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

1 Limitations

One limitation to our proposed method is its increased computational cost. Generative retrieval has n autoregressive text decoding steps, where n is the length of the retrieval template sentence, while contrastive retrieval has one text encoding step. Given the short and fixed-length sentence templates in the attribute learning context, the computational complexity of generative retrieval is $n\times$ constrastive ($n = 2$ to 4). In addition, the text-only attribute embeddings in contrastive retrieval can be precomputed and cached in advance, which would make contrastive retrieval take 0 encoding steps at inference time. This is not possible for generative retrieval, as it is not possible to precompute a part of the likelihood of generating a image-object-attribute triple. Another limitation to the generative retrieval approach is that is is specifically designed for tasks where the assumed lengths of answers or prompts are similar. Since the sum of log probabilities in $L^{(gen)}$ is influenced by the length of the text, the approach is biased towards shorter answers. In the context of attribute prediction tasks, the assumption of similar lengths holds true, allowing us to treat attribute prompt optimization as joint probability optimization in a graph model. This task formulation sets it apart from VQA tasks, which typically involve multiple-choice questions with answers of varying lengths. It is worth noting that this limitation does not undermine our main contribution, which is the development of a novel formulation and framework that connects knowledge from large-scale prefixLM pre-training to the method of generative retrieval for attribute recognition problems.

2 More qualitative examples

We provide more examples to compare our zero-shot retrieval methods, we also include the results from the fully-supervised method SCoNE [14] trained on the VAW dataset. Fig. 1 at the end of the supplementary material shows the results. Some interesting observations can be made. First, VAW is still a closed domain

* Equal contribution.
† Work done at Google.

Table 1: 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 which finetunes the generative prompts. For mA, we report mA@threshold=0.005 as we cross-validated.

Methods	Overall			
	mAP	mR@15	mA	F1@15
ResNet-Bas.-CE	56.4	55.8	50.3	61.5
LSEP	61.0	50.7	67.1	62.3
PartialBCE+GNN	62.3	52.3	68.9	63.9
ResNet-Bas.	63.0	52.1	68.6	63.9
ML-GCN	63.0	52.8	69.5	64.1
sarafianos2018deep	64.6	51.1	68.3	64.6
SCoNE	68.3	58.3	71.5	**70.3**
TAP (w/o in-domain PT)	65.4	54.2	67.2	66.4
TAP (in-domain PT)	73.4	63.3	73.5	71.1
Ours*"{A}{O}"*	70.8	61.8	73.7	68.3
Ours*"{O} is {A}"*	**72.0**	**62.1**	**74.7**	**68.7**
Ours*"{A}{O} is {A}"*	**71.9**	**62.6**	**74.4**	**68.7**

dataset, lacking in the coverage of long-tailed attributes. In example (2), our generative retrieval predicts "decorative", "antique", and "bamboo", which are visually salient and grammatically correct. However, the ground-truth annotation does not include these two options. Second, compared to others, generative retrieval can surface some of the most significant attributes in the examples. For example, "in the background", "decorative", "worn", or "closed". However, many predictions of the contrastive retrieval method are visually imperceptible or incorrect, such as arch-shaped, standing, partially-eaten, water.

3 Additional Evaluation Results

We include additional results on the VAW experiments in Tab. 1, including the less comparable metrics of mR$^{@15}$ and F1$^{@15}$, which were omitted in the main text due to space constraints. Our method achieves the second place only slightly behind TAP, despite focusing more on cross-domain knowledge extraction and not on constructing task-specific models, which may involve fitting to the evaluation dataset at hand using specialized modules, training procedures, or special training data like segmentation masks that are expensive or impossible to scale.

