CFM: Language-aligned Concept Foundation Model for Vision

Kai Wittenmayer, Sukrut Rao†, Amin Parchami-Araghi†, Bernt Schiele, and Jonas Fischer

Max Planck Institute for Informatics, Saarland Informatics Campus
Saarbrücken, Germany
{kai.wittenmayer,sukrut.rao,amin.parchami}@mpi-inf.mpg.de
{schiele,jonas.fischer}@mpi-inf.mpg.de
† Equal contribution

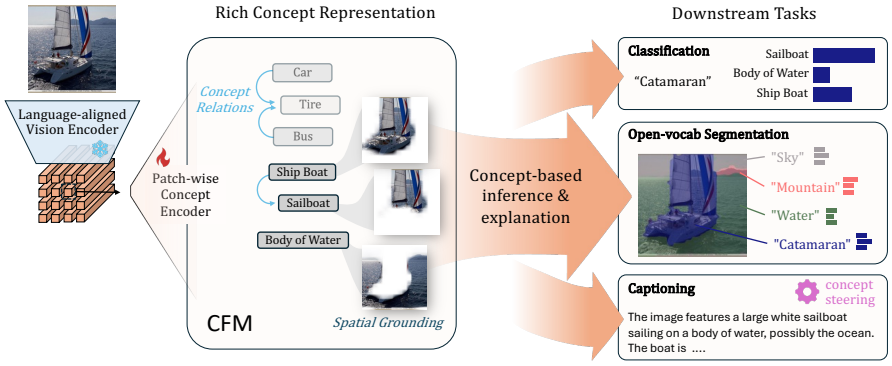

Fig. 1: CFM provides rich conceptual explanations for vision foundation model tasks. Our model provides a **hierarchical concept representation** (*e.g.* 'Sailboat' concept is a child of 'Ship Boat') with **local, spatially grounded** concepts that are automatically **named**. These concepts enable CFM to provide concept-based explanations for decision-making across vision tasks.

Abstract. Language-aligned vision foundation models perform strongly across diverse downstream tasks. Yet, their learned representations remain opaque, making interpreting their decision-making difficult. Recent work decompose these representations into human-interpretable concepts, but provide poor spatial grounding and are limited to image classification tasks. In this work, we propose CFM, a *language-aligned concept foundation model for vision* that provides fine-grained concepts, which are human-interpretable and spatially grounded in the input image. When paired with a foundation model with strong semantic representations, we get explanations for *any of its downstream tasks*. Examining local co-occurrence dependencies of concepts allows us to define concept relationships through which we improve concept naming and obtain richer explanations. On benchmark data, we show that CFM provides performance on classification, segmentation, and captioning that is competitive with opaque foundation models while providing fine-grained, high quality concept-based explanations. Code at https://github.com/kawi19/CFM.

1 Introduction

Language-aligned vision foundation models such as CLIP [49,61,72] serve as powerful feature extractors that perform strongly across tasks such as zero shot classification and open-vocabulary semantic segmentation. Such models also serve as vision encoders for large vision-language models [12, 31, 34] and generative models [73]. However, their representations are opaque, which hinders interpreting their decision-making. Interpretability is, however, key for human users, especially for safety critical applications, and explanations of decision-making can help to understand whether a model is right for the right reasons, increase trust in a model's prediction, investigate why a model failed, and steer it away from harmful outputs.

Concept Bottleneck Models have emerged as a widely successful approach to make model predictions more interpretable, with recent works [1,2,32,53,71] aiming to decompose representations of foundation models into human-interpretable concepts and achieve transparency at an unprecedented scale. However, extracted concepts are not often properly grounded in the input image [2, 53, 71], harming interpretability. Moreover, the existing approaches only provide a limited range of granularity of concepts [1, 2, 32, 53] and their spatial coverage [2, 53, 71], and inter-concept dependencies are not examined even when using hierarchical priors [71]. Textual concept labels, when assigned [2, 53], often stem from a limited vocabulary that is not tailored for large-scale and fine-grained conceptual diversity.

Besides these limitations of what concepts can capture, such models have not been used in the foundational setting, i.e. such concept representations have only been used for global tasks, typically classification. The lack of granular, well-grounded, and non-spurious concepts hinders application to local downstream tasks, such as segmentation or large vision-language model reasoning, whereas the underlying foundation model backbone would be capable of this.

In this work, we address these issues and propose CFM, a *language-aligned concept foundation model* that provides a unified concept representation applicable to local image regions, thus *enabling any downstream vision foundation model task previously not possible with concept-based explanations.* The discovered concepts are at varying granularities, spatially grounded in the image, and organized into hierarchies with explicit inter-concept dependencies. To achieve this we employ a semantically smoothened image encoding through DINOised CLIP features [64] and learn a sparse autoencoder with a matryoshka objective encouraging hierarchical structures of the concept representations [5, 71]. By learning the concept representation across local image regions (e.g., patches) we achieve spatially localized and human-aligned concepts from coarse to fine granularities and avoid spurious concepts common in the existing global concept bottlenecks. By examining co-occurrences of local concept activations we discover parent-child relationships between concepts that provide further insights into the learned semantics. We use these relations to refine concept labeling towards more accurate and meaningful textual concept labels. Our model enables interpretable decision-making across foundation model downstream tasks typi-

cal for language-aligned vision foundation models (see Fig. 1). Our contributions can be summarized as follows:

1. We propose a **concept foundation model** (CFM), lifting concept bottleneck models to language-aligned vision foundation models **applicable to any downstream task** beyond just image classification through unified local concept representations.
2. We show how to extract **explicit hierarchical relations** between foundational concepts that further improve the explanations of model decisions.
3. We propose an **improved concept labeling scheme** that explicitly takes these concept relationships into account.
4. Putting this together, CFM yields **human-interpretable concepts** at **diverse granularities** that are **spatially localized and well-grounded** in the input.

On benchmark data and through a human user study we show that in contrast to state-of-the-art, CFM yields concepts that are more quantitatively and qualitatively consistent, spatially more localized (Fig. 3), well organized into hierarchies (Fig. 4), and meaningfully named (Fig. 5). We further show that this added interpretability does not come with a degradation of performance across various downstream tasks. On image classification tasks CFM yields performances competitive with the state-of the-art concept-based and opaque baseline models (Fig. 6). On an image captioning task, CFM performs similar to its opaque backbone and enables steering (Fig. 7). CFM also enables interpretable open-vocabulary segmentation that is on par with opaque foundation model alternatives (Tab. 1, Fig. 8).

