ArtVLM: Attribute Recognition Through Vision-Based Prefix Language Modeling

William Yicheng Zhu[1]*, Keren Ye[1]*, Junjie Ke[1], Jiahui Yu[2]†, Leonidas Guibas[1], Peyman Milanfar[1], and Feng Yang[1]

[1] Google Research [2] OpenAI

Abstract. Recognizing and disentangling visual attributes from objects is a foundation to many computer vision applications. While large vision-language representations like CLIP had largely resolved the task of zero-shot object recognition, zero-shot visual attribute recognition remains a challenge because CLIP's contrastively-learned vision-language representation cannot effectively capture object-attribute dependencies. In this paper, we target this weakness and propose a sentence generation-based retrieval formulation for attribute recognition that is novel in 1) explicitly modeling a to-be-measured and retrieved object-attribute relation as a conditional probability graph, which converts the recognition problem into a dependency-sensitive language-modeling problem, and 2) applying a large pretrained Vision-Language Model (VLM) on this reformulation and naturally distilling its knowledge of image-object-attribute relations to use towards attribute recognition. Specifically, for each attribute to be recognized on an image, we measure the visual-conditioned probability of generating a short sentence encoding the attribute's relation to objects on the image. Unlike contrastive retrieval, which measures likelihood by globally aligning elements of the sentence to the image, generative retrieval is sensitive to the order and dependency of objects and attributes in the sentence. We demonstrate through experiments that generative retrieval consistently outperforms contrastive retrieval on two visual reasoning datasets, Visual Attribute in the Wild (VAW), and our newly-proposed Visual Genome Attribute Ranking (VGARank).

1 Introduction

Understanding attributes associated with objects in an image is essential for many computer vision applications, including content recommendation, visual reasoning, and text-to-image generative models. While supervised learning techniques such as classification [24, 33, 61], detection [23, 54, 55], and segmentation models [8, 56] have made significant progress in object recognition tasks, directly adding a branch for object-agnositic attribute prediction [18, 19, 46] can result in incorrect and counterfactual outputs since it fails to model the co-dependency

* Equal contribution.
† Work done at Google.

between attributes and the objects. Other existing attribute learning methods rely heavily on human annotations [2, 3, 25, 34, 77] to address this dependency, but this makes them expensive and hard to scale. All things considered, how to properly establish object-attribute relationship at scale remains an open problem.

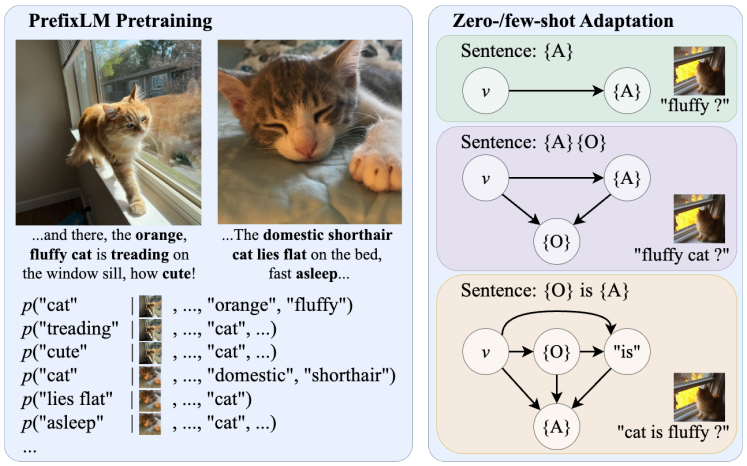
Large-scale image-text foundation models such as CLIP [49] and ALIGN [26] inspired us to explore their potential for attribute learning. These models have learned from a vast amount of noisy image-text pairs from the web, adequately utilizing self-supervised learning to benefit from easily accessible data sources. They have shown exceptional performance in zero-shot object recognition [38, 39, 48, 52, 60, 68, 72, 78] through image-text similarity measurement, a method which we refer to as "contrastive retrieval".

However, naively applying contrastive retrieval to attribute prediction tasks is suboptimal due to its two inherent problems. First, treating input text as an unstructured whole to be aligned with images results in insufficient learning on attributes if the object alone is distinguishable enough in the image to match the image-text pair, as often is the case in CLIP training data. This creates a discrepancy between the pre-training and the downstream tasks: the model learned to primarily differentiate between objects but is later asked to understand finer attributes. Second, contrastive prompting cannot model the co-dependency between objects and attributes. Since contrastive pre-training does not capture word sequence order, as opposed to language model pre-training (Fig. 1 (left)), the model is unable to correctly measure the likelihood of counterfactual combinations such as "bell-shaped sky" or "graffitied sky" (see Fig. 4). These challenges emphasize the necessity for better methods in modeling object-attribute dependency in the context of large image-text foundation models.

This paper presents a novel approach to address the two aforementioned problems in applying image-text foundation models to attribute learning. The approach consists of two parts: prefix language modeling (prefixLM) [6, 69] as the pre-training foundation, and a novel, sentence generation-based formulation of attribute retrieval that allows for the extraction of pre-training knowledge for structural reasoning (see Fig. 1). During pre-training, the prefixLM is trained to predict the next token based on visual inputs and previous text tokens, which inherently captures diverse combinations of object-attribute dependencies in the sentence. In the downstream attribute recognition task, we measure the object-attribute alignment in an image by evaluating the probability of generating a sentence capturing the relations. We refer to this approach as "generative retrieval". In particular, this method enables flexible retrieval for a wide range of attribute relations (associative, possessive, or further modified with temporal words like "currently" or "already") through building arbitrary conditional dependency models for downstream tasks at inference time (Fig. 1 (right)), which are effectively "meta-models" .

