

DEAL: Disentangle and Localize Concept-level Explanations for VLMs

Tang Li , Mengmeng Ma , and Xi Peng 

Deep-REAL Lab: <https://deep-real.github.io/>

Department of Computer & Information Science, University of Delaware
{tangli, mengma, xipeng}@udel.edu

Abstract. Large pre-trained Vision-Language Models (VLMs) have become ubiquitous foundational components of other models and downstream tasks. Although powerful, our empirical results reveal that such models might not be able to identify fine-grained concepts. Specifically, the explanations of VLMs with respect to fine-grained concepts are entangled and mislocalized. To address this issue, we propose to DisEntAngle and Localize (DEAL) the concept-level explanations for VLMs without human annotations. The key idea is encouraging the concept-level explanations to be distinct while maintaining consistency with category-level explanations. We conduct extensive experiments and ablation studies on a wide range of benchmark datasets and vision-language models. Our empirical results demonstrate that the proposed method significantly improves the concept-level explanations of the model in terms of disentanglability and localizability. Surprisingly, the improved explainability alleviates the model’s reliance on spurious correlations, which further benefits the prediction accuracy.

Keywords: eXplainable Machine Learning · Vision-Language Models

1 Introduction

Since the introduction of large-scale pre-trained Vision-Language Models (VLMs), *e.g.*, CLIP [44] and FLAVA [52], they become ubiquitous “foundations” for other models [5, 28] and downstream tasks [23, 50]. Although these models exhibit significant capabilities, our empirical results reveal that they might not be able to identify fine-grained concepts. As shown in Fig. 1, VLMs often associate related concepts to the main object without further distinctions. This would yield severe consequences in safety-critical downstream tasks. For example, in autonomous vehicles, a model that entangles shapes, colors, or contexts with specific road signs, might misinterpret or fail to recognize temporary or region-specific signs. This could lead to unsafe driving decisions, endangering passengers and other road users. A question naturally arises: *Can we develop VLMs that disentangle and localize fine-grained linguistic concepts on images?*

¹ Our source code and pretrained weights: <https://github.com/tangli-udel/DEAL>.

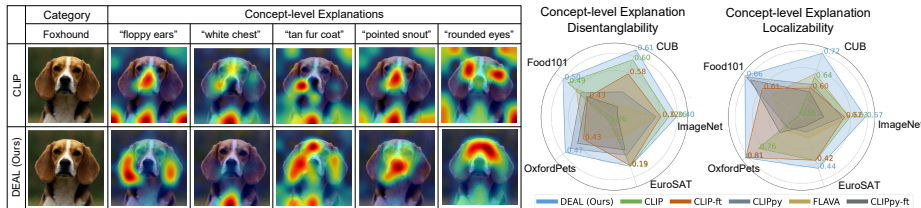


Fig. 1: We visualize the CLIP [44] model’s explanations for fine-grained concepts. The comparison is conducted using Chefer et al. [9] explanation method on the ViT-B/32 vision backbone. (Left) CLIP’s explanations *w.r.t.* visually distinct concepts might highlight the same region, or even mislocalize concepts to spurious factors, *e.g.*, background. In contrast, the concept-level explanations of our model are well-disentangled and correctly localized. (Right) Our method significantly improves the model’s concept-level explanations in terms of disentanglability and localizability compared to state-of-the-art VLMs on different benchmark datasets. This figure is best viewed in color.

This task is non-trivial due to: (i) learning objective mismatch and (ii) supervision gap. On the one hand, the learning objective of VLMs aims to align the image with the entire description, rather than distinguishing individual concepts within the description [17, 44, 53]. This entangles the model’s understanding of different concepts [33, 58, 60]. On the other hand, existing works for finer-grained alignment of VLMs rely on human annotations, *e.g.*, bounding boxes, to localize different objects [24, 61]. However, their supervision is limited to the object-level (*i.e.*, category-level) granularity. Obtaining human annotations at concept-level is prohibitively expensive, or even impossible [26, 55]. Therefore, how to train VLMs that can localize fine-grained concepts remains an open question.

In this paper, we propose a straightforward but effective method to **DisEnt**-**Angle** and **Localize** (DEAL) the concept-level explanations for solving this challenging task. The key idea is to harness the discrepancy and consistency between explanations at the concept and category levels for self-supervised learning. This will enable us to generate supervisory signals for concept learning without manual annotations. Our method works as a plug-in module, seamlessly integrating with the existing model architectures without structural changes. After training, our method provides explanations for the predictions using human-understandable concepts based on reliable visual evidence.

We evaluate the proposed method on a wide range of benchmark datasets and different vision backbones. The extensive comparisons and ablation studies prove the effectiveness of the proposed method. Quantitatively, we propose two metrics, concept-level explanation disentanglability and localizability, to evaluate the explanation quality without human annotations. The empirical results demonstrate the superiority of our concept-level explanations on both metrics. Qualitatively, in contrast to existing VLMs that exhibit an entangled understanding of concepts, our model can visually disentangle and correctly localize fine-grained concepts. The results indicate that our model understands the semantic meanings of concepts even though their appearance drastically changed

across categories. Furthermore, our model exhibits superior performance in localizing object parts, as validated by ground truth segmentation masks.

Surprisingly, the proposed method also significantly improves VLM’s prediction accuracy. This is different from most interpretability methods that come with a compromised benchmark performance [10, 22]. We argue that the disentangled concepts would alleviate the model’s reliance on spurious correlations [31, 42], thereby enabling better generalization capability to unseen data.