2 Related Work

Vision-language models (VLMs) such as CLIP [13, 26, 49], ALIGN [27], CoCa [69], and SigLIP [61,72] learn vision and text encoders that share a semantic latent space. Such models are typically trained on billions of paired image-text data through contrastive learning, yielding powerful but opaque language-aligned visual representations enabling tasks such as open-vocabulary classification and semantic segmentation. Our proposed method CFM is such a language-aligned vision encoder, but trained to provide a human-interpretable concept representation space.

Concept extraction methods aim to decompose learned representations from encoders into human-interpretable concepts through, e.g., clustering [22, 30, 41], non-negative matrix factorization [20] or a learned concept dictionary that is image-specific [2] or shared across images [32,46,53,71]. The latter, in particular, have been applied to VLMs, but (1) lack spatial grounding of concepts to the image [2, 53, 71], (2) learn a flat hierarchy of concepts [2, 32, 53], or (3) use a limited vocabulary and imprecise names [53]. Our CFM, in contrast, learns a rich set of spatially grounded concepts with explicit parent-child relationships

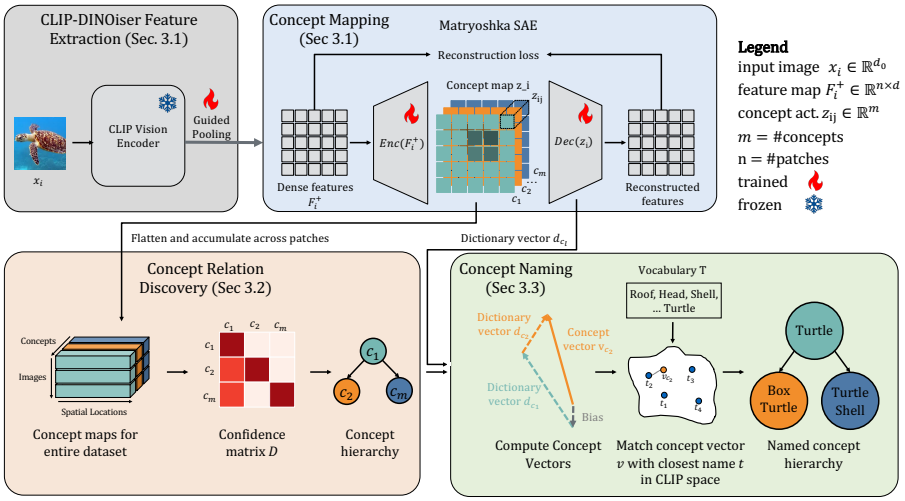

Fig. 2: Overview of CFM. We visualize our approach to represent human-interpretable, well localized, and consistent concepts of different granularities in a language-aligned vision foundation model. As backbone, we use pretrained CLIP-DINOiser [64] and extract concepts patch-wise by training a Matryoshka Sparse Autoencoder (SAE) without any supervision. From these learned concept representations, we discover concept relations through examining the co-occurrences of concepts within patches that allow more fine-grained interpretation of decision-making on downstream tasks. We further name the concepts with a Matryoshka-aware matching between concept representations and a large vocabulary of labels, taking the discovered concept relations into account.

that are named using a large fine-grained vocabulary taking such relationships into account, yielding a significantly more accurate, localized, and interpretable concept decomposition.

Sparse autoencoders (SAEs) [3,15] are models that aim to learn a dictionary of human-interpretable concepts from a latent space. These consist of an affine encoder and decoder that map features to a sparse high-dimensional space of concepts typically trained through a reconstruction and sparsity loss. SAEs have been shown a useful tool to understand large language models (LLMs) [3, 15], vision models [53, 71], and vision-language models [43]. Recent works suggested alternative sparsity objectives [4, 21, 50, 51] and architectural priors for concept hierarchies [5, 71] to improve SAE representations. Here, we combine the best of both worlds and use Matryoshka SAEs [5] with a BatchTopK objective [4] to extract high quality, localized, and hierarchical concepts.

Concept-based classification models such as concept bottleneck models (CBMs) [1,29,40,53,70] and visual prototype-based methods [10,18,39,67] learn concepts from a latent feature space to yield interpretable classification predictions. Where CBMs extract labeled, global concepts that have been shown to often be incorrectly grounded in the image [25], prototypes are local visual

features that have to be investigated by a user to understand their conceptual meaning. Our approach bridges the gap between the two and helps construct inherently interpretable concepts that are labeled, can be global as well as local, and are well grounded in the input. Moreover, in the spirit of foundation models, CFM enables downstream tasks beyond just classification.

3 CFM: A Language-aligned Concept Foundation Model for Vision

To lift concept bottleneck models to be applicable to any vision foundation model downstream task we need accurate, spatially local concept representations. Here, we propose a language-aligned vision foundation model with explicit concept representations of local image regions, such as patches or pixels, that enable transparent, human-interpretable reasoning. In a nutshell, we learn a *local* mapping from dense, semantically smoothened image representations to a human-interpretable concept space. This concept space is encouraged to represent *concepts at varying levels of granularity*, so that we are able to discover locally consistent *concept hierarchies* describing part- or specification-relationships between concepts based on patch-wise concept co-occurrences (Fig. 2).

Specifically, as backbone we leverage CLIP-DINOised image embeddings [64], a CLIP architecture that encodes images such that semantically similar regions, informed through DINO [7], share similar representations. To map this latent CLIP encoding to meaningful concepts, we train a Sparse Autoencoder (SAE) on patch embeddings to learn *local* interpretable concept representations in an unsupervised manner at scale (Sec. 3.1). Through a Matryoshka objective, we further encourage to learn concepts at different levels of granularity. From these interpretable concept representations, we show how to extract relations between concepts that reflect both specifications as well as part-relations through an analysis of local co-occurrence patterns (Sec. 3.2). Finally, we extend the existing concept vocabularies used for labeling, show how to improve concept labeling by leveraging our discovered concept relationships, and show that an architecture-aware text encoding is necessary for accurate labels (Sec. 3.3).

3.1 From Global to Spatially Localized Concept Representations

We first extract human interpretable, spatially localized, and visually grounded concepts from latent features of the CLIP [49] vision encoder. We do this using sparse autoencoders (SAEs) [3], which we discuss first, and then describe our approach.