There are two immediate applications for the proposed prefixLM + generative retrieval framework: (1) describing objects through their visual appearance, state of being, or relationship to other objects in the image. And conversely,

Fig. 1: Prefix language modeling and generative prompting. During pretraining, the image-conditioned prefix language model (prefixLM) learns to generate the captions associated with images, and through this way it curates knowledge and learn to reason on object-attribute composition and dependency present in the sentence. In the downstream attribute recognition task, we propose a novel generative retrieval strategy to extract and apply the knowledge acquired from the prefixLM's large-scale pretraining. Different from contrastive retrieval, generative retrieval models the conditional dependency in a sentence, hence is more aligned with the actual language semantics. {A} and {O} are placeholders for attributes or objects in the sentence.



(2) recognizing objects based on their various visual attributes such as color, shape, size, and so on. In addition, our method can be further applied towards many other visual tasks that require structural reasoning. We summarize the contributions as follow:

1. We formally **reframe the attribute recognition problem** as a task of learning and modeling the image-object-attribute conditional probabilities in a large visual-language modeling setting.
2. We establish the effectiveness of using prefixLM as a foundational model for capturing complex object-attribute relationships in **pretraining**, and propose the generative retrieval method to flexibly **distill pretraining knowledge for downstream attribute recognition tasks**.
3. We demonstrate the limitations of purely using contrastive learning for attribute recognition and show the superior zero-shot and finetuning performance of our method.
4. We introduce Visual Genome Attribute Ranking (VGARank), a novel benchmark combining attribute and object recognition tasks into an unified, and therefore directly comparable setting, to demonstrate the generalizability of the proposed approach.

2 Related Work

We first introduce studies on attribute learning, which mostly rely on handcrafted probabilistic models without the use of large language models. Then, we summarize existing works on language modeling to specifically introduce PrefixLM to the attribute learning tasks. Finally, we provide an overview of existing prompt learning techniques to introduce our novel approach of generative retrieval, which distillate information from the pretrained PrefixLM.

Visual attribute recognition involves identifying the properties of visual objects, such as their color, material or shape. Early works had focused on object description (*img*→*att*) using classification [18, 19, 46] or relative ranking models [9, 31, 45, 67, 74] to learn attribute presence independent of object category. Some works use attributes as a foundation [2, 25, 34] for zero-shot object recognition (*img,att*→*obj*; e.g., recognizing "zebra" by the attributes "black", "white", "striped"). These works learn an attribute vector space and use it to recognize unseen visual objects based on the marginal probability. In visual object detection, some models [3, 77] train additional attribute prediction branches using the Viusal Genome dataset [32] to improve model diversity and to create models with multi-tasking capabilities. These models concatenate the visual feature with the ground-truth object class embedding and feed them into an attribute prediction branch (*img,obj*→att).

Vector space-based approaches has also been studied for attribute recognition. For example, [7, 22, 48] apply the CLIP [49]. They use the CLIP embedding to compare visual information against predefined attribute prompts (*img*↔*obj,att*), to determine if the image contains those attributes. In addition to CLIP, [40–43] allow objects and attributes to be projected into the same feature space, while the decoration of attributes on objects is modeled as an operator (*img*↔*obj* OP *att*, operator OP could be ± or linear transform).

Our innovation lies in the novel view of treating attribute recognition as a language modeling problem. We integrate probability modeling for image, object class, and attribute prediction in an unified image-text model, while leveraging LLM's foundational pre-training.

Language modeling (LM) estimates the probability of a sequence of words being observed in a sentence. Early language models use dense neural networks [6] or recursive neural networks [70], while recent large language models (LLMs) [15, 16, 49, 69, 75] are based on the transformer architecture [66] because of its strong generalizability. LM has many applications in both NLP and computer vision, including question answering(QA) [51, 71], conversational question answering [53], visual captioning [1, 12, 59, 73], and visual question answering [4, 21, 76]. These applications can be categorized into three main types of LM: (1) image-text matching [20], (2) masked language modeling [16], and (3) prefix language modeling [6].

Attribute recognition is a condensed VQA problem that requires predicting the visual attribute of an query object. The foundational methods proposed in the VQA domain mostly combine image-text matching and masked lan-

guage modeling. Examples include LXMERT [63], UNITER [13], OSCAR [35], VinVL [77], ViLT [29], VLMo [5].

Different from these works, we show that prefix language modeling (prefixLM) [6,10,11,69,75]) can approximate masked language modeling (see Sec. 3.3) in the attribute tasks. With a novel prompting scheme, prefixLM exhibits even greater expressive power than MLM, making it a powerful tool for deep visual reasoning [65, 72].

Prompt learning originates in the field of natural language processing (NLP), where tasks like question-answering are frequently formulated as a "fill-in-the-blank" problem. Notable examples include BERT [16] which employs masked language modeling, and GPT [50] that uses prefix language modeling. While large language models (LLMs) [15, 44, 64] have been widely explored in NLP for fact-based reasoning, their application in the computer vision domain is relatively unexplored.

Prompt learning in computer vision has gained attention following the success of CLIP [49]. Numerous works [38, 39, 48, 52, 60, 68, 72, 78] have focused on designing CLIP-prompts or utilizing the pre-trained CLIP checkpoint. Approaches such as [79, 80] learn the prompting vectors instead of manually designing text prompts.

Our approach focus on its application towards attribute learning, which the aforementioned contrastive learning based methods are ill-suited for. Our proposed generative prompting is based on image-conditioned prefix language modeling [69, 75], which takes sequence ordering into consideration and is therefore well-suited for modeling the dependence between visual objects and attributes. The proposed method has potential applications in other visual reasoning problems such as visual relation detection [37] or scene graph generation [28].

