Localization *vs*. Semantics: Visual Representations in Unimodal and Multimodal Models

Zhuowan Li[1] **Cihang Xie**[2] **Benjamin Van Durme**[1] **Alan Yuille**[1]
[1] Johns Hopkins University [2] University of California, Santa Cruz

Abstract

Despite the impressive advancements achieved through vision-and-language pretraining, it remains unclear whether this joint learning paradigm can help understand each individual modality. In this work, we conduct a comparative analysis of the visual representations in existing vision-and-language models and vision-only models by probing a broad range of tasks, aiming to assess the quality of the learned representations in a nuanced manner. Interestingly, our empirical observations suggest that vision-and-language models are better at label prediction tasks like object and attribute prediction, while vision-only models are stronger at dense prediction tasks that require more localized information. We hope our study sheds light on the role of language in visual learning, and serves as an empirical guide for various pretrained models. Code will be released at https://github.com/Lizw14/visual_probing.

1 Introduction

The joint learning of vision and language offers mutual benefits. As evident by the recent advancements in vision-and-language pretraining (VLP) models (Radford et al., 2021; Jia et al., 2021; Wang et al., 2022; Singh et al., 2022), they attain not only impressive performance on multi-modal tasks like visual question answering, but also on specialized uni-modal vision tasks like ImageNet classification (Deng et al., 2009), or language tasks GLUE language understanding (Wang et al., 2019).

Despite the superior performance, there is little understanding of *how multimodal learning can help visual representations*. Therefore, we hereby are motivated to compare the visual representations in existing vision-and-language (VL) models and vision-only (V) models from a probing perspective. Specifically, we probe the visual representations through a range of probing tasks that evaluate different properties, including semantics knowledge and localized information, in order to gain a fine-grained understanding of the visual representations. This is inspired by recent works on multimodal feature probing (Ilharco et al., 2021; Zhang et al., 2022), which studies the opposite question to ours, *i.e*., the role of vision in language models.

Fig. 1 illustrates our probing pipeline. We first extract image features using different pretrained models, and then train a simple prediction head to align the model's representation space with the label space of interest. We make the head as simple as possible based on the intuition that less expressive heads can more selectively reflect the quality of the representations (Hewitt and Liang, 2019). The probing is done on various tasks and datasets: object name classification on the Visual Genome dataset (Krishna et al., 2017), attribute prediction on the VAW dataset (Pham et al., 2021), object detection and instance segmentation on the MSCOCO dataset (Lin et al., 2014), and semantic object part segmentation on the PartImageNet dataset (He et al., 2022a). With these probing tasks, we compare vision-and-language pretrained models including OFA (Wang et al., 2022), FLAVA (Singh et al., 2022) and CLIP (Radford et al., 2021) with advanced vision-only models including MAE (He et al., 2022b) and MOCOv3 (Chen et al., 2021).

Interestingly, our experiments suggest that VL models are much better at the label prediction tasks (*e.g*., object class and attribute prediction), while vision-only models are stronger at dense prediction tasks like object detection and segmentation. In other words, multimodal models encode more semantic information in visual representations to better predict fine-grained labels, but fail to enrich the localization information that is required by spatial-aware tasks. This finding is further verified by a more detailed analysis of the segmentation and attribute prediction results, which reveals intriguing properties of the unimodal and multimodal representations.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2378–2390