In summary, our main contributions:

- Our empirical results reveal that the existing VLMs cannot distinguish between fine-grained concepts. Their explanations with respect to fine-grained concepts are entangled and mislocalized.
- We propose a straightforward but effective method, called DEAL, to disentangle and localize the concept-level explanations for VLMs without human annotations. Our method works as a plug-in module without changing the model architecture.
- Empirical results in a wide range of benchmark datasets and VLMs demonstrate that our method significantly improves the concept-level explanation in terms of disentanglability and localizability by 8.8% and 10.9% on average. Furthermore, the enhanced explainability also improves the model’s prediction accuracy by 1.6% on average.

2 Related Works

eXplainable Machine Learning (XML). The emerging field of XML aims to bring transparency to today’s powerful but opaque deep learning models. XML usually consists of two categories. (1) Intrinsic methods, whose explanations are inherent to the model’s design and training, provide explanations along with output, such as joint training [19, 32] and prototyping [10] methods. However, such interpretable methods usually sacrifice their prediction performance [10, 46]. (2) Post hoc methods, which give insight into the learned associations of a model that are not readily interpretable by design. Such methods usually leverage back-propagation or local approximation to offer saliency maps as explanations, *e.g.*, Vanilla Gradient [51], Grad-CAM [49], LIME [45], and SHAP [30]. However, such saliency-based explanations are inherently ambiguous and require ML expertise to understand [25, 46]. For example, given a saliency map that highlights the correct object on the image, it remains unclear whether the model’s prediction stems from the texture or color. In contrast, the proposed method offers human-understandable explanations while maintaining high prediction performance.

Concept-based explanations. Instead of delving directly into raw features, there have been existing attempts to correlate the internal representations of machine learning models with high-level, semantically meaningful concepts familiar to humans. They either provide quantification of the influence of predefined concepts like “stripes” on the model’s predictions, *e.g.*, TCAV [20], Concept Bottleneck [22], or manually translate the abstract feature maps learned by the model into human-understandable terms, *e.g.*, ProtoPNet [10], CRP [1]. However, their

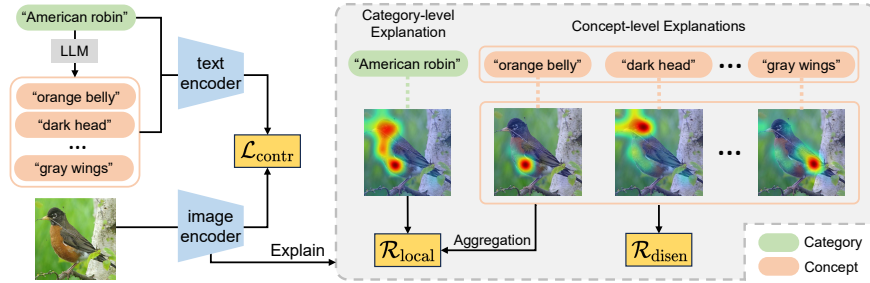


Fig. 2: Overview of the proposed method to Disentangle and Localize (DEAL) concept-level explanations for VLMs. Our method consists of the concept-level explanation disentanglement and localization constraints for training Vision-Language Models (VLMs). First, we query the Large Language Model (LLM), *e.g.*, GPT-3.5 [40], with the category name to obtain discriminative visual concepts for distinguishing the category. Then, we feed the category and concepts as a sentence into the text encoder, and the image into the image encoder to calculate the standard contrastive learning loss ($\mathcal{L}_{\text{contr}}$). Next, we calculate the explanations *w.r.t.* the category name and each of the concepts within the category. We constrain the disentanglement between the explanations of all concepts within the category ($\mathcal{R}_{\text{disen}}$). Finally, we constrain localization by the consistency between the category-level explanation and the aggregation of concept-level explanations ($\mathcal{R}_{\text{local}}$). This figure is best viewed in color.

explanations might not be faithful to the model. The former relies on subjective user-defined concepts that might not align with the model’s understanding, and is unable to localize their presence [15]. The latter is limited by human bias, it is impractical to interpret the vast and complex feature spaces of deep models [34].

Vision-Language Models (VLMs). Vision-Language Models have surged in popularity as they aim to bridge the domains of computer vision and natural language processing. These models are designed to understand descriptions for visual content. OpenAI’s CLIP [44] and Google’s ViLBERT [29] have showcased the potential of combining vision and language pre-training on large-scale datasets, resulting in state-of-the-art performance across a variety of benchmarks. Decomposing categories into visual descriptors can further improve the prediction accuracy of VLMs [35, 43, 59]. However, recent works reveal that they might lack compositional reasoning to understand the objects and their relationships [33, 38, 60]. Our empirical results further prove that VLMs cannot distinguish fine-grained concepts, their explanations with respect to these concepts are entangled and mislocalized. Although there are recent attempts on fine-grained alignment, *e.g.*, X-VLM [61] and PyramidCLIP [13], they heavily rely on expensive human annotations and are limited to category (or object) level granularity.