3 Approach

3.1 Image-Conditioned Language Modeling

Our proposed generative retrieval is based on image-conditioned prefix language modeling, i.e. image captioning. Given an image v, we aim to generate the corresponding text $x = (s_1, ..., s_n)$ by modeling the probability $p(x|v)$ using Eq. 1. This equation factors $p(x|v)$ into the product of conditional probabilities [6, 50], where at each time step, the model predicts the next token s_i based on the visual input v and previous tokens $(s_0, ..., s_{i-1})$ (s_0 is the start-of-sentence token "<s>").

$$p(x|v) = \prod_{i=1}^{n} p(s_i|v, s_1, \ldots, s_{i-1}) \tag{1}$$

The factorization provided by Eq.1 is advantageous as it breaks down the word generation process into individual probability factors. In Fig. 1 (left), we show that the model can capture various object-attribute compositions during pre-training. As a result, in downstream attribute-related tasks, we can leverage

this factorization to address reasoning questions such as $p(w_{att}|v, w_{obj})$, which represents the probability of observing an attribute w_{att} (e.g., "orange", "fluffy") given the visual input v and object w_{obj} (e.g., a "cat").

3.2 Generative Retrieval for Attribute Classification

We formalize the simplest attribute classification task as a common foundation for both generative retrieval and contrastive retrieval. Specifically, given an image v and sentence $t^{(1)}, \ldots, t^{(C)}$ (C is number of classes), retrieval-based classification involves designing a loss function $L(v, t)$ to measure the cost of aligning image v and text $t^{(i)}$ ($1 \leq i \leq C$). Thus, zero-shot classification can be achieved through finding the class label $c = \text{argmin}_{1 \leq i \leq C}\{L(v, t^{(i)})\}$.

Contrastive retrieval builds on the fact that paired image-text are projected into the same feature space through contrastive learning during pre-training. Assuming the image is encoded as $f(v)$ and the text is encoded as $g(t)$, the contrastive learning objective aims to maximize the inner product between the matched image-text embeddings while minimizing the unmatched ones. This encourages paired image-text samples to have a high similarity while pushing unpaired samples apart. Under the common assumption of unit norm in the embeddings [26, 49, 62], this can be equivalently represented by using the L2 loss to measure the distance between image and text, denoted as $L^{(con)}(v, t) = \|f(v) - g(t)\|_2$.

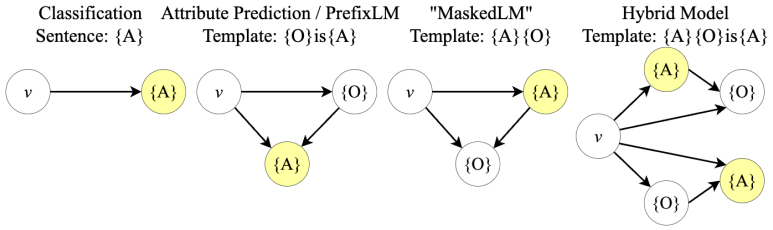
Generative retrieval is our proposed approach for visual attribute recognition, which utilizes cross-entropy to evaluate the image-text alignment loss, represented as $L^{(gen)}(v, t) = -\sum_{i=1}^{N} \hat{p}(t_i) \log q_\theta(v, t_{j|j<i})$. Here, $\hat{p}(t_i) \in \mathbb{R}^{1 \times V}$ ($1 \leq i \leq N$, N is the length) represents the one-hot representation of the i-th token of sentence t. To generate the information at the i-th step, the model q_θ relies on the image v and all previous text tokens $t_{j|j<i}$ to produce a probability distribution $q_\theta(v, t_{j|j<i}) \in \mathbb{R}^{V \times 1}$ over the vocabulary V. The term $-\hat{p}(t_i) \log q_\theta(v, t_{j|j<i}) \in \mathbb{R}^1$ represents the cross-entropy between the i-th token in the sentence t and the model's prediction at the i-th step. Fig. 3 (middle) provides a visual representation of this equation.

3.3 Modeling the Conditional Dependence

In generative retrieval, we can build different probabilistic models for visual attribute recognition by changing word ordering in the to-be-measured sentence (see Fig. 2). Our key contribution to the community is proposing and showcasing its versatility in enabling the design of arbitrary probabilistic model by engineering different structure for the measured sentence. Below, we use *{A}* to indicate attribute, and *{O}* to indicate object.

Sentence *"{A}".* This sentence models the simplest dependency for predicting attribute based on the image. In this dependency model, we focus on the cross-entropy of classifying the image as having a specific attribute, which can be achieved through a simple classification model. This approach aligns with early

Fig. 2: Conditional dependencies modeled by different sentence templates. Attribute recognition is modeled as a fill-in-the-blank problem for the highlighted *"{A}"* in the graph. Our proposed method optimizes or approximates the joint probability of observing these graph meta-modals, all while only relying on the prefixLM pre-training.



methods [9, 18, 19, 31, 45, 46, 67, 74] that describe attributes rather than naming the objects.

Sentence *"{O} is {A}".* This sentence template models the prediction of attributes based on both an image and an object, approximating $p(\text{"}\{A\}\text{"} \mid v, \text{"}\{O\}\text{"})$. In this dependency model, all sentences share the same prefix *"{O} is"* (e.g., "cat is orange", "cat is fluffy", "cat is cute", etc.). Therefore, the only factor that matters in generative retrieval becomes $-\hat{p}(\text{"}\{A\}\text{"})q_\theta(v, \text{"}\{O\}\text{"}, \text{"is"})$, which quantifies the loss associated with classifying an attribute given the image and object. This dependency model characterizes recent attribute works such as [3, 47, 48, 77].