Furthurmore, to qualitatively demonstrate our model's superior performance on the less frequent categories in the distribution long tail of the Medium (72.0% mAP vs 64.8% mAP) and Tail (60.6% mAP vs 48.0% mAP) attribute classes, we show below Tab. 2 of model performance on the least frequent attributes in VAW:

Table 2: Model performance on the least frequent attributes in VAW

Methods	Model	
	SCoNE mAP	Our mAP
nylon	0.6984	0.5333
bell shaped	0.6955	**0.9167**
braided	0.3893	**0.7046**
styrofoam	0.3591	0.3354
spiral	0.2294	**0.8605**
kissing	0.0409	**0.4085**
wallpapered	0.5293	**0.8956**
smoking	0.1966	**0.3671**
stucco	0.3774	**0.5914**
cubed	0.1102	**0.4258**
TAIL MEAN	0.4800	**0.5940**

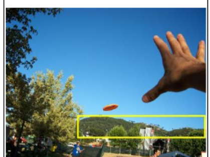
4 Image Attribution

In this paper we display several images from the VAW dataset. The Flickr links and the license information for these images can be found in Tab. 3. We thank the original photographers for sharing their photos.

Table 3: Flickr links and license of the images.

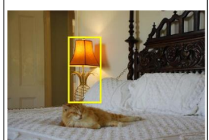
Flickr link	User	License
Paper Fig. 4 (from left to right, top to bottom)		
`flickr.com/photos/mount_otz/31929683/`	mount_otz	CC BY-NC-SA 2.0
`flickr.com/photos/jenny-pics/2381135314/`	jenny-pics	CC BY 2.0
`flickr.com/photos/worldofjan/2984166899/`	worldofjan	CC BY-NC 2.0
`flickr.com/photos/23909838@N02/3363471858/`	23909838@N02	CC BY-SA 2.0
Supplementary materials Fig. 1 (from top to bottom)		
`flickr.com/photos/felipelopez/2660779383/`	felipelopez	CC BY-NC 2.0
`flickr.com/photos/afagen/2269170288/`	afagen	CC BY-NC-SA 2.0
`flickr.com/photos/nbarcet/2172355975/`	nbarcet	CC BY 2.0
`flickr.com/photos/dammit_jack/1523816737/`	dammit_jack	CC BY-NC 2.0
`flickr.com/photos/mjhagen/4347200481/`	mjhagen	CC BY 2.0

(1) Object: mountain
GT Attributes: tree-covered



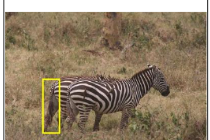
Generative	Contrastive	SCoNE[14]
-21.891 in the background	0.197 arch shaped	**0.998 tree-covered**
-23.587 green	**0.197 tree-covered**	0.994 green
-23.649 for sale	0.196 stucco	0.989 grassy
-23.744 blue	0.196 red striped	0.972 in the background
-24.557 water	0.194 cylindrical	0.963 full
-24.598 small	0.193 partially visible	0.943 far away
-24.677 red	0.193 trimmed	0.901 wide
-24.993 relaxing	0.193 side view	0.893 tall
-25.057 white	0.191 statue	0.688 lush
-25.085 closed	0.191 displayed	0.655 dense
-25.109 orange	0.191 graffitied	0.631 large
-25.17 open	0.19 looking down	0.587 dark
-25.304 clear	0.189 looking up	0.575 rocky
-25.328 in the air	0.189 rolled up	0.434 leafy
-25.396 sleeping	0.187 wallpapered	0.427 small

(2) Object: lamp
GT Attributes: vertical, amber, orange



Generative	Contrastive	SCoNE[14]
-19.623 decorative	0.252 wicker	0.982 standing
-19.842 bronze	0.251 trimmed	**0.979 orange**
-19.907 antique	0.249 displayed	0.973 bright
-19.908 for sale	0.246 tucked in	0.959 illuminated
-20.316 used	0.245 wallpapered	0.946 golden
-20.316 white	0.244 decorative	0.935 shaded
-20.322 wooden	0.244 wispy	0.93 modern
-20.338 golden	0.243 cushioned	0.921 thin
-20.452 small	0.243 resting	0.921 yellow
-20.576 painted	0.242 pinned	**0.901 vertical**
-20.606 open	0.242 pinstriped	0.871 small
-20.715 bamboo	0.241 upholstered	0.778 brown
-20.717 yellow	0.241 buttoned	0.771 rounded
-20.771 on the wall	0.241 unlit	0.701 tall
-20.811 hanging	0.241 bamboo	0.667 tiny