3 Methods

In this section, we first present how to prompt Large Language Models (LLMs) for discriminative concepts, then provide the problem formulation, and then

Bald Eagle	Tabby Cat	Bald Eagle	Tabby Cat
<ul style="list-style-type: none"> • large, powerful bird • talons • wingspan 2.5 meters • white head and tail • two wings • ... 	<ul style="list-style-type: none"> • four-legged animal • long tail • wet nose • large eyes • furry body • ... 	<ul style="list-style-type: none"> • white head and tail • dark brown body • large, hooked beak • long, broad wings • yellow feet and beak • ... 	<ul style="list-style-type: none"> • striped fur pattern • dark "M" on forehead • white paws and chest • dark ears and tail tip • round face shape • ...
Query w/o prompt engineering		Ours	

Fig. 3: Directly query GPT-3.5 [40] model would yield concepts that are not visually measurable. In contrast, our concepts are visually distinctable. The **bold** words are the category names in *ImageNet* [12], and the list follows are the generated concepts.

introduce a novel method to disentangle and localize concept-level explanations. Fig. 2 shows an overview of our method.

3.1 Prompting LLMs for Discriminative Concepts

The recent breakthroughs in LLMs demonstrate their capability to provide commonsense knowledge in language [36]. Therefore, LLMs are implicit knowledge bases that can be easily queried with natural language, even for non-experts. Inspired by recent works in language prompting [35, 59], we can extract the desired concepts for each category from LLMs efficiently for large-scale datasets.

However, as shown in Fig. 3, our empirical results show that the concept quality in terms of discriminative capacity is highly sensitive to the prompts. To address this issue, we leverage the In-context learning [8] capability of LLMs to provide Chain-of-Thought (CoT) [56] instructions. Specifically, we provide the GPT-3.5 [40] model with exemplary queries and responses, followed by the subsequent question:

Q: What are the discriminative visual features with minimum overlap for identifying a [CATEGORY] in an image?
A: The discriminative visual features that identify a [CATEGORY] in an image are:

In contrast to direct query, our concepts align with the typical human rationale for predictions. Fig. 3 shows examples of our generated concepts for the corresponding category, which are more visually distinctable.

3.2 Disentangling and Localizing Explanations

Problem formulation. The contrastive learning objective of typical VLMs, *e.g.*, CLIP [44], is to learn a mapping function $f \in \mathcal{F}$ such that for each pair $(I, T) \sim P(\mathbf{I}, \mathbf{T})$, where $P(\mathbf{I}, \mathbf{T})$ is the distribution of image-text pairs, the function f aligns the embeddings of I and T in a shared space if they are a correct match. This process is typically guided by a contrastive loss function $\mathcal{L}_{\text{contr}}(\cdot)$. As we discussed in Secs 1 and 2, such learning objective is coarse-grained and ignores the distinction between fine-grained concept-level explanations. To address this, we propose an effective approach to refine the learning process.

Let $g(\text{[TEXT]})$ be a post-hoc explanation method (e.g., Grad-CAM [49]) that calculates the explanation heatmap for an image I corresponding to the textual input [TEXT] , based on their similarity in the embedding space measured by f . Considering all [CONCEPT] belong in to the same [CATEGORY] , we propose to *disentangle and localize (DEAL) concept-level explanations* by optimizing:

$$\begin{aligned} \min_{f \in \mathcal{F}} \text{Risk}(f) &:= \mathbb{E}_{(I,T) \sim P} [\mathcal{L}_{\text{contr}}(f(I,T))] &< \textbf{Contrast} \\ \text{s.t. } \text{Dist}(g(\text{[CONCEPT]}), g(\text{[CONCEPT]}')) &\geq \epsilon, &< \textbf{Disentangle} \\ \text{Dist}(\sum g(\text{[CONCEPT]}), g(\text{[CATEGORY]})) &\leq \delta. &< \textbf{Localize} \end{aligned} \quad (1)$$

Intuitively, our goal is to ensure that the explanation of each concept within a category is distinct, while the collective explanations for all concepts are consistent with the category’s overall explanation. Next, we will explain the Contrast, Disentangle, and Localize in detail.

Contrastive learning objective. Following the standard contrastive learning methodology between image and text used in CLIP [44], we employ the InfoNCE loss [37] as our backbone loss. Denote $\{(I_i, T_i)\}_{i=1}^N$ as a batch of N paired images and texts, $(v_i, t_i) = f(I_i, T_i)$ as the vision and text embeddings, τ as a temperature scaling factor, then we define our contrastive learning loss as:

$$\mathcal{L}_{\text{contr}} = \frac{1}{N} \sum_{i=1}^N -\log \frac{\exp(v_i \cdot t_i / \tau)}{\sum_{m=1}^N \exp(v_i \cdot t_m / \tau)} \quad (2)$$

In contrast to the typical practice that simply uses category as the text description [14, 57], we augment them with the corresponding concepts obtained in Sec. 3.1. For example, we can apply a template “An image of [CATEGORY] with [CONCEPT_1] , [CONCEPT_2] , ..., and [CONCEPT_K] ”.

Disentangling concept-level explanations. As we previously discussed in Secs. 1 and 2, the key challenge for learning correct concept-level explanations is the absence of human annotations. Let $\{C_k\}_{k=1}^K$ be the set of K concepts for the category of image I as obtained in Sec. 3.1, then $g(C_k)$ calculates the explanation heatmap for image I with respect to concept C_k . Denote $\text{Dist}(\cdot, \cdot)$ as a distance metric, e.g., ℓ_1 distance, we propose to disentangle the concept-level explanations by imposing regularization on the model for each image using:

$$\mathcal{R}_{\text{disen}} = - \sum_{k=1}^K \sum_{j=1, j \neq k}^K \text{Dist}(g(C_k), g(C_j)) \quad (3)$$

Intuitively, this regularization term constrains the model’s explanations for each image *w.r.t.* different concepts to have minimum overlap. Note that $g(\cdot)$ is not restrictive, including arbitrary gradient-based explanation implementation for VLMs, such as Chefer et al. [9] and GradCAM [49]. Different from existing methods that rely on category-level region annotations [24, 61], our method provides finer-grained concept-level supervision without human annotations.