Sentence *"{A}{O}".* This sentence is similar to Masked Language Modeling (MLM) [16] as it involves filling in the blank in a sentence like "an image of a [MASK] cat". However, there are two key distinctions: (1) $p(\text{"}\{A\}\text{"} \mid v)$ requires the attribute must be easily recognizable from the image, and (2) $p(\text{"}\{O\}\text{"} \mid v, \text{"}\{A\}\text{"})$ requires that the attribute can be employed to modify the object. In contrast, MLM uses all contextual information to predict the masked token (attribute), regressing to the earlier sentence *"{O} is {A}"*). The probabilistic modeling derived from the sentence *"{A}{O}"* closely resembles the approaches in [2, 25, 34], where attributes were utilized for object recognition.

Sentence *"{A}{O} is {A}".* This sentence produces unconventional sentences such as "fluffy cat is fluffy". We highlight this sentence template to showcase the versatility of generative retrieval. In essence, this likelihood formulation includes all three previously discussed conditional probability terms: (1) $p(\text{"}\{A\}\text{"} \mid v)$ – classification; (2) $p(\text{"}\{O\}\text{"} \mid v, \text{"}\{A\}\text{"})$ – object-attribute compatibility; and (3) $p(\text{"}\{A\}\text{"} \mid v, \text{"}\{O\}\text{"})$ – attribute prediction based on image and object. We present an approximate probability graph in Fig. 2 (right), where we duplicate both the attribute and object nodes. With the duplicated *"{A}"* in the sentence, the resulting modeling accounts for the co-dependency between object and attribute. For example, attributes preceding objects: "red car", "blue sky"; and objects preceding attributes: "kid is smiling", "cat is lying". This construction further bridges the gap between pre-training and zero-shot inference.

Discussion. Our proposed generative retrieval is novel for two reasons. Firstly, from the language modeling perspective, we offer a new solution for

training using prefixLM, enabling the model to mimic MLM or more advanced LM in a zero-shot manner for downstream tasks. Secondly, our generative retrieval produce dependency models at inference time that serves as meta-models for attribute recognition, since we can flexibly modify the probabilistic modeling and conditional dependence through changes in sentence templates. In our experiments, we show the results for the four different probabilistic attribute models (Fig. 2).

3.4 Finetuning on Attribute Tasks

Since the attribute class names have similar lengths, their cross-entropy scores $L^{(gen)}(v,t)$ with the image are expected to fall within a similar range of values (see Fig. 4). Therefore, an intuitive way to adapt the knowledge in a few-shot manner is to learn to "rescale" the retrieval scores to adapt to the new dataset priors during finetuning. Specifically, if $t^{(c)}$ is the sentence for class c, we introduce learnable parameters bias μ_c and scaling factor σ_c to adjust the $L^{(gen)}(v,t^{(c)})$, resulting in a transformed probability $p_c = \text{sigmoid}\big(-\frac{L^{(gen)}(v,t^{(c)})-\mu_c}{\sigma_c}\big)$. This probability p_c represents the likelihood of the image-object pair being associated with attribute c. During finetuning, p_c can be optimized using cross-entropy loss. In Sec. 4.2, we also provide the baseline results of finetuning the contrastive retrieval, using the same approach (but with loss score $L^{(con)}$ instead of $L^{(gen)}$).

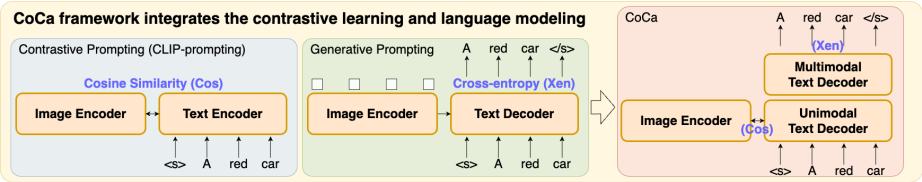
For the additional parameters μ_c and σ_c, we initialize μ_c using -15.0 and σ_c using 0.5, inspired by the values we observed in Fig. 4 (which shows $-L^{(gen)}(v,t)$ for sorting purpose). The initial values roughly transform the logits p_c to zero mean and scale the standard deviation to 6.0. To finetune the model, we use a batch size of 4, a maximum text length of 16, a weight-decay of 0.01. We use the Adafactor optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of 1e-5 linearly decayed to zero in 100k training steps, which are roughly 1.8 training epochs. All experiments are conducted on single machine with TPUv3 with the average time to finetune a model being 7 hours.

4 Experiments

4.1 Implementation Details

We use CoCa [75] pretrained on the LAION [58] as the foundation model. CoCa combines multimodal contrastive learning with image-conditioned prefix language modeling, as illustrated in Fig. 3. Its text decoder consists of (1) a unimodal text decoder trained on a contrastive learning objective with an image encoder, and (2) a multimodal text decoder trained to generate image captions by cross-attending to the image encoder. We adopt CoCa as the foundation model as it allows for performing both contrastive and generative retrieval with one model trained on the same image-text data, ensuring a fair comparison. In our experiments, we use the CoCa Base model, which consists of a ViT [30] image encoder with 12 transformer layers, a unimodal text decoder with 6 layers,

Fig. 3: Overview of Coca. CoCa integrates both contrastive learning and prefix language modeling. While its text decoder as a whole (Unimodal+Multimodal) learns to caption images, the first few layers (Unimodal) can be used for contrastive learning.



and a multimodal text decoder with an additional 6 layers. The image resolution is set to 224×224 pixels with a patch size of 16×16 pixels. All transformer layers have hidden dimensions of 768 and MLP size of 3,072.

