mask, for instance in the second examples in Fig. 10 where it assigns "woman" to the hair and the wall behind the person, or "bottle" to the arm holding the bottle in the same example. We highlight that, nevertheless, SAM + LLaVA 1.5 (7B) is the model with the highest number of parameters overall, using two large-scale foundation models to address VSS.

Differently from SAM, SAN tends to segment whole objects instead of parts, *e.g.*, first and second example in Fig. 8, where it divides the image into foreground and background, or completely merge floor and ceiling of the airplane. Moreover, given the labels selected by SAN, it looks like the model tend to assign the most out-of-context words compared with SAM with LLaVA 1.5 (7B) and Dense-

CaSED. As an example, SAN assigns the word "Erin" to the person in the second example of Fig. 10, or "Cumberland Gap" to the landline in the second example in Fig. 9. Other examples include "pantile" or "westbound" assigned to the sky in the first example of Fig. 8 and in the second example of Fig. 9. Some of these issues are fixed when using CaSED to generate the vocabulary, as shown by the better quantitative results achieved in Tab. 4-6 by SAN + CaSED. Finally, compared to the SAM segmentation masks, the outputs presents slightly more noise on the edges, where regions are not perfectly segmented but are more approximate.

Compared to the previous approaches, our method shows more coarse segmentation masks mainly due to the lack of any segmentation model to generate them. This is noticeable in, *e.g.*, the first example of Fig. 8, where the chimney and the sky have no clean separation, or in the second example of Fig. 9, where the sky is not completely segmented as a single object. Moreover, it tends to propose larger amount of regions, *e.g.*, first example of Fig. 9,

	Method		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	Avg.
CLIP RN50	CLIP	WordNet	30.3	18.3	**22.5**	13.2	47.8	31.4	45.2	26.0	14.2	31.2	28.0
		English Words	24.8	17.5	18.5	13.4	49.5	23.1	36.6	22.2	15.5	27.1	24.8
	Caption	Closest Caption	9.7	7.1	13.3	8.4	21.2	6.2	8.7	6.5	12.8	14.9	10.9
		CaSED	**44.6**	23.9	12.5	15.3	58.8	48.7	50.1	32.8	24.6	**33.9**	34.5
		UpperCaSED	44.4	**24.3**	17.9	_16.7_	_62.2_	_49.1_	_55.1_	33.5	_29.0_	33.8	**36.6**
	CLIP upper bound		82.1	41.5	33.5	19.6	63.1	74.6	78.9	55.6	54.9	58.4	56.2
CLIP ViT-L/14	CLIP	WordNet	34.0	20.1	16.7	16.7	58.3	40.9	52.0	29.4	18.6	39.5	32.6
		English Words	29.1	19.6	22.1	15.9	64.0	30.9	44.4	24.2	19.3	34.5	30.4
	Caption	Closest Caption	12.8	8.9	16.7	13.3	28.5	13.1	15.0	8.6	20.0	17.8	15.5
		CaSED	**51.5**	29.1	**23.8**	22.8	68.7	58.8	60.4	37.4	31.3	**47.7**	43.1
		UpperCaSED	51.3	**29.3**	21.0	_24.4_	_70.4_	_61.2_	_60.9_	_37.7_	38.5	46.6	_44.1_
	CLIP upper bound		87.6	52.9	47.4	31.8	78.0	89.9	88.0	65.3	76.5	72.5	69.0
SigLIP ViT-L/16	SigLIP	WordNet	35.0	23.0	26.5	20.7	59.2	52.4	59.8	30.8	37.7	39.1	38.4
		English Words	28.6	21.4	21.4	19.0	60.3	41.2	53.6	24.1	38.0	33.8	34.1
	Caption	Closest Caption	15.6	9.7	26.2	23.9	36.5	19.3	17.2	8.9	36.3	26.9	22.0
		CaSED	56.3	_30.6_	**32.3**	24.6	66.4	_67.0_	**68.1**	**40.2**	45.2	48.5	_47.9_
		UpperCaSED	**58.1**	29.5	25.4	_29.5_	_69.7_	62.8	64.7	39.9	44.4	**49.6**	47.4
	CLIP upper bound		96.3	60.6	44.8	48.3	90.9	94.2	95.4	69.6	92.7	82.0	77.5
	Caption	BLIP-2 (ViT-L)	26.5	11.7	23.3	5.4	23.6	12.4	11.6	19.5	14.8	25.7	17.4
		BLIP-2 (ViT-g)	37.4	13.0	25.2	10.0	29.5	19.9	15.5	21.5	27.9	32.7	23.3
		LLaVA 1.5 (7B)	41.1	11.1	19.7	10.4	13.4	11.1	12.8	14.0	12.1	29.5	17.5
	VQA	BLIP-2 (ViT-L)	60.4	20.4	21.4	8.1	36.7	21.3	14.0	32.6	28.8	44.3	28.8
		BLIP-2 (ViT-g)	62.2	23.8	22.0	15.9	57.8	33.4	23.4	36.4	57.2	55.4	38.7
		LLaVA 1.5 (7B)	76.2	30.6	38.9	3.0	5.8	22.7	7.7	27.5	2.6	48.0	26.3