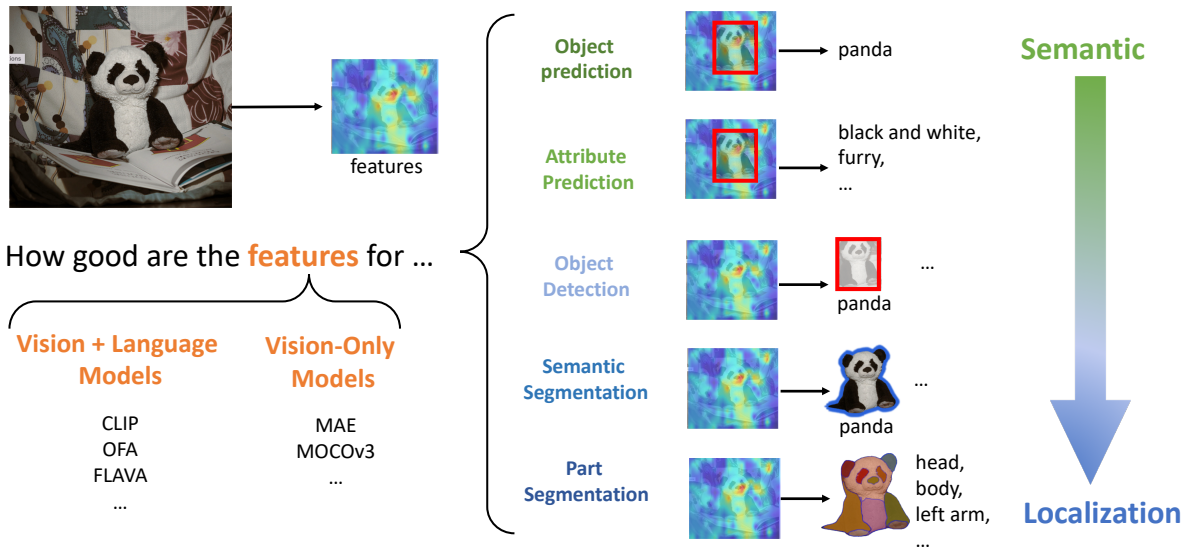

Figure 1: We compare the visual representations from unimodal and multimodal models on five tasks, in order to probe the semantics and localization knowledge encoded in the representations.

In summary, we probe the visual representations in popular VL and vision-only pretrained models on a broad spectrum of tasks and suggest that multi-modal representations encode better semantics. We hope our extensive probing results can serve as a fine-grained benchmark for the publicly released pretrained models, which provides an empirical guide to help researchers choose which model to use for different downstream tasks. Moreover, by offering these insights into the role of language in multi-modal learning, we hope to catalyze future explorations in this direction.

2 Related work

Vision-and-language pretraining (VLP). VLP methods perform well on multi-modal downstream tasks like visual question answering (Antol et al., 2015) and image captioning (Vinyals et al., 2015) and show potential on single-modal tasks. For example, dual encoders trained with a contrastive loss like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) achieve superior visual learning performance. While earlier VLP methods (like LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), OSCAR (Li et al., 2020b), VinVL (Zhang et al., 2021)) rely on image features extracted by separately trained vision models like Faster-RCNN (He et al., 2017) or Resnet (He et al., 2016), more recent works learn the visual features jointly with language. Representative works include OFA (Wang et al., 2022), Florence (Yuan et al., 2021), FLAVA (Singh et al., 2022), UnifiedIO (Lu et al., 2022), CoCa (Yu et al., 2022), and SimVLM (Wang et al., 2021) etc. We refer readers to (Gan et al., 2022) for more details.

Vision and language benefit each other. Several recent works in NLP suggest that multimodal learning can help language understanding. Vokenization (Tan and Bansal, 2020) suggests vision improves the grounding ability of language models. Gordon and Van Durme (2013) shows reduced reporting bias in multimodal world. Z-LaVI (Yang et al., 2022) and VIDLANKD (Tang et al., 2021) show language understanding performance can be improved by better visual imagination or knowledge distillation from videos. Recent work (Zhang et al., 2022) analyzes language and multi-modal models and shows that vision can help language models learn better commonsense knowledge and mitigate reporting bias. However, there is little understanding of the opposite question, *i.e.* how does the visual learning differ in multimodal and unimodal models.

Probing. Probing is a widely used strategy in NLP for interpreting representations (Shi et al., 2016; Belinkov and Glass, 2019). Various works use probing to show that language representations encode a broad range of properties like part-of-speech (Belinkov et al., 2017), syntax (Hewitt and Manning, 2019), semantics (Li et al., 2021), sentence length (Adi et al., 2017), etc., and to compare different language models in those properties (Tenney et al., 2019). Probing has also been adopted to understand multimodal representations in terms of the

capacity for instance retrieval (Ilharco et al., 2021), inter-modality knowledge (Salin et al., 2022), understanding of verbs (Lindström et al., 2020), entity and syntactic grounding (Li et al., 2020a), and visual commonsense knowledge (Zhang et al., 2022), etc. With probing, multi-modal VL models are compared with uni-modal language models to assess the advantage of multi-modal learning. However, probing has not been widely explored for visual representations, despite as a fast on-the-fly metric for model evaluation (Dosovitskiy et al., 2020; He et al., 2022b; Chen et al., 2021) complementary to fine-tuning. To our knowledge, we are the first to compare VL models and vision-only models using probing.