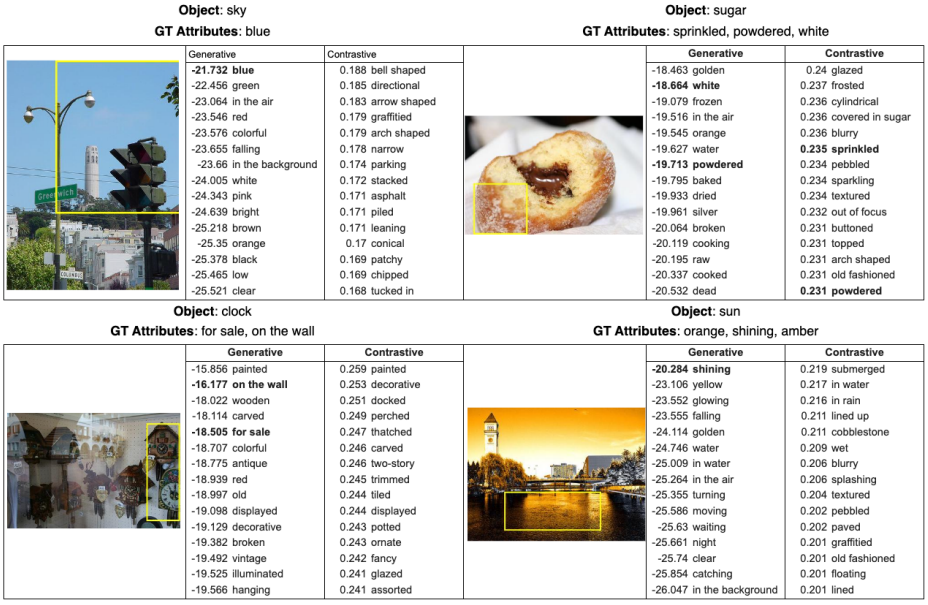
The following two datasets are used for evaluation:

Visual Attribute in the Wild (VAW) [47] is a large dataset of images with explicitly labeled positive and negative attributes. The associated task requires a model to predict a set of visual attributes given an object's name and the image it is on. VAW contains 216,790 objects from 58,565 images for training, 12,286 objects from 3,317 images for validation, and 31,819 objects from 10,392 images for testing. We use the test set to report results after validating the model.

Visual Genome Attribute Ranking (VGARank) is a modified version of the Visual Genome (VG) dataset [32] also designed to evaluate a model's ability to recognize visual attributes. The proposed dataset is different from VAW in that it is 1) an open-vocabulary ranking task, instead of a fixed vocabulary domain classification task, and 2) has two variants, VGARank-Attribute or VGARank-Object focusing on either attribute recognition given an object or object recognition given an attribute. This allows us to investigate how pretraining knowledge differs between attribute concepts and object concepts.

For VGARank-Attribute, each ranking problem is formulated with respect to one anchor object, with N ground truth attributes paired with that object in Visual Genome's annotations and $(50-N)$ additional false attributes. VGARank-O mirrors this design, but is formulated with respect to an anchor attribute present on the image. To make the problem challenging, these false pairings are selected in accordance to the dataset's conditional probability $P(object|attribute)$ or $P(attribute|object)$, i.e. for a given object, we select the attributes most likely to co-occur with the object in the Visual Genome distribution but which are not true for the given bounding box. Additionally, if any of the selected fake pairing exists on the current image as part of another bounding box instance, we would not include it in the set of $(50 - N)$ fake pairings. In the case where the given object or attribute does not appear often enough in Visual Genome and there is not enough fake pairing candidates from the conditional probability $P(object|attribute)$ or $P(attribute|object)$ to make up the set of $(50 - N)$, we

Fig. 4: Zero-shot attribute prediction - qualitative results on the VAW dataset. Images are cropped using the yellow bounding boxes, and models only see the areas inside the boxes.



would select fake object or attribute according to the dataset prior $P(object)$ or $P(attribute)$ to fill the rest of the choices. We obtain a dataset with 770,721 ranking problems for training, 7,997 for validation, and 32,299 for testing, and the dataset is available on our GitHub page.

4.2 Results on the VAW Dataset

First, we show that generative retrieval is better than contrastive retrieval, then analyze various conditional models, and finally compare our results to the state-of-the-art and analyze in particular its much superior performance on the less frequently seen categories.

Generative v.s. contrastive retrieval. The VAW dataset and the following metrics were used: rank (average rank of the correct predictions out of all 620 choices), mR@15 (mean recall over all classes at top 15 predictions for each instance), and mAP (mean average precision over all classes). We use average rank as the primary metric as it is more direct and comprehensive at describing overall ranking performance in a large candidate space.

Tab. 1 and 2 show the results of the zero-shot and fine-tuning settings, respectively. Generative retrieval outperforms contrastive retrieval in both settings, demonstrating a stronger ability to model fine-grained associations between ob-

Table 1: Zero-shot results on VAW. Generative retrieval vs contrastive retrieval, across different sentence templates / dependency meta-models. **Blue** and **black** numbers in bold represent the best and second best, respectively. The best performing setting is highlighted in gray.

	Dependency	Rank↓	mR$^{@15}$↑	mAP↑
Con	*"{A}"*	95.1	32.0	52.5
	"{A}{O}"	149.8	22.4	47.1
	"{O}is{A}"	151.4	23.2	45.9
	"{A}{O}is{A}"	141.0	23.7	48.3
Gen	*"{A}"*	82.1	28.7	**53.8**
	"{A}{O}"	63.9	**35.9**	47.7
	"{O}is{A}"	**61.9**	**32.9**	46.1
	"{A}{O}is{A}"	**56.0**	31.7	**49.9**

Table 2: Finetuning results on VAW. Generative retrieval vs contrastive retrieval, across different sentence templates / dependency meta-models. **Blue** and **black** numbers in bold represent the best and second best, respectively. The best performing setting is highlighted in gray.