TABLE 13

Cluster Accuracy on the ten datasets. Green is our method, gray shows the upper bound. **Bold** represents best, underline indicates best considering also image captioning and VQA models.
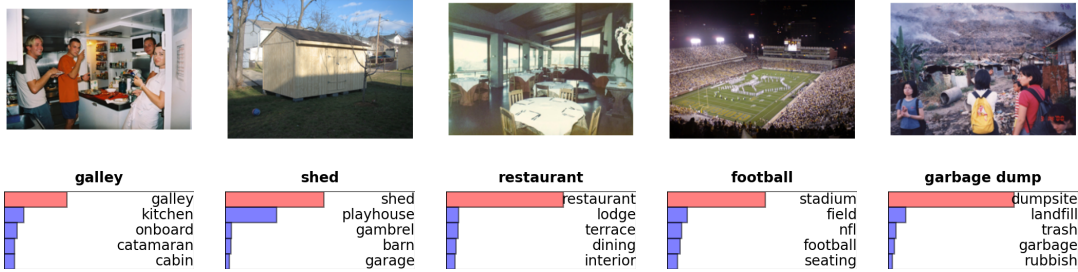

Fig. 7. Qualitative results of CaSED on SUN397. The first three samples represent success cases, the last two shows failure cases.

where the cat is separated into "cat", "ear", and "whisker", or in the second example of Fig. 10, where the woman is segmented as "hair" and "shirt". Note, however, that DenseCaSED is the only method to correctly classify and segment subtle details like the "cup"s in the left part of the second example of Fig. 10, or the horse and grass in the first example of the same figure. Despite the coarse segmentation masks, our model provides more contextualised labels, e.g., "plane", "ceiling", "airline", "aisle", and "aircraft" for the airplane aisle in the second example of Fig. 8 (SAM and SAN selects labels such as "chair", "air travel", and "hair"), or "cat", "whisker", and "ear", in the first example of Fig. 9 (SAN selects "tabby", "gitana", "Freya", and "soporific" as labels).

To conclude, all methods show their peculiarities, with SAM presenting the cleanest segmentation masks but sometimes undersegmenting objects, SAN having the most efficient implementation but showing sub-optimal performance with vocabularies out-of-distribution (i.e., outside of the common words used for semantic segmentation, as also shown by its weakness against "distractor" words), and DenseCaSED recognising also smaller objects but resulting in the most coarse segmentation masks. These results confirm the challenges of VSS, and the need of developing methods tailored to this task.

(a) SAM + LLaVA 1.5 (7B)
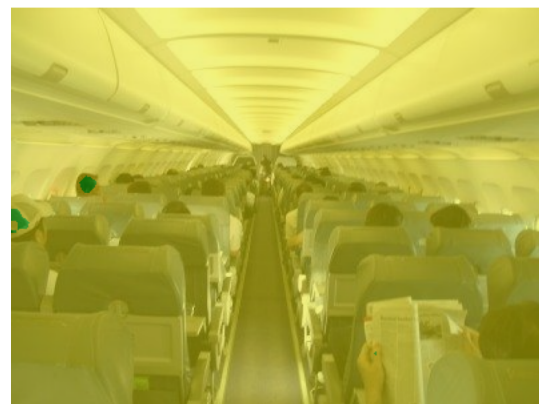
(b) SAN + WordNet