3 Method

To analyze the capacity of the learned representations of different models, we choose a set of tasks to probe the models. For each task, we first extract features using the pretrained models, then we train a simple standard head to predict the results. Mathematically, for every image $I \in \mathbb{R}^{3\times w\times h}$, we extract its features $f \in \mathbb{R}^{C\times W\times H}$ using the off-the-shelf visual encoders in the pretrained models. Here (w, h) is the size of the input image and (C, W, H) is the size of the feature. Then a prediction head P is trained to predict the task-specific results based on feature f. In the whole process, only the head P is trained while the pretrained model (*i.e.*, feature extractor) is frozen.

In this section, we will first describe the probing tasks, datasets and the prediction head for each task (Sec. 3.1), then we describe the evaluated models (Sec. 3.2), and finally how to make the comparison settings fair for every model (Sec. 3.3).

3.1 Probing tasks and datasets

We choose five probing tasks: object name prediction, attribute prediction, object detection, instance segmentation and semantic segmentation for object parts. Among the five tasks, object name and attribute prediction focus more on predicting the semantic labels, while the others are dense prediction tasks that highly rely on spatial information.

Object name prediction. Understanding object names is critical in various multi-modal downstream tasks like VQA and image captioning, in which text descriptions refer to objects by their names. Given an image and a bounding box, object name prediction requires predicting the name of the object in the box. We use the Visual Genome dataset (Krishna et al., 2017) for training and evaluation in this task. Images in Visual Genome mostly come from MSCOCO (Lin et al., 2014) and contain multiple objects. For each object, the annotations provide its bounding box, name and attributes (color, material, etc.). The annotations cover 151 object classes for 1.3M objects in 108k images.

A simple linear classifier is used to predict object names. More specifically, for each object, we first use ROI-Pooling (Ren et al., 2015) to average pool the features according to its box, then use a linear layer on top of the pooled features to predict the name class of the object. Cross entropy loss is used to train the head. Note that the ground-truth bounding box coordinates are provided to the head for both training and testing.

Object attribute prediction. Similar to object name prediction, attribute prediction requires predicting attributes for the object in the given bounding box. As shown in (Zhang et al., 2021), visual features with better-encoded attribute information can substantially improve the performance of multi-modal tasks. This motivates us to treat the attribute as an important axis for evaluating visual representation. The VAW dataset (Pham et al., 2021) is used for object attribute prediction. VAW improves the noisy attribute annotations in Visual Genome. VAW annotates 620 attributes belonging to 8 categories, including color, shape, size, material, texture, action, state, and others. Every attribute is annotated as positive, negative, or unknown for each instance. The annotation covers 260k instances from 72k images, which is a subset of Visual Genome images. Mean average precision (mAP) is used to evaluate the prediction results following (Pham et al., 2021).

Since attribute prediction is formulated as a multi-label classification problem, the prediction head is similar to object name prediction, but has several differences. First, binary cross entropy loss is used for training instead of cross entropy. Second, since the attributes naturally come with a long-tailed distribution, to prevent the rare attributes (*e.g.*, playing) from being overriden by the frequent ones (*e.g.*, black), we assign higher weights to rare attributes and lower weights to frequent ones. Third, for the attributes labeled as unknown, we treat them as negative labels with a small (0.001) weight. Those strategies are borrowed from (Pham et al., 2021).

Task	Dataset	# of classes	Metric	Prediction head
object name prediction	Visual Genome (Krishna et al., 2017)	151	accuracy	linear classifier on ROI features
attribute prediction	VAW (Pham et al., 2021)	620	mAP	linear classifier on ROI features
part semantic segmentation	PartImageNet (He et al., 2022a)	40	mIOU	head from Segmenter (Strudel et al., 2021)
object detection	MSCOCO (Lin et al., 2014)	80	mAP	head from VitDet (Li et al., 2022)
instance segmentation	MSCOCO (Lin et al., 2014)	80	mAP	head from VitDet (Li et al., 2022)

Table 1: The details of {dataset, number of classes, metric, prediction head} for the five probing tasks.