Background: Sparse Autoencoders (SAEs). Sparse autoencoders provide an efficient and unsupervised method to learn a concept mapping, crucial at the scale of foundation models. These consist of a light-weight linear layer encoder, $\pi^{enc} : \mathbb{R}^d \to \mathbb{R}^{d'}$ that describes an affine mapping from a d-dimensional input

space of CLIP features to a sparse representation where each dimension ideally encodes a single interpretable concept. An affine mapping $\pi^{dec} : \mathbb{R}^{d'} \to \mathbb{R}^d$ then maps back to the input space. The SAE is then trained through a reconstruction loss such that $\pi^{dec}\pi^{enc}(x) \approx x, \quad x \in \mathbb{R}^d$ and a sparsity regularization term on the encoder representations, typically via an ℓ_1 [3], TopK [21], or BatchTopK [4] constraint. Recent work [71] proposed Matryoshka SAEs to learn concepts across granularities. Specifically, let $C = [1 \dots m]$ be the index set for the m neurons in the SAE concept layer and $l_j \in C$ indices until which we define a matryoshka shell. Then

$$\mathcal{L}_{rec}(F_p) = \sum_j \left\| \pi^{dec}(\ \pi^{enc}(F_p)[1 : l_j]\ ) - F_p \right\|_2^2, \tag{1}$$

where F_p are the latent features for patch p given by the backbone model and $\pi^{enc}(F_p)[1 : l_j]$ sets all activations of neurons after index l_i to zero. Thus, each Matryoshka shell consisting of the first l_j neurons is trained to reconstruct the *full* latent features, which encourages that the smallest subset of neurons learns coarse, high-level information and higher-index neurons learn the *residual* fine-grained information on top of this coarse representation.

Extracting Semantic, Spatially Localized, and Grounded Concepts. While useful for learning concepts across granularities, concepts from such an SAE are not spatially localized, visually grounded to the input, and may not map well to human interpretable semantic features. To address this, we propose to use the following two approaches (Fig. 2, top).

Local Concept Extraction. Instead of encoding pooled visual features, we encode *each patch token* from the final layer of the CLIP vision encoder through the Matryoshka SAE, which helps encode spatially localized concepts. This also allows visual grounding, since concept strengths across spatial regions in the image can now be viewed as a heat map. We also use BatchTopK [4] instead of TopK for sparsity regularization as they provide more flexibility across samples.

DINOised CLIP Features. It has been shown that [77] fine-grained local representations can be extracted from CLIP by repurposing the patch token value vectors from the CLS attention as dense local embeddings. However, despite preserving a degree of spatial locality, they are noisy, which hurts concept extraction. To address this, we adopt the guided pooling approach introduced by CLIP-DINOiser [63], which smoothens the representations of semantically similar patches using affinity patterns of self-supervised features such as those from DINO [7]. Intuitively, this pooling acts as a "voting system" that enforces consistency among semantically similar patches while "attenuating noisy features" [63]. For efficient inference via a single forward pass, a small 3×3 convolutional layer is trained on patch tokens to approximate the DINO affinity map. In our case, we DINOise the final layer patch tokens from CLIP before using them as inputs to the Matryoshka SAE; we refer to App. B for more details.

As compared to using pooled CLIP features, our approach helps obtain localized and better human-aligned concepts, since these DINOised local embeddings

Methods	Loc. ↑ Part	Loc. ↑ Stuff	Cons. ↑ Part	Cons. ↑ Stuff	Impur. ↓ Part	Impur. ↓ Stuff	C^2-Sc. ↑ IMN
SALF-CBM [1]	16.4	19.5	9.9	6.7	0.841	1.143	0.298
PatchSAE [32]	32.2	34.6	13.1	12.7	0.353	0.663	0.295
CFM	**44.3**	**44.5**	**15.3**	**13.1**	**0.333**	**0.539**	**0.465**

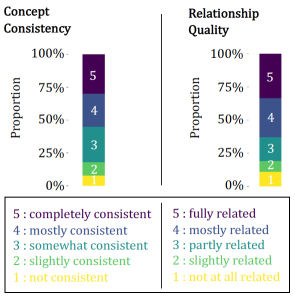

Fig. 3: CFM discovers consistent, well-grounded, and well-named concepts. ***Left:*** We quantitatively evaluate the consistency and grounding of our concepts using annotation-based and annotation-free metrics. Specifically, we compare against state-of-the-art concept extraction methods on part-annotated PartImageNet and COCO-Stuff, using metrics from [47], and measure localization of concepts, consistency of localization across images, and concept impurity across parts. We also evaluated C^2-Score [46] over ImageNet, which measures consistency of attributions independent of human annotations. ***Right:*** User study results on concept consistency (left) and relationship accuracy (right) for ∼1000 neurons each covered with 5 users in Amazon MTurk. We find our concepts to be highly consistent and with accurate relations.

are pooled from semantically coherent regions of the image which frequently correspond to meaningful object parts or localized visual attributes. Moreover, these extracted local embeddings remain well aligned with CLIP's language space, allowing them to still be directly associated with natural language descriptions through CLIP's pretrained text encoder.

3.2 Discovering Concept Relationships

The concept extraction procedure from Sec. 3.1 provides localized concepts at diverse granularities, but does not reveal inter-concept relationships. For example, a concept could be a specialization (*e.g.* 'leatherback turtle') or part (*e.g.* 'turtle shell') of a more general parent concept (*e.g.* 'turtle'). Identifying such relationships would help interpretability and allow for more effective steering of model decisions. Despite being trained to encode concepts at multiple granularities, extracting explicit relations from Matryoshka SAE concepts has so far not been explored. Other prior works rely on predefined concept hierarchies or leverage LLMs [45, 59] to order concepts in a hierarchical manner, but these may not reflect the structure that emerges in the data or the model's reasoning process. In contrast, by discovering concept relations directly from data, we avoid mismatches between imposed taxonomies and the representations that the model actually uses.