(c) DenseCaSED

Fig. 8. Qualitative results of Vocabulary-free Semantic Segmentation methods on ADE20K-150. For visualisation purposes, we visualise only the top-6 most prominent regions.
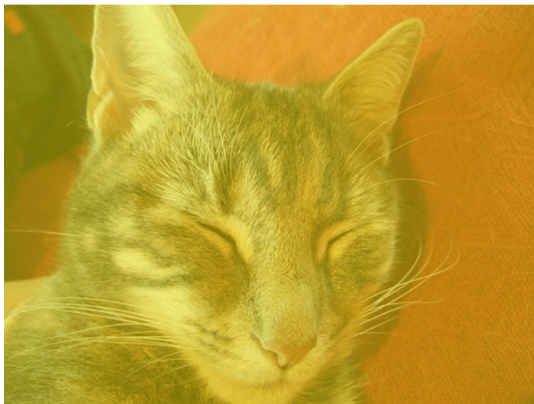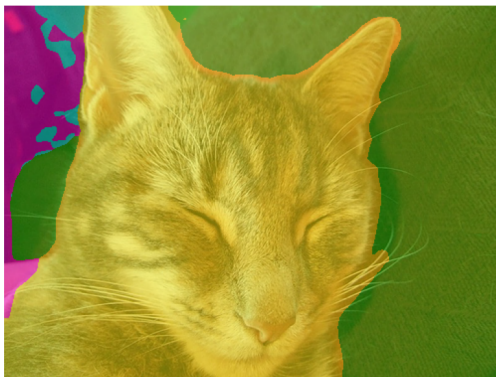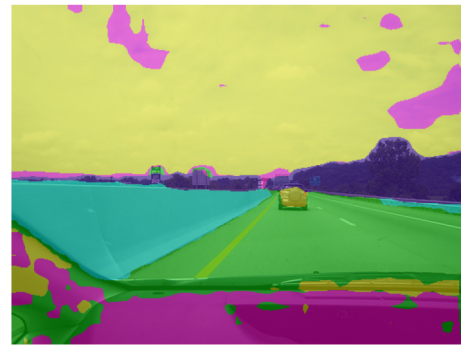
(a) SAM + LLaVA 1.5 (7B)

(b) SAN + WordNet
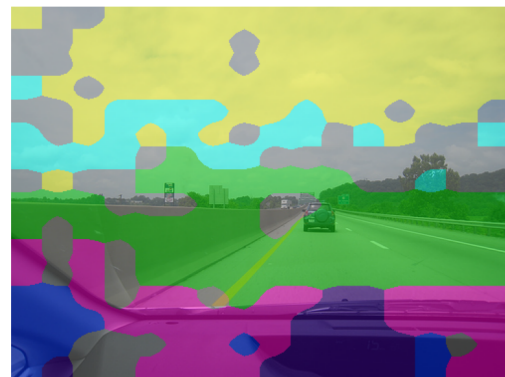
(c) DenseCaSED

Fig. 9. Qualitative results of Vocabulary-free Semantic Segmentation methods on PASCAL Context-59. For visualisation purposes, we visualise only the top-6 most prominent regions.

(a) SAM + LLaVA 1.5 (7B)

Legend (left): slide · hair · plant · ladder

Legend (right): bacardi · car · bottle · woman · shirt · picture

(b) SAN + WordNet

Legend (left): neigh · antrorse · cheval-de-frise

Legend (right): Erin · lauryl alcohol · bottlefeed · alcoholize · bottle-shaped

(c) DenseCaSED

Legend (left): horse · cropped · deer · grass · fence · green
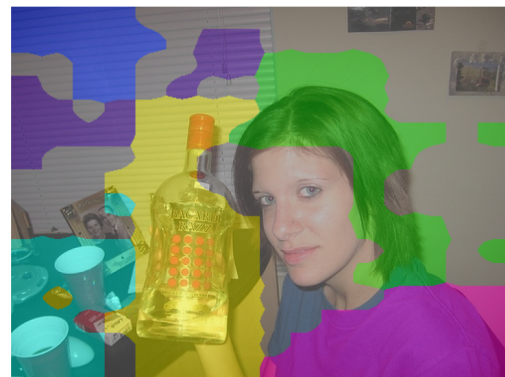
Legend (right): bottle · shirt · window · hair · cup · paper
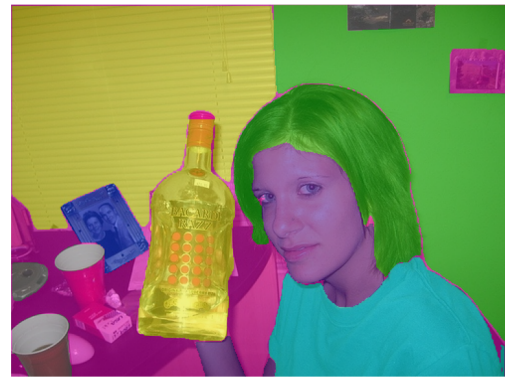
Fig. 10. Qualitative results of Vocabulary-free Semantic Segmentation methods on PascalVOC-20. For visualisation purposes, we visualise only the top-6 most prominent regions.