Localizing concept-level explanations. Our empirical results (Tab. 4) indicate that recklessly optimizing Eq. (3) could easily fall into local minimums.

Algorithm 1: The proposed DEAL algorithm.

Input: Image-Text distribution $P(\mathbf{I}, \mathbf{T})$; step size η
Output: Learned VLM parameters θ

```

1 for all  $[\text{CATEGORY}]$  do
2    $\{[\text{CONCEPT}]\} \leftarrow \text{LLM}(\text{prompt}([\text{CATEGORY}]))$ ; // Generate concepts
3 end
4 while not converged do
5   Sample  $(I, T) \sim P(\mathbf{I}, \mathbf{T})$ ; // Sample Image-Text pairs
6   Calculate Risk( $f$ ) via Eq. (5); // Calculate overall risk
7   Update  $\theta$  via  $\theta^{t+1} = \theta^t - \eta \nabla \text{Risk}(f)$ ; // Update model parameters
8 end

```

For example, the explanations highlight well-separated parts of the image that are not related to the concepts.

“The meaning of the whole is a function of the meanings of its parts” is held to be a key characteristic of human intelligence [11, 33]. Following the same spirit, we propose to further regularize the concept-level explanations with corresponding category-level explanations. Denote $g([\text{CATEGORY}])$ as the explanation heatmap for image I with respect to its category name, we localize the concept-level explanations by regularizing the model using:

$$\mathcal{R}_{\text{local}} = \text{Dist}\left(\frac{1}{K} \sum_{k=1}^K g(C_k), g([\text{CATEGORY}])\right) \quad (4)$$

Intuitively, this regularization term constrains the aggregation of all concept-level explanations within the same category, to be consistent with the category-level explanation. The proposed two constraints work in an adversarial-like manner. This mechanism ensures distinct concept-level explanations are collectively coherent within their category, thereby forcing effective localization.

3.3 Solving the Constrained Optimization

Solving such a constrained optimization problem in Sec. 3.2 often leads to a non-convex problem, wherein methods like stochastic gradient descent (SGD) cannot guarantee constraint satisfaction. To address this problem, we leverage Karush–Kuhn–Tucker conditions [7] and introduce Lagrange multipliers λ and γ to convert the constrained problem into its unconstrained counterpart. Our overall learning objective is formulated as follows:

$$\min_{f \in \mathcal{F}} \{\text{Risk}(f) := \mathbb{E}_{(I, T) \sim P(\mathbf{I}, \mathbf{T})} [\mathcal{L}_{\text{contr}}(f(I, T))] + \lambda \mathcal{R}_{\text{disen}} + \gamma \mathcal{R}_{\text{local}}\} \quad (5)$$

The entire training pipeline is summarized in Alg. 1. The proposed method has the following merits. (1) In contrast to the pertaining objective of VLMs which is limited to coarse-grained alignment [17, 44, 53], the proposed learning

objective can disentangle and localize fine-grained concepts. (2) Different from existing category-level alignment works that rely on human annotations [24, 61], the proposed method leverages explanations to achieve an even finer, concept-level alignment in a self-supervised manner. (3) The proposed method introduces no additional parameters, it works as a plug-in module for training VLMs via concept-level supervision without changing the model architecture.

4 Experiments

To best validate the explanation quality and prediction performance, we conduct a series of experiments to compare the proposed DEAL method with other learning methods on different vision backbones. The experimental results prove that our method achieves superior concept-level explanation disentanglability and localizability on a wide scope of benchmark datasets. In contrast to most interpretability methods that come with a compromise on the benchmark performance, our method benefits the model’s prediction.

4.1 Datasets and Implementation Details

Datasets. We conduct experiments on five typical image recognition datasets that have been used for evaluating the CLIP [44] model. (1) *ImageNet* [12] dataset stands as one of the typical benchmarks in the field of computer vision. The dataset spans 1,000 object classes and 1,281,167 training images from the WordNet hierarchy, ensuring a vast and varied semantic landscape. (2) *CUB* [54], short for Caltech-UCSD Birds-200-2011, is a highly specialized dataset designed to distinguish between subcategories of birds. The dataset contains 11,788 images of 200 bird species. (3) *Food-101* [6] dataset is designed for the task of food recognition. The dataset comprises 101,000 images split across 101 food categories. (4) *Oxford-Pets* [39] dataset includes a set of 7,349 images divided across 37 distinct pet categories, including a balanced collection of 12 dog breeds and 25 cat breeds. (5) *EuroSAT* [18] dataset is derived from Sentinel-2 satellite images, including a diverse range of European land use and land cover categories. The dataset contains 27,000 geo-referenced samples of 10 different classes.

Implementation details. For all datasets, we use the implementations of CLIP [44] for different vision backbones. Due to the computational cost of training CLIP from scratch, we focused on finetuning experiments using ViT-B/32 and ResNet-50 backbones. We fine-tune the full model leveraging the proposed method, and compare the results with zero-shot and representative fine-tuning strategies of the CLIP model, including full model and vision-encoder-only fine-tuning. For all the training, we use our augmented text of the category as the text description. We split the training data of each dataset into 80% and 20% splits, and use the larger splits for training, and the smaller splits for validation and model selection. We evaluate the model performance on the testing data provided by each dataset. All models are trained using Adam [21] optimizer until convergence. We crop the images of random size and aspect ratio, resizing

Table 1: Comparison of explanation and prediction performance with state-of-the-art VLMs. We conduct experiments on five datasets using the Vision-Transformer (ViT) encoders. The models are tested on hold-out testing data. Our results are on the average of three trials of experiments. We highlight the **best results** and the second best results. On average, our method exhibits superior concept-level explanation in terms of disentanglability and localizability. The improved explainability also benefits the model’s prediction accuracy. [†] Reported in [44], [‡] reported in [27]. See Supplementary for the full table and the results using ResNet-50 backbone.