To represent hierarchies, we identify pairs of parent and child concepts (Fig. 2, bottom left), where parents represent more general concepts and children capture more fine-grained or part-based specializations. In such a relation, the parent

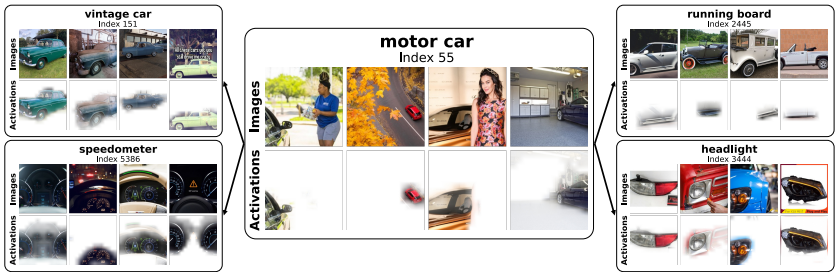

Fig. 4: A concept relationship subgraph. The parent concept, 'motor car', is shown in the center with children concepts arranged around it. For each concept, we provide the assigned label (top) and the top-4 most activating images (first row) along with their spatial localization (second row) based on concept activation strength per patch.

concept must also be active (*i.e.*, above a threshold, see App. B) whenever the specialized child concept is active. To identify such pairs, we construct a co-occurrence matrix $C \in [0,1]^{V \times V}$, such that C_{ij} denotes the weighted conditional probability of concept i being active given that concept j is active, *i.e.*

$$C_{ij} = \frac{\sum_{p:i,j \in B(p)} A_{jp}}{\sum_{p:j \in B(p)} A_{jp}} \tag{2}$$

where A_{jp} denotes the activation strength of concept j for patch p and $B(p)$ is the set of concepts active for patch p. We then use a threshold τ to prune weak associations, *i.e.*, a concept u is assigned to be a parent of concept v if $C_{uv} \geq \tau$, and construct our hierarchy.

For this to work, our design with *local, well-grounded* concepts is crucial as otherwise spatially unrelated concepts (e.g., *field* and *cow*) may incorrectly be assigned a relationship merely due to spurious correlations in their image-level activations. The discovered relations refine explanations through inter-concept dependencies and serve as a prior that allows to learn better concept labels discussed next.

3.3 Hierarchy-aware Concept Labeling

For human-understandable textual explanations of decision-making, prior work [53] assign labels for each concept neuron in an automated manner. They leverage the language-alignment of CLIP to assign labels l_i to a concept i, by finding the label v from a pre-defined vocabulary V whose CLIP text embeddings $\mathcal{T}(v)$ have the highest cosine similarity to the concept dictionary vector, given by the SAE decoder weights π_i^{dec}, *i.e.*

$$l_i = \text{argmax}_{v \in \mathcal{V}} \cos\left(\pi_i^{dec} + b, \mathcal{T}(v)\right), \tag{3}$$

where b is the SAE decoder bias. While this performs well for general concepts in the first Matryoshka group of the SAE, it fails for subsequent groups with fine-grained concepts since it does not take into account inter-concept dependencies

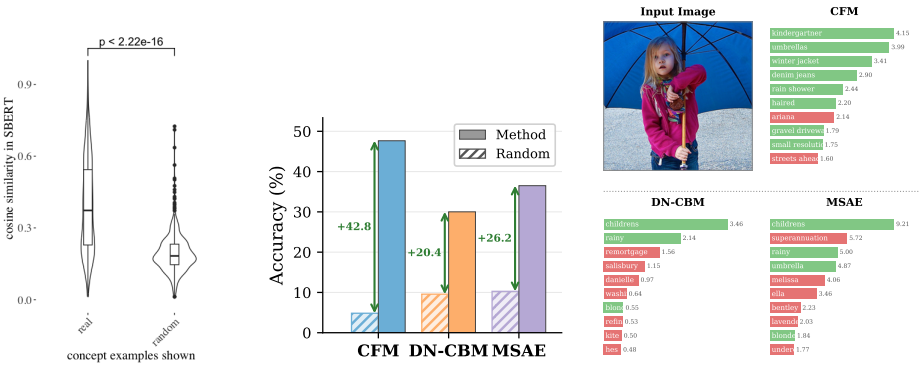

Fig. 5: CFM assigns concept names accurately. ***Left:*** We ask human annotators to provide a 1-2 word label for each concept, and compute cosine similarities with the concept name assigned by CFM using SBERT [54], and find a significant improvement over a random baseline. ***Middle:*** We use gpt-5-mini [58] as a judge to evaluate whether the top-10 concepts predicted by CFM and baselines are present in the image. CFM provides the most accurate concept names and has the biggest improvement over the random baseline. ***Right:*** A qualitative example of concept name accuracy. Green (red) indicate concepts present (absent) in the image as per the judge.

(see App. B). To account for this, we propose a hierarchy-aware naming scheme (Fig. 2, bottom right) that offsets child concept dictionary vectors with those of their parents, *i.e.*

$$l_j = \text{argmax}_{v \in \mathcal{V}} \cos\left(\left(\sum_{(i \to j) \in G} \pi_i^{dec}\right) + \pi_j^{dec} + b, \mathcal{T}(v)\right), \tag{4}$$

where $(i \to j)$ denotes a parent-child relationship in the extracted hierarchy G.

Putting all pieces together (Fig. 2), we obtain CFM, a concept foundation model which extracts language-aligned, spatially grounded, visually localized human-interpretable concepts across granularities for bringing interpretability to diverse downstream tasks.

4 Experimental Evaluation

We construct a CFM for a CLIP ViT-B/16 and use CC12M [9] to train the SAE with 8192 concepts and $K = 12$ concepts per patch for our main experiments. For more details including ablations on the number of concepts, see App. B. In Sec. 4.1, we evaluate the quality and interpretability of our extracted concepts, *i.e.* how localized, consistent, and pure they are via quantitative metrics and a human user study. We also evaluate accuracy of concept relationships and assigned names from CFM and compare them against baselines. Then, in Sec. 4.2, we evaluate the utility of CFM as a foundation model for diverse downstream

tasks. We first evaluate for image classification and show comparable performance to baselines, while enhancing interpretability. Next, we move to tasks that require local concept representations and show that CFM can be applied to yield meaningful image captions that are steerable through our concept layer. Finally, we use CFM for Open Vocabulary Segmentation (OVS) and show that it provides performance competitive with baselines while being the only method to provide explanations for the predicted segmentation.

4.1 Concept Quality and Interpretability

Locality, Consistency, Purity. We quantitatively evaluate concepts extracted from CFM (Sec. 3.1) for locality, consistency, and purity using metrics adapted from Pham *et al*. [47], on the part-annotated PartImagenet [23] and COCO-Stuff [6] datasets, and compare against state-of-the-art grounded concept extraction methods—SALF-CBM [1] and PatchSAE [32]. Specifically, to measure *locality*, we evaluate how well the concepts represent a human-annotated part by finding the maximal IoU between regions where the concept is active and the part annotation, across all parts. To measure *consistency*, *i.e*. how consistent parts map to the same concepts across images, we compute the maximal sum of part-specific IoUs across all parts. Lastly for the *impurity*, we evaluate how *semantically impure* a concept is by computing the per-concept entropy of normalized activation values across all parts. To avoid bias from dataset-annotated parts, we also compute the C^2-Score [46], which measures grounding consistency in DINOv2 [42] feature space, on ImageNet [16]. We find that CFM consistently outperforms (Fig. 3-left) all baselines across datasets and metrics, and in particular provides a significant improvement for locality. For more details, see App. A, and for comparisons with non-grounded baselines using post-hoc attribution methods, see App. C.