	Dependency	Rank↓	mR$^{@15}$↑	mAP↑
Con	*"{A}"*	18.3	48.6	69.6
	"{A}{O}"	12.8	59.8	65.7
	"{O}is{A}"	12.3	58.9	66.7
	"{A}{O}is{A}"	12.2	59.6	67.3
Gen	*"{A}"*	18.0	50.5	71.7
	"{A}{O}"	11.4	61.8	70.8
	"{O}is{A}"	**11.1**	**62.1**	**72.0**
	"{A}{O}is{A}"	**10.6**	**62.6**	**71.9**

jects and attributes. Generative retrieval achieves a rank of 56.0 with its best sentence template, compared to 95.1 (↓ lower is better) for contrastive retrieval in the zero-shot setting (Tab. 1) and similarly achieves 10.6 vs 12.2 in the finetuning setting. (Tab. 2). As previously mentioned, there are two underlying reasons for generative retrieval's superiority. First, it captures true visual attributes, while contrastive retrieval may learn superficial connections through object identities (as shown in Tab. 1, adding object hints in contrastive retrieval makes it perform worse). Second, it explicitly models the object-attribute relations through their dependencies and interactions, which eliminates counterfactual attribute-object pairs. Fig. 4 shows some qualitative examples of differences between generative retrieval and contrastive retrieval. In the top-left example, the contrastive retrieval ranks highly the attributes "sky is bell shaped" and "sky is graffitied", which are strongly associated with other objects present in the bounding box but which are not applicable to the entity of sky. This shows that contrastive retrieval can surface attributes based on image-level associations acquired from contrastive pre-training, which is highly undesirable for attribute recognition.

Conditional dependence modeling. In Tab. 1 and 2, we also investigate the four types of graphical models (see Fig. 2) that generative retrieval approximate. As finetuning results shows similar trends, we present the zero-shot results in Tab. 1 (bottom).

The simple classification sentence template *"{A}"* does not model the important object prior and achieves the worst rank of 82.1. The PrefixLM sentence template *"{A}{O}"* produces a better model with a rank of 63.9, as it first classifies attributes then checks whether the attributes fit the *"[MASK] {O}"*, and the MLM sentence template *"{O} is {A}"* has a similar rank of 61.9. We want to highlight that while all baselines on the VAW in Tab. 3 are analogous to this formulation, it is not the best among the four graphical models. Therefore, improving the probability modeling in these SOTA methods can potentially further

Table 3: Comparing to the SOTA on the VAW dataset. The top rows show the baseline models; the last three rows shows the results of our method, finetuned with generative retrieval sentences. On the best baseline model, TAP, we primarily compare against the version without in-domain pretraining (LSA) on the evaluation dataset. The performance of the version of TAP with in-domain pretraining is included for completeness. For metric mA, we report mA@threshold=0.005 as cross-validated. Due to space constraints, we are moving the overall mR@15 and f1@15 to the supplementary material.

Methods	Overall		Class Imb. mAP			Attribute Type mAP						
	mAP	mA	Head	Med.	Tail	Colr.	Mat.	Shap.	Size	Txtr.	Actn.	Other
ResNet-Bas.-CE	56.4	50.3	64.6	52.7	35.9	54.0	64.6	55.9	56.9	54.6	47.5	59.2
LSEP	61.0	67.1	69.1	57.3	40.9	56.1	67.1	63.1	61.4	58.7	50.7	64.9
PartialBCE+GNN	62.3	68.9	70.1	58.7	40.1	57.7	66.5	64.1	65.1	59.3	54.4	65.9
ResNet-Bas.	63.0	68.6	71.1	59.4	43.0	58.5	66.3	65.0	64.5	63.1	53.1	66.7
ML-GCN	63.0	69.5	70.8	59.8	42.7	59.1	64.7	65.2	64.2	62.8	54.7	66.5
sarafianos2018deep	64.6	68.3	72.5	61.5	42.9	62.9	68.8	64.9	**65.7**	62.3	56.6	67.4
SCoNE	68.3	71.5	**76.5**	64.8	48.0	70.4	**75.6**	68.3	**69.4**	68.4	60.7	69.5
TAP(w/o in-domain)	65.4	67.2	-	-	-	-	-	-	-	-	-	-
TAP(in-domain PT)	73.4	73.5	77.6	72.9	58.8	71.6	74.5	71	73.4	70.4	67.9	77.3
Ours *"{a}{o}"*	70.8	73.7	74.0	71.0	58.2	73.1	75.0	70.9	61.8	72.2	68.8	70.7
Ours *"{o}is{a}"*	**72.0**	**74.7**	74.9	**72.0**	**60.6**	**75.2**	**76.0**	**72.6**	62.9	**72.7**	**69.6**	**72.0**
Ours *"{a}{o}is{a}"*	**71.9**	**74.4**	**75.0**	**72.1**	**59.4**	**75.7**	75.3	**71.2**	62.8	**72.2**	**70.3**	**71.8**

improve their performance, and our generative retrieval offers an easy way to do so. Finally, the hybrid sentence template *"{A}{O} is {A}"* performs the best with an average rank of 56.0. This is because it jointly considers three important factors: $p($*"{A}"* | v$)$, $p($*"{O}"* | v, *"{A}"*$)$, and $p($*"{A}"* | v, *"{O}"*$)$, all captured by the proposed generative retrieval. In particular, it is difficult for MLM to capture the latter two factors simultaneously, and this shows how our proposed prefixLM + generative retrieval is a more flexible approach.