(3) Object: tail
GT Attributes: patterned, spotted, hanging

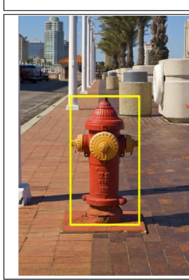


Generative	Contrastive	SCoNE[14]
-22.257 striped	0.276 striped	0.992 hairy
-22.457 spotted	0.257 blue striped	**0.969 hanging**
-22.583 upside down	0.257 barred	0.951 long
-22.66 brown	0.257 red striped	0.915 small
-22.994 running	**0.25 spotted**	0.907 black
-23.208 jumping	0.242 partially eaten	0.898 extended
-23.252 flying	0.241 male	0.897 dark colored
-23.628 walking	0.24 resting	0.896 bushy
-23.692 falling	0.24 lined up	0.896 dark
-23.692 broken	0.24 camouflage	0.878 fluffy
-23.752 white	0.239 hiding	0.803 brown
-23.852 open	0.239 pinstriped	**0.775 patterned**
-24.181 black	0.238 slender	0.768 gray
-24.192 dead	0.238 piled	0.621 curved
-24.303 painted	0.238 horned	0.595 fuzzy

(4) Object: shoes
GT Attributes: athletic

Generative	Contrastive	SCoNE[14]
-24.524 black	0.204 skateboarding	0.998 black
-24.528 broken	0.201 circular	0.987 raised
-24.53 worn	0.2 cylindrical	0.949 used
-24.707 leather	0.198 bell shaped	0.926 dark
-24.89 in the air	0.198 bending	0.844 worn
-24.936 dead	0.197 knocked over	**0.749 athletic**
-24.998 cut	0.197 bent	0.709 leather
-25.037 white	0.196 pulled back	0.669 trimmed
-25.515 old	0.195 pinned	0.666 dark colored
-25.597 used	0.195 holed	0.579 gray
-25.7 flying	0.194 operating	0.567 close
-25.738 falling	0.193 cooked	0.472 shiny
-25.881 painted	0.193 skating	0.456 brown
-25.894 flat	0.192 cutting	0.421 old
-26.036 vintage	0.192 stopped	0.415 wet

(5) Object: hydrant
GT Attributes: tall, clean, close, thin, red, painted, metal, yellow, hard



Generative	Contrastive	SCoNE[14]
-15.886 red	0.302 displayed	0.909 water
-17.218 closed	0.3 light skinned	0.858 buried
-17.521 broken	0.299 tagged	**0.857 metal**
-17.627 painted	0.297 vertical	0.838 colorful
-17.72 for sale	0.297 pinstriped	**0.83 tall**
-17.77 empty	0.296 lined	0.83 old
-17.992 open	0.296 lined up	**0.814 red**
-18.023 old	0.295 modern	**0.806 painted**
-18.031 orange	0.295 docked	**0.766 thin**
-18.285 water	0.295 neat	0.686 standing
-18.309 dead	0.295 amber	0.627 large
-18.377 mounted	**0.295 painted**	0.617 shiny
-18.509 upside down	0.295 old fashioned	0.602 bright
-18.703 funny	0.295 full	0.492 tagged
-18.723 in the background	**0.295 tall**	0.465 dirty

Fig. 1: More qualitative examples on the VAW dataset, **zero-shot vs. fine-tuned**. The generative and contrastive columns use zero-shot retrieval, while the baseline column SCoNE [14] is fine-tuned on the VAW dataset.