While relying on human-annotated part data for quantitative evaluation is an established benchmark in the field and provides a proxy for interpretability, ultimately we are interested in how well humans perceive and understand our discovered concepts. To do this, we conduct a large-scale user study via Amazon MTurk for concept consistency, with 1000 concepts and five annotators per task, which ask annotators to rate concepts on a five point scale. Each concept is shown via its top activating images and the regions it activates at. We show aggregated results in Fig. 3-right and provide detailed study design and results in App. A. We find that users rate concepts mostly or completely consistent for more than half of the shown concepts. More than 80% of annotators assign *at least* somewhat consistent, *i.e*. score 3 or higher. Taken together, we show that CFM **provides more interpretable and well-grounded concepts**.

Relationship Quality. We measure accuracy of our discovered hierarchy (Sec. 3.2) via a similar user study as above, where we show annotators pairs of two concepts and ask them to rate how well related they are. As shown in Fig. 3-right, we find that like with consistency, users rate concept hierarchies from CFM to be

Model	Spatial Grounding	IMN	Places
Linear Probe	-	80.2	55.1
Zero Shot	-	68.6	41.2
LF-CBM [40]	✗	75.4	50.6
LaBo [68]	✗	78.9	-
CDM [44]	✗	79.3	52.6
DCLIP [37]	✗	68.0	40.3
DN-CBM [53]	✗	**79.5**	55.1
D-CBM [48]	✗	70.5	50.9
CF-CBM [45]	●	78.5	-
SALF-CBM [1]	✓	76.3	49.4
CFM	✓	78.9	**55.4**

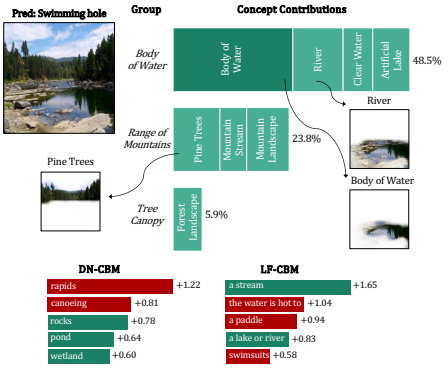

Fig. 6: CFM shows competitive classification performance while providing spatially grounded concepts. ***Left:*** We compare CFM with baselines on ImageNet (IMN) and Places365 (Places). "Spatial Grounding" indicates whether methods extract spatially grounded concepts (✓), partially local concepts (●), or global concepts only (✗). Note that 'Linear Probe' and 'Zero-Shot' refer to the standard CLIP [49] baselines. We show linear probe performance of CFM on max- and mean-pooled concept activations across patches (see App. B for specific aggregation details). Baseline results taken from Prasse *et al.* [48], we retrained CFM on the same OpenAI-CLIP backbone, best method in **bold**. ***Right:*** For an image of the "Swimming hole" Class of Places365 we show explanations of CFM against existing methods. For CFM, we provide the parent concept (dark green) and its contributing children (light green) with percentage of contribution and the spatial grounding of the top-3 most contributing concepts. In contrast to our work, existing methods (bottom) show spurious concepts.

'mostly' or 'fully' related for most pairs. Additionally, we evaluate the accuracy of the directionality of the relationship via a gpt-5.5 model judge, by providing top-5 images and their corresponding activating regions for parent-child pairs and asking the model to evaluate relationship direction. For consistent and related concepts, we obtain an accuracy of 79.79% for pairs found by CFM. For more details, see App. A.

We also show a qualitative example of discovered concept relations in Fig. 4, which shows visually well-grounded and meaningfully related concepts of varying granularities, and provide more examples in App. C. Overall, we show that **CFM organizes concepts into meaningful hierarchies** aiding interpretability.

Naming Accuracy. We perform two evaluations for the accuracy of concept names assigned by CFM (Sec. 3.3). First, we perform a user study, where we ask annotators to provide a 1-2 word description for a concept given top activating images with grounding. We then measure the cosine similarity between user-provided names and names assigned by CFM using SentenceBERT [54], and observe a strong statistically significant ($p < 2.22 \times 10^{-16}$ Wilcoxon rank-sum test) difference (Fig. 5-left) between our assigned names versus randomly drawn names using the set of concept names discovered by our method.

Metrics ↑	CLIP-DINO	CFM
CLIP-Sc [24]	0.72	0.72
RefCLIP-Sc [24]	0.73	0.72
Capture [17]	0.35	0.35

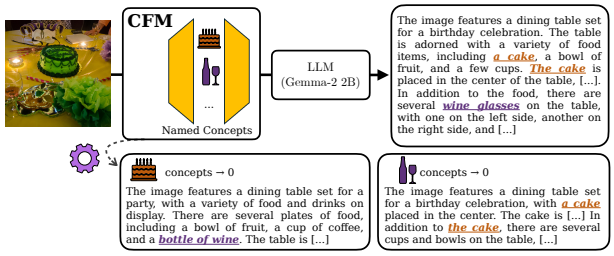

Fig. 7: CFM yields competitive and steerable captioning. ***Left:*** Captioning capabilities of CFM compared to its opaque backbone. ***Right:*** Through named concepts, CFM allows for concept-level intervention, steering the image captioning. For example, the 'cake' concept can be removed from the caption while preserving the 'drink' concept, and vice-versa. See also Apps. B and C for further details and examples.

To evaluate its utility as a foundation model that can provide concepts, we also evaluate the accuracy of concept names found by CFM against baselines for images from the COCO [33] validation set. We use gpt-5-mini [58] as a judge, and ask it to decide whether each concept name from the top-10 concepts reported by each of the methods is present in the image. We find that CFM provides far more accurate concepts (Fig. 5-middle) as compared to the random baseline of randomly sampling concepts from the concept sets. This is also reflected qualitatively in Fig. 5-right. Overall, we show that **CFM assigns names that meaningfully align with human labeling** and more accurately describe what is in an image as compared to prior work.