Metrics	Models	#Param.	Datasets					
			ImageNet	CUB	Food101	OxfordPets	EuroSAT	Avg.
Concept-level Explanation Disentanglability \uparrow	CLIP [44]	151M	0.361	0.596	0.487	0.363	0.192	0.400
	FLAVA [52]	241M	0.298	0.541	0.463	0.429	0.115	0.369
	DeCLIP [27]	186M	0.032	0.071	0.042	0.056	0.013	0.043
	PyramidCLIP [13]	153M	0.048	0.116	0.080	0.085	0.026	0.071
	CLIPpy [2]	196M	0.300	0.557	0.449	0.461	0.162	0.386
	DEAL (Ours)	151M	0.397	0.608	0.501	0.475	0.192	0.435
Concept-level Explanation Localizability \uparrow	CLIP [44]	151M	0.633	0.638	0.511	0.762	0.423	0.593
	FLAVA [52]	241M	0.630	0.650	0.589	0.668	0.361	0.580
	DeCLIP [27]	186M	0.366	0.367	0.318	0.369	0.295	0.343
	PyramidCLIP [13]	153M	0.662	0.672	0.644	0.700	0.302	0.596
	CLIPpy [2]	196M	0.612	0.614	0.656	0.657	0.345	0.577
	DEAL (Ours)	151M	0.673	0.718	0.660	0.809	0.444	0.661
Prediction Accuracy (%)	CLIP [44]	151M	[†] 63.2	52.6	[†] 84.4	[†] 87.0	[†] 41.1	65.7
	FLAVA [52]	241M	55.1	49.4	79.7	57.7	28.2	54.0
	DeCLIP [27]	186M	[†] 66.2	35.9	57.0	59.1	27.3	49.1
	PyramidCLIP [13]	153M	46.0	43.8	49.3	36.0	20.0	39.0
	CLIPpy [2]	196M	45.3	18.9	53.8	47.5	18.9	36.9
	DEAL (Ours)	151M	70.8	69.6	86.9	89.3	77.4	78.8

to 224×224 pixels, random horizontal flips, random color jitter, grayscaling the image with 10% probability, and normalization using the *ImageNet* channel statistics. We use learning rate = $5e-7$. See Supplementary for more details.

4.2 Evaluation Metrics

Concept-level Explanation Disentanglability. The disentanglability of concept-level explanations is evaluated through the $\mathcal{R}_{\text{disen}}$ regularization term that is also applied in our model’s training phase. While this term was optimized on training data, the disentanglability of concept-level explanations on the hold-out test data remains essential for evaluating the success of our goal. For better comparison, its values are normalized to $[0, 1]$. A higher metric score is expected for concept-level explanations that exhibit better disentanglement.

Concept-level Explanation Localizability. Due to the absence of region annotations for concepts, we use the explanation fidelity [41] *w.r.t.* the concepts to measure their localizability. This metric quantifies the increase in predictive score probability with the sequential inclusion of pixels, sorted in descending order of their importance as derived from the explanation [4]. Different from the conventional category-level version, our metric measures the fidelity at the concept level using the predictive score of concepts. We calculate the area under

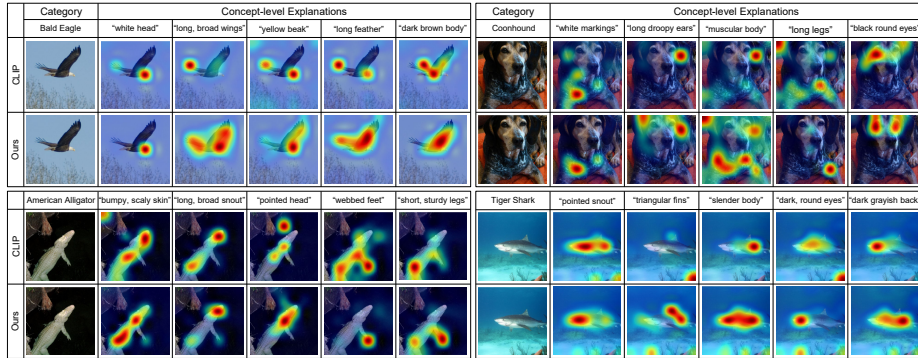


Fig. 4: Disentanglability comparison using images from the *ImageNet* [12] dataset. The CLIP [44] model’s explanations *w.r.t.* different concepts are entangled. They might highlight the same region for visually distinct concepts, or even mislocalize concepts to backgrounds. In contrast, our concept-level explanations are not only visually well-disentangled, but also correctly localized. This indicates that our model understands the semantic meanings of fine-grained concepts, and can visually distinguish them.

the insertion curve (iAUC), capturing the cumulative probability increase as features from the original input are incrementally inserted into a reference input. We employ a blurred canvas as the reference input to minimize the introduction of distributional shifts. To compare across models, we follow [48] to normalize iAUC scores to $[0, 1]$. A higher iAUC score is expected for an explanation that is faithful with respect to the concept.