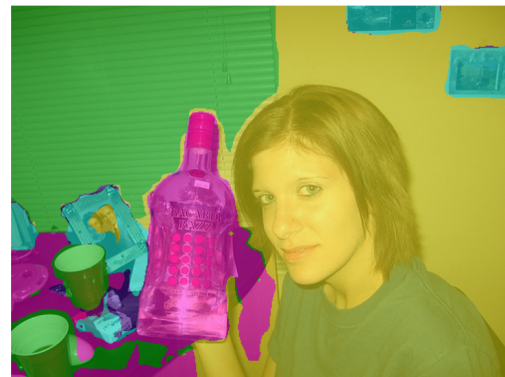|	Method		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	Avg.
			Semantic IoU (%) ↑										
CLIP RN50	CLIP	WordNet	9.8	1.9	**0.9**	0.0	21.2	5.5	14.1	7.3	3.5	2.6	6.7
		English Words	5.1	0.8	0.0	0.1	12.4	1.8	13.2	7.8	2.5	1.3	4.5
	Caption	Closest Caption	3.4	0.5	0.1	1.5	6.6	2.2	2.2	2.1	9.1	0.7	2.8
		CaSED	31.1	2.9	0.1	__2.8__	29.7	15.1	27.6	15.0	**13.4**	__5.2__	14.3
		UpperCaSED	**33.0**	__3.7__	0.4	2.7	**30.4**	**15.2**	**29.7**	**15.7**	12.9	4.7	**14.8**
		CLIP upper bound	81.5	41.4	33.9	16.5	60.2	74.6	78.9	56.9	66.5	57.1	56.8
CLIP ViT-L/14	CLIP	WordNet	15.0	3.0	1.3	0.5	31.3	7.8	14.7	9.0	4.8	3.8	9.1
		English Words	8.0	2.0	0.0	1.1	16.4	2.0	17.2	8.1	2.7	1.8	5.9
	Caption	Closest Caption	4.5	0.8	1.3	1.9	5.9	3.1	3.0	2.3	11.4	1.0	3.5
		CaSED	35.4	__5.1__	2.3	__4.8__	33.1	__19.4__	__35.1__	17.2	**16.2**	__8.4__	17.7
		UpperCaSED	**37.8**	4.6	**5.2**	4.4	**35.2**	18.9	34.9	__17.8__	16.0	7.8	__18.3__
		CLIP upper bound	86.0	52.2	51.5	28.6	75.7	89.9	88.0	66.6	84.5	71.3	69.4
SigLIP ViT-L/16	SigLIP	WordNet	18.0	__5.7__	1.3	1.0	29.1	12.7	19.6	11.0	9.5	3.0	11.1
		English Words	9.3	3.1	0.0	1.3	12.3	1.7	16.8	8.1	6.0	0.4	5.9
	Caption	Closest Caption	5.0	0.5	0.5	2.1	5.9	3.1	3.5	2.6	13.8	0.8	3.8
		CaSED	__43.0__	5.1	5.1	__6.0__	__40.2__	__19.0__	__36.2__	18.5	17.7	__7.9__	__19.9__
		UpperCaSED	41.5	5.1	**5.8**	5.5	39.0	17.6	33.7	__19.2__	17.8	5.9	19.1
		CLIP upper bound	93.3	60.0	47.1	47.3	88.6	94.2	95.4	70.8	95.7	81.9	77.7
	Caption	BLIP-2 (ViT-L)	13.4	1.4	4.8	0.0	7.5	4.7	1.7	4.7	11.6	1.1	5.1
		BLIP-2 (ViT-g)	16.8	1.8	4.1	0.1	13.9	7.9	2.9	5.7	24.7	1.9	8.0
		LLaVA 1.5 (7B)	13.8	0.9	5.1	0.0	2.9	1.7	0.3	3.5	4.6	1.4	3.4
	VQA	BLIP-2 (ViT-L)	36.1	1.8	7.0	0.1	21.5	3.7	5.7	11.5	18.9	2.5	10.9
		BLIP-2 (ViT-g)	41.5	2.4	7.5	2.0	38.0	8.6	10.2	13.8	33.2	2.8	16.0
		LLaVA 1.5 (7B)	41.4	0.4	6.1	0.0	5.0	3.0	0.7	6.6	1.3	2.0	6.6

TABLE 14

Semantic IoU on the ten datasets. Green is our method, gray shows the upper bound. **Bold** represents best, <u>underline</u> indicates best considering also image captioning and VQA models.