Comparing to the SOTA. We compared our fine-tuned model to the state-of-the-art methods in Tab. 3 using the following metrics from [47]: mAP (mean average precision over all classes), mA (mean balanced accuracy over all classes), mR@15 (mean recall over all classes at top 15 predictions in each instance), and F1@15 (overall F1 at top 15 predictions). The table below focuses on fine-grained mAP metrics as the best comparison for retrieval performance. Results on mR@15 and F1@15 can be found in the supplemental. The following baselines were considered: ResNet-Bas.-CE [3, 27], ResNet-Bas. [46], LSEP [36], [57], PartialBCE + GNN [17], ML-GCN [14], SCoNE [47], and TAP [48]. We thank the authors from [47] for reimplementing/adapting all the baselines.

By leveraging vision-based PrefixLM pretraining, our method ranks first among methods without in-domain pretraining on the target evaluation dataset. Even compared to TAP with in-domain pretraining, our method ranks only slightly behind TAP in overall metric and outperforms TAP on most long-tail categories.

Compared to SCoNE and TAP, our method focuses on cross-domain knowledge extraction instead of task-specific modules or training procedures to fit to the dataset domain. The SCoNE and TAP work both rely on object mask supervision and large custom datasets during training, incorporates specialized modules, and are trained and tested in the same domain. In particular, the main metric gain in the TAP work is through in-domain pretraining on LSA [48]. Our method, on the other hand, is designed to be generalist and compatible to large-scale pretraining, while at the same time incorporating flexible probabilistic modeling in the architecture. One major advantage of our method is its significantly better performance in the distribution long tail — the rarer attributes of the Medium (72.0% mAP) and Tail (60.6% mAP) attribute classes, as shown in table 3. While SCoNE and TAP's has higher overall mAP and Head mAP, it does not necessarily mean they are better models overall, since these numbers are biased towards frequently observed attributes where it is easy for models to directly fit to the dataset, especially when it is in the same domain. Our model's strong performance on the long tail demonstrates its ability infer on rarely seen attributes and indicates that it fits to the priors carried by the foundation models, instead of to the dataset distribution.

Table 4: Zero-shot results on VGARank-Attribute. Generative retrieval vs. contrastive retrieval, across different sentence templates / dependency meta-models.

	Dependency	Rnk↓	R@1↑	R@5↑	R@10↑
Con	*"{**A**}"*	17.3	6.3	22.7	38.4
	*"{**A**}{O}"*	16.1	9.7	28.9	43.6
	*"{O}is{**A**}"*	17.2	8.7	26.4	40.7
	*"{**A**}{O}is{**A**}"*	16.5	9.0	27.9	42.8
Gen	*"{**A**}"*	14.0	8.9	34.2	53.0
	*"{**A**}{O}"*	**13.0**	13.9	41.6	**58.6**
	*"{O}is{**A**}"*	13.1	**15.2**	**42.3**	**58.6**
	*"{**A**}{O}is{**A**}"*	**12.0**	**17.6**	**46.6**	**62.2**

Table 5: Zero-shot results on VGARank-Object. Generative retrieval vs. contrastive retrieval, across different sentence templates / dependency meta-models.

	Dependency	Rnk↓	R@1↑	R@5↑	R@10↑
Con	*"{**O**}"*	6.0	32.4	70.2	83.2
	*"{**O**}is{A}"*	5.9	34.1	70.2	82.9
	*"{A}{**O**}"*	5.9	35.3	70.9	83.2
	*"{A}{**O**}is{A}"*	6.0	34.7	70.0	82.5
Gen	*"{**O**}"*	6.1	31.3	70.3	83.2
	*"{**O**}is{A}"*	**6.0**	38.9	**73.2**	**83.6**
	*"{A}{**O**}"*	**5.8**	**40.6**	**74.2**	**84.4**
	*"{A}{**O**}is{A}"*	6.4	**41.6**	72.3	82.0

4.3 Results on the VGARank Dataset

Generative vs Contrastive Retrieval. We make similar observations as on the VAW dataset, shown in Tab. 4 and 5. We observe that generative retrieval sentence variations significantly outperformed the contrastive counterparts on both datasets. The best generative retrieval sentence template on VGARank-Attribute is *"{A}{O} is {A}"*, achieving a rank of 12.0, while the best one on VGARank-Object is *"{A}{O}"*, achieving a rank of 5.8. This again verifies our claim that generative retrieval is more optimal for attribute recognition than contrastive retrieval.

Conditional dependence modeling. Tab. 4 and 5 show the results on VGARank-A and VGARank-O. We boldface the targets to be predicted, which are *"{**A**}"* for VGARank-A, and *"{**O**}"* for VGARank-O.