Prediction Accuracy. The typical practice for VLMs to make predictions is to directly rank the similarity between the image and all category name embeddings, and select the top-1 category as prediction [44]. In contrast, to better evaluate the model’s understanding of fine-grained concepts, we propose using the similarity between the image and all concept embeddings to make predictions. Specifically, as described in Sec. 3.1, each category consists of a set of K concepts. The category similarity will be determined by the average embedding similarity between the image and all concepts within that category:

$$s(I, [\text{CATEGORY}]) = \frac{1}{K} \sum_{k=1}^K \text{similarity}(f(I, C_k)) \quad (6)$$

This has been proved by recent works [35, 43] that can improve the zero-shot prediction performance of VLMs. In this paper, we use it as a natural metric.

4.3 Comparison with State-of-the-Art VLMs

Takeaway 1: DEAL can disentangle and localize fine-grained concepts. As aforementioned, the existing learning objectives of VLMs are limited to the

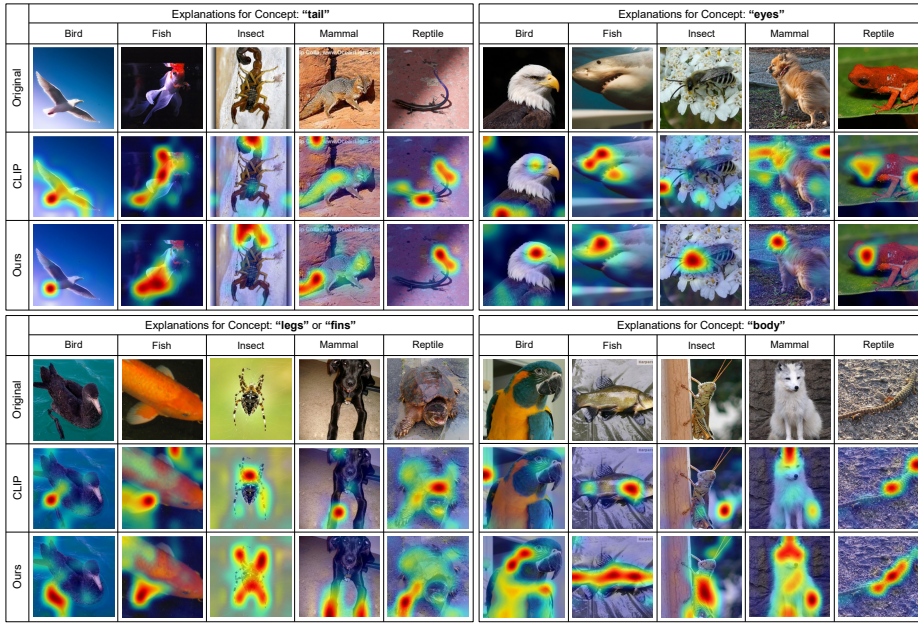


Fig. 5: Localizability comparison per concept using different animal species groups from the *ImageNet* [12] dataset. The CLIP [44] model’s explanations *w.r.t.* the same concept might vary across different species, they usually entangle with other concepts or even spurious factors. In contrast, our model consistently localizes the region *w.r.t.* the semantic meaning of the concept. Note that we explicitly select images from five diverse animal species groups, *i.e.*, Bird, Fish, Insect, Mammal, and Reptile. The results indicate that our model can generalize the learned concepts across categories even though their appearance drastically changed. This figure is best viewed in color.

category (or object) level alignment, while our method achieves finer alignment at the concept level. Furthermore, existing works typically rely on expensive annotations of objects, while our method leverages self-supervised learning thereby eliminating the need for human annotations. In this section, we compare our method with representative VLMs (*i.e.*, CLIP [44], FLAVA [52], and DeCLIP [27]) and state-of-the-art VLMs using for category-level alignment (*i.e.*, PyramidCLIP [13] and CLIPpy [2]).

Evaluation on disentanglability. Quantitatively, our model significantly outperforms VLMs and their fine-tuned versions in terms of disentanglability. As shown in Tab. 1, on the average of five benchmark datasets, our method outperforms the second-best results by 8.8%. Qualitatively, in Fig. 4, we visualize five different concept-level explanations for each image from the *ImageNet* [12] dataset. As shown, the concept-level explanations of the CLIP model are entangled, exhibiting a high degree of similarity between them. In contrast, the concept-level explanations of our model are visually disentangled. Furthermore, our concept-level explanations accurately segment the contour of the correspond-

Table 2: Comparison with typical fine-tuning on CLIP [44] model. The abbreviations “zs”, “ft-full”, and “ft-vision” correspond to the zero-shot, full model fine-tuning, and vision-encoder-only fine-tuning. Our results demonstrate that adapting pretrained VLMs to downstream tasks without constraints might deteriorate the model’s understanding of concepts and yield compromised performance. In contrast, our method achieves superior performance through the disentanglement and localization of fine-grained concepts. [†] Reported in [44]. See Supplementary for results of other datasets.

Models	ImageNet			CUB		
	Disentanglability \uparrow	Localizability \uparrow	Acc.	Disentanglability \uparrow	Localizability \uparrow	Acc.
RN50-zs	0.677	0.608	[†] 59.6	0.679	0.613	48.7
RN50-ft-full	0.452	0.641	68.8	0.608	0.647	55.0
RN50-ft-vision	0.572	0.642	61.8	0.641	0.615	54.7
RN50-ours	0.690	0.661	70.0	0.725	0.651	57.4
ViT-B/32-zs	0.361	0.633	[†] 63.2	0.596	0.638	52.6
ViT-B/32-ft-full	0.323	0.618	68.9	0.579	0.602	67.5
ViT-B/32-ft-vision	0.207	0.617	66.4	0.549	0.624	62.1
ViT-B/32-ours	0.397	0.673	70.8	0.608	0.718	69.6

ing concepts, even though most concepts lack clear borders with adjacent concepts. Our results reveal that the global alignment learning objectives of existing VLMs entangle the model’s understanding of fine-grained concepts.