4.2 Foundation Model Applications

Image Classification. We evaluate CFM for image classification on ImageNet [56] and Places365 [75] by training a linear classification head for each. We compare with recent CLIP ViT-B/16 based concept bottleneck models, *i.e.* LF-CBM [40], LaBo [68], CDM [44], DCLIP [37], DN-CBM [53], D-CBM [48], and CF-CBM [45]. We additionally include the ResNet-50 based SALF-CBM [1]. All baselines except SALF-CBM and CFM only provide image-level concepts. We find that **CFM yields competitive classification accuracies (Fig. 6-left) while also providing more interpretable and spatially grounded concepts (Fig. 6-right)**. Notably, all the baselines compared against are trained explicitly and only for classification, while CFM can do much more, which highlights the versatility concept foundation model.

Image Captioning. Next, we go beyond classification and show the utility of CFM's interpretable concept-based representation for image captioning. We follow the LLaVA setup [35], by learning a projection for spatial output tokens from CFM to a language model through an MLP adapter and fine-tune both the adapter and LLM (Gemma-2-2B Instruct [60]) captioner on CC12M DreamLIP

Table 1: CFM provides strong open-vocabulary semantic segmentation. We report mIoU between segmentations and annotated parts for each dataset considering evaluation with and without a 'background' prompt as discussed in App. A. We took baseline results from Wysoczańska *et al.* [64] and rerun (*) for inference consistency. First and second best method are **bold** and underlined, respectively.

Methods$^{+(\text{Variants/Extra Backbones})}$	No background prompt					w/ background			
	VOC20	C59	Stuff	City	ADE	Context	Object	VOC	Avg.
Build prototypes per class									
ReCo [57]	57.8	22.3	14.8	21.1	11.2	19.9	15.7	25.1	23.5
OVDiff [28]$^{+(\text{DINO \& SD})}$	81.7	33.7	-	-	14.9	30.1	34.8	67.1	-
*Text/image alignment **training** with captions*									
GroupViT [65]	79.7	23.4	15.3	11.1	9.2	18.7	27.5	50.4	29.4
ZeroSeg [11]	-	-	-	-	-	21.8	22.1	42.9	-
SegCLIP [36]	-	-	-	11.0	8.7	24.7	26.5	52.6	-
TCL [8]	77.5	30.3	19.6	23.1	14.9	24.3	30.4	51.2	33.9
CLIPpy [52]	-	-	-	-	13.5	-	32.0	52.2	-
OVSegmentor [66]	-	-	-	-	5.6	20.4	25.1	53.8	-
*Use **Frozen** CLIP*									
CLIP-DIY [63]$^{+(\text{DINO})}$	79.7	19.8	13.3	11.6	9.9	19.7	31.0	59.9	30.6
MaskCLIP [77]	61.8	25.6	17.6	25.0	14.3	22.9	16.4	32.9	27.1
MaskCLIP [77]$^{+(\text{ref}) [8]}$	71.9	27.4	18.6	23.0	14.9	24.0	21.6	41.3	30.3
CLIP-DINOiser* [64]	80.8	36.0	24.9	**39.4**	20.5	32.5	34.7	62.2	41.4
CLIP-DINOiser* [64]$^{+(\text{AnyUp}) [62]}$	81.6	37.2	**25.4**	39.1	20.9	33.8	**35.2**	**64.0**	**42.2**
*Use **Frozen** CLIP with **Interpretability***									
CFM	80.7	36.5	24.2	38.5	20.7	33.1	34.7	62.2	41.3
CFM$^{+(\text{AnyUp}) [62]}$	**81.8**	**37.6**	24.4	38.1	**21.1**	**34.3**	33.9	63.6	41.9

Long captions [9, 74]. In Fig. 7-left we observe that CFM maintains captioning performance similar to the opaque, dense representation of the backbone, while providing an explicit concept representation. Importantly, **this representation can then be used to steer the captioning model**, as shown in Fig. 7-right.

Open-Vocabulary Segmentation (OVS). Finally, we use CFM for the challenging fine-grained task of semantic segmentation. Its language alignment allows for open-vocabulary segmentation (OVS), *i.e.* can be used as is without expensive pixel-level data annotation limited to a fixed set of classes. To do this, we compute the cosine similarity between the embedding of each word in the segmentation vocabulary and the reconstructed *language aligned embedding* of the SAE. We consider two ways to obtain dense predictions. In one approach, we reconstruct patch-wise CLIP embeddings with the SAE and compute similarities to the vocabulary, after which predictions are upsampled to the pixel level using bilinear interpolation. Alternatively, we upsample the spatial concept activations directly to the pixel level using AnyUp [62] and reconstruct the CLIP embeddings with the SAE directly at pixel resolution before computing cosine similarities. In both cases, predictions are interpretable: the reconstructed CLIP embeddings are linear combinations of concepts through the SAE decoder, allowing either patch or pixel-level segmentation decision to be explained by a linear combination of concepts.

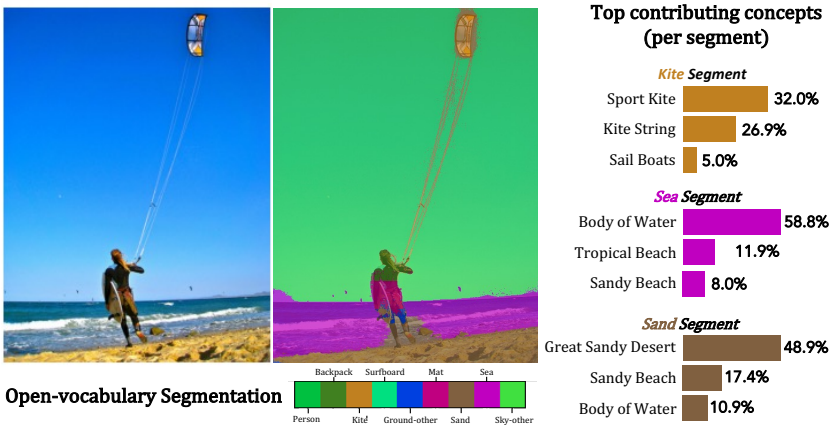

Fig. 8: Explainable Open-Vocabulary Segmentation. $\text{CFM}^{+(\text{AnyUp})}$ pixel-level open-vocabulary segmentation (left) with concept-based explanations for each segment (right).

We compare against the reported performance of prototype-based [28, 57], segmentation-specific [8,36,52,55,65,66], as well as MaskCLIP and CLIP-DINOiser versions [64, 77] on a variety of benchmark datasets, including PASCAL VOC and VOC20 [19], PASCAL Context and Context59 [38], COCO Object [33] and Stuff [6]), Cityscapes [14], and ADE20K [76]. We report mIoU in Tab. 1 with and without a background prompt following established benchmarking protocols [64].