As in VAW, the classification template *"{**A**}"* or *"{**O**}"* is the least effective, with a rank of 14.0 on VGARank-A and 6.1 on VGARank-O. The PrefixLM template *"{**A**}{O}"* or *"{**O**} is {A}"* performs better with rank of 13.0 and 6.0, which is expected as it first classifies the target token then checks whether the target token fits the context. However, the more optimally ordered MLM template *"{O} is {**A**}"* or *"{A}{**O**}"* mostly outperforms the previous approach, which aligns with conclusions drawn by earlier works like [2, 25, 34] that suggest attributes help the classification of uncommon objects, hence our model's better-than-SOTA performance on the mid and long-tail categories on VAW. Notably, *"{A}{**O**}"* achieves the best performance on the VGARank-O. The Hybrid template *"{**A**}{O} is {**A**}"* or *"{A}{**O**} is {A}"* performs the best on VGARank-A with a rank of 12.0, but it falls behind the *"{A}{**O**}"* variant on VGARank-O. We attribute this to the more challenging nature of attribute recognition as compared to object recognition, where the former can benefit from the more complex dependency modeling in our method, while the latter still relies more on salient information. The VGARank-A/VGARank-O experiments highlights the versatility of the proposed method in predicting attributes from objects and vice versa. This flexibility demonstrates the foundational nature of the prefixLM approach through generative retrieval — by simply making changes to the sentence template, we can construct various explainable probabilistic models, expanding the possibilities for modeling complex relationships between objects and attributes.

5 Conclusion and Broader Impact

Our work reformulates attribute learning as a probabilistic modeling problem and a knowledge extraction process from pretraining to downstream tasks, and we in turn propose the generative retrieval method on a vision-based prefixLM foundation. By leveraging the knowledge on complex word dependencies captured by the foundation model during pretraining, our work enables the explicit modeling of various object-attribute dependencies in downstream attribute tasks. We also showcase the method's flexibility in emulating various conditional probabilistic dependencies, and we envision that vision-based prefixLM, through the proposed generative retrieval method, can serve as a universal framework to construct meta-models for representing and measuring complex logical relations.

As our method studies the visual attribute recognition problem in the context of large pretrained models, advancements in large vision-language models will directly translate to stronger performance on this domain. This work also benefits the broader community on image generative models by providing a better metric than CLIP for image-text alignment at a fine-grained, attribute level. This additional metric can lead to the creation of cleaner text-image datasets that has higher standards for caption details on object-attribute correctness, which would greatly benefit the training of generative models in the community.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
2. Al-Halah, Z., Tapaswi, M., Stiefelhagen, R.: Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5975–5984 (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV) (2015)
5. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems **35**, 32897–32912 (2022)
6. Bengio, Y., Ducharme, R., Vincent, P.: A neural probabilistic language model. Advances in neural information processing systems **13** (2000)
7. Chen, K., Jiang, X., Hu, Y., Tang, X., Gao, Y., Chen, J., Xie, W.: Ovarnet: Towards open-vocabulary object attribute recognition. arXiv preprint arXiv:2301.09506 (2023)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
9. Chen, S., Grauman, K.: Compare and contrast: Learning prominent visual differences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1267–1276 (2018)
10. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al.: Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565 (2023)
11. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
12. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
13. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. In: European Conference on Computer Vision (ECCV) (2020)
14. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
15. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)

16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
17. Durand, T., Mehrasa, N., Mori, G.: Learning a deep convnet for multi-label classification with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
18. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1778–1785. IEEE (2009)
19. Ferrari, V., Zisserman, A.: Learning visual attributes. Advances in neural information processing systems **20** (2007)
20. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf
21. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
22. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
23. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
25. Jayaraman, D., Grauman, K.: Zero-shot recognition with unreliable attributes. Advances in neural information processing systems **27** (2014)
26. 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: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
27. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
28. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
29. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
30. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., Zhai, X.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

31. Kovashka, A., Parikh, D., Grauman, K.: Whittlesearch: Image search with relative attribute feedback. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2973–2980. IEEE (2012)
32. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
34. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 951–958. IEEE (2009)
35. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (August 2020)
36. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
37. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (2016)
38. Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., Xie, X.: Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18051–18061 (June 2022)
39. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16410–16419 (June 2022)
40. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
41. Naeem, M.F., Xian, Y., Tombari, F., Akata, Z.: Learning graph embeddings for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 953–962 (June 2021)
42. Nagarajan, T., Grauman, K.: Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 169–185 (2018)
43. Nan, Z., Liu, Y., Zheng, N., Zhu, S.C.: Recognizing unseen attribute-object pair with generative model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8811–8818 (2019)
44. OpenAI: Gpt-4 technical report (2023)
45. Parikh, D., Grauman, K.: Relative attributes. In: 2011 International Conference on Computer Vision. pp. 503–510. IEEE (2011)
46. Patterson, G., Hays, J.: Coco attributes: Attributes for people, animals, and objects. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 85–100. Springer (2016)
47. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13018–13028 (June 2021)

48. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Improving closed and open-vocabulary attribute prediction using transformers. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
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: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
50. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
51. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
52. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18082–18091 (June 2022)
53. Reddy, S., Chen, D., Manning, C.D.: Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics **7**, 249–266 (2019)
54. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
55. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
56. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
57. Sarafianos, N., Xu, X., Kakadiaris, I.A.: Deep imbalanced attribute classification using visual attention aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 680–697 (2018)
58. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
59. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1238, https://aclanthology.org/P18-1238
60. Shi, H., Hayat, M., Wu, Y., Cai, J.: Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9611–9620 (June 2022)
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.1556
62. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)

63. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
64. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.T., Jin, A., Bos, T., Baker, L., Du, Y., et al.: Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
65. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
67. Wang, Y., Wang, S., Tang, J., Liu, H., Li, B.: Ppp: Joint pointwise and pairwise image label prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
68. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11686–11695 (June 2022)
69. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
70. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
71. Yang, Y., Yih, W.t., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2013–2018 (2015)
72. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
73. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014). https://doi.org/10.1162/tacl_a_00166, https://aclanthology.org/Q14-1006
74. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 192–199 (2014)
75. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
76. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and Yang: Balancing and answering binary visual questions. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
77. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5579–5588 (June 2021)
78. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16793–16803 (June 2022)
79. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16816–16825 (June 2022)
80. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)