Evaluation on localizability. Quantitatively, as shown in Tab. 1, our model outperforms VLMs and their fine-tuned versions in terms of localizability. On the average of five benchmark datasets, our method outperforms the second-best results by 10.9%. Qualitatively, in Fig. 5, we visualize the explanations of five different images for each concept. The images are from five diverse animal species groups in the *ImageNet* [12] dataset. As shown, the CLIP model usually associates the concepts with the entire object, sometimes even highlighting the spurious correlations (*e.g.*, watermarks). On the contrary, our model can differentiate fine-grained concepts and accurately highlight regions that align with the concepts’ semantic meanings. The results demonstrate the superior capability of our model in understanding fine-grained concepts, even though the presentation of these concepts significantly differs among species groups.

Evaluation on prediction accuracy. Our model, as shown in Tab. 1, outperforms state-of-the-art VLMs in terms of prediction accuracy. On average of five benchmark datasets, our method outperforms the second-best results by 13.1%. This finding is surprising since interpretability methods usually come with a compromise on the benchmark performance [10, 46]. The result indicates that, through improving concept-level explanations, our method alleviates the model’s reliance on spurious correlations [3], *e.g.*, watermark highlights in Fig. 5, thereby benefiting its prediction performance on unseen data.

4.4 Comparison with Fine-tuned Models

Takeaway 2: The disentanglability of DEAL can boost prediction accuracy. To better evaluate the effectiveness of the proposed constraints, we com-

Table 3: Evaluation using ground truth object part annotations. For all models, the mean Intersection over Union (mIoU) is calculated between the explanations of part and their corresponding ground truth segmentation masks. The results are on the average of 10 different parts for the *CUB-part* and 40 parts for the *PartImageNet* datasets.

Datasets	ViT-B/32			RN50		
	CLIP	CLIP-ft	Ours	CLIP	CLIP-ft	Ours
CUB-Part [47]	10.49	10.24	15.69	6.31	6.52	12.04
PartImageNet [16]	8.79	7.25	12.73	4.92	5.19	10.04

pare our method with zero-shot and representative fine-tuning methods (*i.e.*, full model fine-tuning, and vision-encoder-only fine-tuning of CLIP) on the five benchmark datasets. We use Grad-CAM [49] for ResNet-50 and Chefer et al. [9] for ViT-B/32 to generate explanations *w.r.t.* concepts. As shown in Tab. 2, our method outperforms the second-best results by 6.8%, 5.5%, and 2.1% on average of five datasets and two backbones in terms of disentanglability, localizability, and prediction accuracy. See Supplementary for the full table and details.

4.5 Evaluation using Ground Truth Object Part Annotations

Takeaway 3: DEAL can accurately localize fine-grained object parts. Object parts represent a specific subset of fine-grained concepts that can evaluate the model’s localization capability. We evaluate our model on the *CUB-Part* [47] and *PartImageNet* [16] datasets that provide ground truth masks of parts, such as “beak” and “head”, for 299 and 24,000 images. As shown in Tab. 3, our model significantly improves the localization of explanations for these concepts by 5.20% and 5.52% for two different vision backbones on *CUB-Part*, and 3.94% and 4.85% on *PartImageNet* in terms of mIoU score.

4.6 Ablation Study

In this section, we perform ablation studies to investigate the key components of our method proposed in Eq. (3) and Eq. (4). The empirical results are shown in Tab. 4. Furthermore, as shown in Tab. 5, we compare the scalability between our method and the baseline fine-tuning method.

Ablation on disentanglement constraint. The “DEAL w/o disen.” represents a variant of our method which only optimizes the consistency between the aggregation of concept-level explanations and the corresponding category-level explanation. Without constraints on the disentanglement, the concept-level explanations tend to be similar to easily satisfy the consistency constraint. This leads to entangled concept-level explanations, which also dramatically decrease the prediction performance by 10.4%.

Ablation on localization constraint. The “DEAL w/o local.” represents a variant of our method which only optimizes for the disentanglement of explanations. As we discussed in Sec. 3, recklessly disentangling concept-level explanation might easily fall into a local minimum. As shown in Tab 4, although

Table 4: Ablation study on *ImageNet* [12] dataset. In this study, we use the ViT-B/32 vision backbone. The variant without disentanglement produces entangled explanations thereby drastically deteriorating prediction accuracy. The variant without localization recklessly optimizes for disentanglement, which falls into a local minimum with compromised localizability and prediction accuracy.

Models	Disentanglability \uparrow	Localizability \uparrow	Acc. (%)
CLIP [44]	0.361 \pm 0.00	0.633 \pm 0.00	63.2 \pm 0.0
CLIP-ft-full	0.323 \pm 0.01	0.618 \pm 0.01	68.9 \pm 0.3
DEAL w/o disen.	0.354 \pm 0.01	0.657 \pm 0.02	60.4 \pm 0.8
DEAL w/o local.	0.702 \pm 0.03	0.623 \pm 0.01	62.7 \pm 0.6
DEAL (Ours)	<u>0.397</u> \pm 0.01	0.673 \pm 0.01	70.8 \pm 0.4

Table 5: Scalability comparison for training one epoch on *ImageNet* [12] dataset. In this comparison, we use the ViT-B/32 vision backbone. Note that our training time varies depending on the number of concepts (we use five in this example). The time investment is justified by the benefits, our method significantly outperforms standard training in terms of explanation quality and prediction performance.