We observe that **CFM performs on par with the strongest opaque baselines**, *i.e.* CLIP-DINOiser and OVDiff. It also outperforms existing specialized architectures by a wide margin across datasets. In Fig. 8 we show a qualitative example of OVS on COCO-Stuff highlighting the interpretable downstream decision-making provided by CFM.

Table 2: Dictionary Size Ablation. We evaluate CFM across different dictionary capacities (4k, 8k, 16k) on ImageNet (IMN), Places365 (Places), average Open-Vocabulary Segmentation (OVS Avg), and interpretability metrics. We select 8k as our default to balance segmentation performance and computational efficiency. Best results are **bold**, second best are underlined. Note that the classification results in this table are with the OpenCLIP CLIP ViT-B/16 backbone, in contrast results in Fig. 6 are done with the OpenAI CLIP ViT-B/16 backbone for clear comparability with existing CBMs.

Dict. Size	Class. ↑		OVS ↑	Loc. ↑		Cons. ↑		Imp. ↓		C^2 ↑
	IMN	Places	Avg	Part	Stuff	Part	Stuff	Part	Stuff	Score
4k	78.2	55.5	41.1	43.9	44.2	**15.8**	12.5	.340	**.540**	**0.488**
8k (Default)	**78.6**	55.6	**41.3**	44.3	44.5	15.3	13.1	**.330**	**.540**	0.465
16k	**78.6**	**56.0**	40.8	**44.6**	**44.7**	15.7	**13.6**	.340	**.540**	0.418

4.3 Ablation on Concept Set Size

To validate our hyperparameter choices, we ablate the dictionary capacity of CFM in Tab. 2. We evaluate dictionary sizes of 4k, 8k, and 16k on classification, average performance across datasets for open-vocabulary segmentation, and interpretability metrics. We observe that both downstream performance and concept grounding are highly stable across varying capacities. However, a trade-off emerges: classification accuracy increases with dictionary size, whereas the C^2 score decreases. We select 8k as our default architecture, as it achieves the highest average segmentation performance while maintaining strong classification accuracy and interpretability.

Comprehensive structural ablations—including scaling to the 400M parameter SigLIP-2 backbone, the necessity of the guided pooling mechanism, varying sparsity factors (k), and comparisons against alternative SAE objectives—are detailed in App. B.

5 Discussion and Conclusion

We proposed CFM, a concept-based representation for a language-aligned vision encoder that represents spatially grounded and human-interpretable concepts of differing granularity. As such, CFM sets a new standard in the field of concept-based interpretability by providing meaningful, well-grounded concepts that are well-organized into hierarchies with meaningful concept names.

This work also comes with limitations. For instance, despite being significantly improved as compared CBMs and at foundation model scale (Fig. 5), there is still room for improvement for concept naming. There is also a plethora of evaluation strategies suggested for explanations, yet most of them are not directly applicable to concept-based explanations learned in an unsupervised manner. In this work, we focused on a set of interpretability metrics that cover a range of interpretability aspects and were established recently in the literature.

Building a highly diverse concept representation space by learning on corresponding large (web-scale) datasets would be an interesting direction for future work. Another highly interesting future application is enabling interpretable vision tokens to be used with large vision-language models (LVLMs) for tasks such as visual question answering (VQA).

In summary, CFM is **the first concept-based foundation model unlocking explanations for all its downstream vision tasks**.

Acknowledgements

Sukrut Rao and Bernt Schiele were funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

References

1. Benou, I., Raviv, T.R.: Show and Tell: Visually Explainable Deep Neural Nets via Spatially-Aware Concept Bottleneck Models. In: CVPR. pp. 30063–30072 (2025)
2. Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F.P., Lakkaraju, H.: Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE). In: NeurIPS (2024)
3. Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J.E., Hume, T., Carter, S., Henighan, T., Olah, C.: Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread (2023)
4. Bussmann, B., Leask, P., Nanda, N.: BatchTopK Sparse Autoencoders. arXiv preprint arXiv:2412.06410 (2024)
5. Bussmann, B., Nabeshima, N., Karvonen, A., Nanda, N.: Learning Multi-Level Features with Matryoshka Sparse Autoencoders. In: ICML (2025)
6. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and Stuff Classes in Context. In: CVPR. pp. 1209–1218 (2018)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV (2021)
8. Cha, J., Mun, J., Roh, B.: Learning to Generate Text-Grounded Mask for Open-World Semantic Segmentation from Only Image-Text Pairs. In: CVPR. pp. 11165–11174 (2023)
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. In: CVPR. pp. 3558–3568 (2021)
10. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This Looks Like That: Deep Learning for Interpretable Image Recognition. In: NeurIPS. vol. 32 (2019)
11. Chen, J., Zhu, D., Qian, G., Ghanem, B., Yan, Z., Zhu, C., Xiao, F., Culatana, S.C., Elhoseiny, M.: Exploring Open-Vocabulary Semantic Segmentation from CLIP Vision Encoder Distillation Only. In: ICCV. pp. 699–710 (2023)
12. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: CVPR. pp. 24185–24198 (2024)
13. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible Scaling Laws for Contrastive Language-Image Learning. In: CVPR. pp. 2818–2829 (2023)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: CVPR. pp. 3213–3223 (2016)
15. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse Autoencoders Find Highly Interpretable Features in Language Models. In: ICLR (2024)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR. pp. 248–255 (2009)
17. Dong, H., Li, J., Wu, B., Wang, J., Zhang, Y., Guo, H.: Benchmarking and Improving Detail Image Caption. arXiv preprint arXiv:2405.19092 (2024)
18. Donnelly, J., Barnett, A.J., Chen, C.: Deformable ProtoPNet: An Interpretable Image Classifier using Deformable Prototypes. In: CVPR. pp. 10265–10275 (2022)
19. Everingham, M., Winn, J.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep **8**(5), 2–5 (2011)

20. Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., Serre, T.: CRAFT: Concept Recursive Activation FacTorization for Explainability. In: CVPR. pp. 2711–2721 (2023)
21. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and Evaluating Sparse Autoencoders. In: ICLR (2025)
22. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards Automatic Concept-Based Explanations. In: NeurIPS. vol. 32 (2019)
23. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: PartImageNet: A Large, High-Quality Dataset of Parts. In: ECCV. pp. 128–145 (2022)
24. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)
25. Huang, Q., Song, J., Hu, J., Zhang, H., Wang, Y., Song, M.: On the Concept Trustworthiness in Concept Bottleneck Models. In: AAAI. pp. 21161–21168 (2024)
26. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (Jul 2021)
27. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In: ICML. pp. 4904–4916 (2021)
28. Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Diffusion Models for Open-Vocabulary Segmentation. In: ECCV. pp. 299–317 (2024)
29. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept Bottleneck Models. In: ICML. pp. 5338–5348 (2020)
30. Kowal, M., Wildes, R.P., Derpanis, K.G.: Visual Concept Connectome (VCC): Open World Concept Discovery and their Interlayer Connections in Deep Models. In: CVPR. pp. 10895–10905 (2024)
31. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy Visual Task Transfer. TMLR (2025)
32. Lim, H., Choi, J., Choo, J., Schneider, S.: Sparse Autoencoders Reveal Selective Remapping of Visual Concepts During Adaptation. In: ICLR (2025)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV. pp. 740–755 (2014)
34. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: CVPR. pp. 26296–26306 (2024)
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
36. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: SegCLIP: Patch Aggregation with Learnable Centers for Open-Vocabulary Semantic Segmentation. In: ICML. pp. 23033–23044 (2023)
37. Menon, S., Vondrick, C.: Visual Classification via Description from Large Language Models. In: ICLR (2023)
38. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The Role of Context for Object Detection and Semantic Segmentation in the Wild. In: CVPR. pp. 891–898 (2014)
39. Nauta, M., Schlötterer, J., Van Keulen, M., Seifert, C.: PIP-Net: Patch-based Intuitive Prototypes for Interpretable Image Classification. In: CVPR. pp. 2744–2753 (2023)
40. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-Free Concept Bottleneck Models. In: ICLR (2023)

41. O'Mahony, L., Andrearczyk, V., Müller, H., Graziani, M.: Disentangling Neuron Representations with Concept Vectors. In: CVPRW. pp. 3769–3774 (2023)
42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. TMLR (2024)
43. Pach, M., Karthik, S., Bouniot, Q., Belongie, S., Akata, Z.: Sparse Autoencoders Learn Monosemantic Features in Vision-Language Models. In: NeurIPS (2025)
44. Panousis, K.P., Ienco, D., Marcos, D.: Sparse Linear Concept Discovery Models. In: ICCVW. pp. 2767–2771 (2023)
45. Panousis, K.P., Ienco, D., Marcos, D.: Coarse-to-Fine Concept Bottleneck Models. In: NeurIPS. vol. 37, pp. 105171–105199 (2024)
46. Parchami-Araghi, A., Rao, S., Fischer, J., Schiele, B.: FaCT: Faithful Concept Traces for Explaining Neural Network Decisions. In: NeurIPS (2025)
47. Pham, N., Jesslen, A., Schiele, B., Kortylewski, A., Fischer, J.: Interpretable 3D Neural Object Volumes for Robust Conceptual Reasoning. In: ICLR (2026)
48. Prasse, K., Knab, P., Marton, S., Bartelt, C., Keuper, M.: DCBM: Data-Efficient Visual Concept Bottleneck Models. In: ICML (2025)
49. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
50. Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., Nanda, N.: Improving Sparse Decomposition of Language Model Activations with Gated Sparse Autoencoders. In: NeurIPS (2024)
51. Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramár, J., Nanda, N.: Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. arXiv preprint arXiv:2407.14435 (2024)
52. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual Grouping in Contrastive Vision-Language Models. In: ICCV. pp. 5571–5584 (2023)
53. Rao, S., Mahajan, S., Böhle, M., Schiele, B.: Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery. In: ECCV (2024)
54. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: EMNLP (2019)
55. Rewatbowornwong, P., Chatthee, N., Chuangsuwanich, E., Suwajanakorn, S.: Zero-Guidance Segmentation Using Zero Segment Labels. In: ICCV. pp. 1162–1172 (2023)
56. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015)
57. Shin, G., Xie, W., Albanie, S.: ReCo: Retrieve and Co-segment for Zero-shot Transfer. NeurIPS **35**, 33754–33767 (2022)
58. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 System Card. arXiv preprint arXiv:2601.03267 (2025)
59. Sun, A., Yuan, Y., Ma, P., Wang, S.: Eliminating Information Leakage in Hard Concept Bottleneck Models with Supervised, Hierarchical Concept Learning. arXiv preprint arXiv:2402.05945 (2024)
60. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P.,

Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C.L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C.A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J.P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L.L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L.B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R.A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S.M., Cogan, S., Perrin, S., Arnold, S.M.R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., Andreev, A.: Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118 (2024)

61. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv preprint arXiv:2502.14786 (2025)
62. Wimmer, T., Truong, P., Rakotosaona, M.J., Oechsle, M., Tombari, F., Schiele, B., Lenssen, J.E.: AnyUp: Universal Feature Upsampling. In: ICLR (2026)
63. Wysoczańska, M., Ramamonjisoa, M., Trzciński, T., Siméoni, O.: CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free. In: WACV. pp. 1403–1413 (2024)
64. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: CLIP-DINOiser: Teaching CLIP a Few DINO Tricks for Open-Vocabulary Semantic Segmentation. In: ECCV. pp. 320–337 (2024)
65. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: GroupViT: Semantic Segmentation Emerges from Text Supervision. In: CVPR. pp. 18134–18144 (2022)
66. Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning Open-Vocabulary Semantic Segmentation Models from Natural Language Supervision. In: CVPR. pp. 2935–2944 (2023)
67. Xue, M., Huang, Q., Zhang, H., Cheng, L., Song, J., Wu, M., Song, M.: ProtoPFormer: Concentrating on Prototypical Parts in Vision Transformers for Inter-

pretable Image Recognition. In: IJCAI (2022)
68. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. In: CVPR. pp. 19187–19197 (2023)
69. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive Captioners are Image-Text Foundation Models. TMLR (2022)
70. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc Concept Bottleneck Models. In: ICLR (2023)
71. Zaigrajew, V., Baniecki, H., Biecek, P.: Interpreting CLIP with Hierarchical Sparse Autoencoders. In: ICML (2025)
72. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training. In: ICCV. pp. 11975–11986 (2023)
73. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion Transformers with Representation Autoencoders. In: ICLR (2026)
74. Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., Shen, Y.: Dreamlip: Language-image pre-training with long captions. In: ECCV (2024)
75. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million Image Database for Scene Recognition. IEEE TPAMI (2017)
76. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic Understanding of Scenes through the ADE20k Dataset. IJCV **127**(3), 302–321 (2019)
77. Zhou, C., Loy, C.C., Dai, B.: Extract Free Dense Labels from CLIP. In: ECCV. pp. 696–712 (2022)