	Method		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	Avg.
			Semantic Similarity (x100) ↑										
CLIP RN50	CLIP	WordNet	43.2	29.0	18.5	21.6	46.7	44.6	50.3	42.8	26.4	40.0	36.3
		English Words	36.0	29.5	14.9	20.0	38.1	34.2	40.7	35.4	18.3	32.4	29.9
	Caption	Closest Caption	37.2	22.8	14.2	26.8	38.9	41.2	32.6	37.4	**44.3**	32.4	32.8
	CaSED		**62.3**	36.4	22.6	**28.7**	**52.8**	<u>59.0</u>	57.0	**50.2**	42.9	**46.2**	**45.8**
	UpperCaSED		**62.3**	<u>37.8</u>	23.6	26.0	52.4	58.3	<u>57.9</u>	50.2	41.6	46.1	45.6
	CLIP upper bound		88.1	62.4	52.8	53.9	72.0	83.7	85.8	73.9	81.0	73.9	72.7
CLIP ViT-L/14	CLIP	WordNet	48.6	32.7	24.4	18.9	55.9	49.6	53.7	44.9	28.8	44.2	40.2
		English Words	39.3	31.6	19.1	18.6	43.4	38.0	44.2	36.0	19.9	34.7	32.5
	Caption	Closest Caption	42.1	23.9	23.4	29.2	40.0	46.9	40.2	39.8	**49.2**	40.3	37.5
	CaSED		65.7	40.0	32.0	**30.3**	55.5	64.5	62.5	52.5	47.4	**54.1**	50.4
	UpperCaSED		**66.3**	<u>40.3</u>	**34.9**	27.0	**56.1**	<u>65.0</u>	**62.9**	**53.0**	46.5	53.8	<u>50.6</u>
	CLIP upper bound		90.8	69.8	67.7	66.7	83.4	93.7	91.8	80.5	92.3	83.3	82.0
SigLIP ViT-L/16	SigLIP	WordNet	50.7	40.2	19.1	22.0	53.3	56.2	60.2	46.7	33.7	43.2	42.5
		English Words	40.0	36.2	18.8	18.6	40.1	40.7	49.2	36.1	28.1	34.1	34.2
	Caption	Closest Caption	45.0	24.1	17.2	30.6	41.1	49.1	41.3	41.2	**53.3**	40.1	38.3
	CaSED		**71.3**	<u>41.6</u>	28.0	<u>34.4</u>	<u>61.7</u>	<u>66.2</u>	66.0	54.2	48.4	<u>55.1</u>	<u>52.7</u>
	UpperCaSED		70.7	41.0	**33.6**	27.7	58.2	61.7	63.9	<u>54.3</u>	48.5	54.2	51.4
	CLIP upper bound		97.8	75.6	63.1	80.0	92.1	96.4	96.8	83.2	98.1	89.6	87.3
	Caption	BLIP-2 (ViT-L)	57.8	31.4	39.9	24.4	36.1	44.6	29.0	45.3	46.4	38.0	39.3
		BLIP-2 (ViT-g)	63.0	33.1	36.2	24.3	45.2	51.6	31.6	48.3	61.0	44.6	43.9
		LLaVA 1.5 (7B)	56.8	29.4	40.5	21.3	31.1	36.9	24.8	42.5	38.1	37.9	35.9
	VQA	BLIP-2 (ViT-L)	70.5	34.9	29.7	29.1	48.8	42.0	40.0	50.6	52.4	48.6	44.7
		BLIP-2 (ViT-g)	73.5	36.5	31.4	30.8	59.9	52.1	43.9	53.3	65.1	55.1	50.1
		LLaVA 1.5 (7B)	72.6	36.7	44.1	29.4	41.8	41.1	36.0	41.9	35.3	46.6	42.6

TABLE 15

Semantic similarity on the ten datasets. Green is ours, gray shows the upper bound. **Bold** represents best, <u>underline</u> indicates best considering also image captioning and VQA models.