Models	#Param.	Time	Disentangle \uparrow	Localize \uparrow	Acc (%)
CLIP [44]	151M	-	0.361	0.633	63.2
CLIP-ft-full	151M	4h06m	0.323	0.618	68.9
DEAL (Ours)	151M	8h06m	0.397	0.673	70.8

achieving high disentanglement, such explanations are incorrectly localized (low localizability) and significantly compromise the prediction performance by 8.1%.

Scalability comparison. We compare the scalability of our method with standard training. Training the full CLIP [44] with ViT-B/32 encoder on the *ImageNet* [12] dataset for one epoch using our method requires 8 hours and 6 minutes of wall clock time, compared to 4 hours and 6 minutes for standard training. However, this additional time investment is justified by the benefits. As shown in Tab. 5, our method significantly improves the model’s concept-level explanations in terms of disentanglability and localizability by 22.9% and 8.9%. Without extra parameters, our method improves the accuracy by 1.9%.

5 Conclusion

In this paper, we present comprehensive studies to show that the training objective of existing VLMs ignores the distinction between fine-grained concepts. Their explanations with respect to these concepts are entangled and mislocalized. To address this problem, we propose to Disentangle and Localize (DEAL) the concept-level explanations for VLMs. Our method fully utilizes the discrepancy and consistency between concept- and category-level explanations to provide supervisory signals for concept learning without human annotations. Extensive experiments demonstrate the superiority of our method in terms of the disentanglability and localizability of concept-level explanations.

Acknowledgment

This work is supported by the National Science Foundation through the Faculty Early Career Development Program (NSF CAREER) Award NSF-IIS-2340074 and the Department of Defense under the Defense Established Program to Stimulate Competitive Research (DoD DEPSCoR) Award.

References

1. Achibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From "where" to "what": Towards human-understandable explanations through concept relevance propagation. *arXiv preprint arXiv:2206.03208* (2022)
2. et al., R.: Perceptual grouping in contrastive vision-language models. *ICCV* (2023)
3. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019)
4. Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: *Proceedings of IJCAI* (2021)
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunsell, E., et al.: On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021)
6. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: *European Conference on Computer Vision* (2014)
7. Boyd, S., Boyd, S.P., Vandenberghe, L.: *Convex optimization*. Cambridge university press (2004)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
9. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 397–406 (October 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. *Advances in neural information processing systems* **32** (2019)
11. Cresswell, M.J.: *Logics and languages* (1973)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
13. Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., Ji, R., Shen, C.: Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems* **35**, 35959–35970 (2022)
14. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19338–19347 (2023)
15. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems* **35**, 23386–23397 (2022)

16. 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: European Conference on Computer Vision. pp. 128–145. Springer (2022)
17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
18. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2019)
19. Hind, M., Wei, D., Campbell, M., Codella, N.C., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., Varshney, K.R.: Ted: Teaching ai to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 123–129 (2019)
20. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
23. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
24. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
25. Li, T., Gao, J., Peng, X.: Deep learning for spatiotemporal modeling of urbanization. arXiv preprint arXiv:2112.09668 (2021)
26. Li, T., Qiao, F., Ma, M., Peng, X.: Are data-driven explanations robust against out-of-distribution data? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3821–3831 (2023)
27. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: International Conference on Learning Representations (2021)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36** (2024)
29. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019)
30. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017)
31. Ma, M., Li, T., Peng, X.: Beyond the federation: Topology-aware federated learning for generalization to unseen clients. In: Forty-first International Conference on Machine Learning (2024)
32. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2302–2310 (2021)

33. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023)
34. Meng, C., Trinh, L., Xu, N., Enouen, J., Liu, Y.: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports* **12**(1), 7166 (2022)
35. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2022)
36. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023)
37. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
38. Parcalabescu, L., Gatt, A., Frank, A., Calixto, I.: Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In: Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR). pp. 32–44 (2021)
39. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
40. Peng, A., Wu, M., Allard, J., Kilpatrick, L., Heide, S.: Gpt-3.5 turbo fine-tuning and api updates. <https://openai.com/blog/gpt-3-5-turbo/> (8 2023)
41. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
42. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12556–12565 (2020)
43. Qin, Z., Yi, H.H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. In: The Eleventh International Conference on Learning Representations (2022)
44. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021)
45. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
46. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019)
47. Saha, O., Cheng, Z., Maji, S.: Improving few-shot part segmentation using coarse supervision. In: European Conference on Computer Vision. pp. 283–299. Springer (2022)
48. Schulz, K., Sixt, L., Tombari, F., Landgraf, T.: Restricting the flow: Information bottlenecks for attribution. In: International Conference on Learning Representations
49. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

50. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? In: International Conference on Learning Representations (2021)
51. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
52. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)
53. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
54. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
55. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
56. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837 (2022)
57. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. arXiv preprint arXiv:2109.01903 (2021), <https://arxiv.org/abs/2109.01903>
58. Yamada, Y., Tang, Y., Yildirim, I.: When are lemons purple? the concept association bias of clip. arXiv preprint arXiv:2212.12043 (2022)
59. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)
60. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2022)
61. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. In: International Conference on Machine Learning. pp. 25994–26009. PMLR (2022)