Object detection and instance segmentation. While object name/attribute prediction tests the ability to predict class labels when the object bounding box is given, we are also interested in tasks that focus more on locating the objects. We choose object detection and instance segmentation on MSCOCO (Lin et al., 2014) for this purpose. MSCOCO contains 330K images with 1.5 million object instances in 80 categories. The bounding box and segmentation mask are annotated for each instance. mAP, *i.e.*, mean of average precision for each category, is adopted as the evaluation metric.

Because detection and segmentation cannot be completed using a simple head like a linear layer, we adopt the prediction head in VitDet (Li et al., 2022) as our probing head. While the widely used Mask-RCNN is based on convolutional neural network (CNN) features, Li et al. (2022) propose a variant that is more suitable for non-hierarchical transformer features. Considering the fact that most of our evaluated models are transformer-based, we adopt this VitDet head for probing in our work. Unless specified, all the experiment settings are kept the same as Li et al. (2022).

Part semantic segmentation. While image classification accuracy on ImageNet dataset (Deng et al., 2009) is the most commonly used metric for evaluating visual representations, the recent PartImageNet dataset (He et al., 2022a) provides additional annotations for the ImageNet images, thus enables finer-grained evaluation. PartImageNet annotates segmentation masks of 40 object parts (*e.g.*, head, body, tail) for 11 categories of objects on 24k images. Using this dataset, we perform semantic segmentation of object parts as an additional probing task that requires localization information.

For the segmentation head, we use the mask transformer decoder in Segmenter (Strudel et al., 2021) due to its simplicity and impressive performance on standard datasets. Strudel et al. (2021) adapts transformers for semantic segmentation with the proposed "mask transformer decoder" on top of the embeddings produced by the transformer encoder (standard ViT). In our probing, we replace their transformer encoder with the pretrained models to be evaluated and train the mask transformer decoder to output the semantic segmentation map. Because our goal is to fairly compare different models instead of achieving high performance, we reduce the input image size (from 1024×1024 to 224×224). A linear layer is used to match the feature's dimensions and bilinear upsampling is used to match feature's spatial sizes. All the other training settings are kept the same.

3.2 Evaluated models

We evaluate five models: three representative VL models including CLIP, OFA and FLAVA, and two vision-only models including MAE and MOCOv3. Among the five models, CLIP and MOCOv3 are trained using contrastive loss, while the others are trained with sequence modeling losses. We choose these models because they are representative and highly popular, and their pretrained weights and code are publicly available. In the following, we describe the models, especially their visual components, and how we extract features from them.

CLIP (Radford et al., 2021). CLIP is a dual encoder model trained with contrastive loss using 400M image-text pairs. The image embeddings produced by the image encoder, which can be either a ResNet or a transformer, and the text embeddings produced by the text encoder are trained to be closer with each other in the embedding space when the image and text pair matches. The learned image embeddings are shown to have superior transferability on various downstream tasks. In our study, image features are extracted using the pretrained image encoder.

OFA (Wang et al., 2022). OFA is a unified model that targets both uni-modal and multi-modal tasks. The vision tasks (image classification and object detection), language tasks, and multi-modal tasks (VQA, region/image captioning, visual grounding) are all formulated into a sequence-to-sequence generation problem. In particular, special visual tokens

from discrete-VAE (Van Den Oord et al., 2017; Esser et al., 2021) are used for image infilling and the object bounding box coordinates are also discretized into special tokens. The OFA model first uses a ResNet (Res101 for OFA_{base}) to encode images, then use the transformer encoder and decoder to generate the target sequence from image and text features. Cross entropy loss is used as supervision. OFA is pretrained using 20M image-text pairs with additional uni-modal data. To obtain visual representations, we feed the model with only the image (*i.e.*, empty text), send it through the ResNet, and take the output of the transformer encoder.

FLAVA (Singh et al., 2022). FLAVA is a fully transformer-based unified model. Similar to OFA, the model solves both uni-modal and multi-modal tasks. However, the differences lie in (a) tasks, (b) model architecture, and (c) training loss. (a) FLAVA does not have bounding boxes in the vocabulary, and thus does not support box-related tasks like object detection, visual grounding or region captioning. (b) FLAVA is fully based on transformers; it uses two separate transformer encoders to encode images and texts, then uses several more transformer layers for multi-modal fusion. (c) FLAVA takes multiple losses including CLIP-like contrastive loss, masked image/text/multi-modal modeling losses, and image-text matching loss. FLAVA is pretrained on 70M image and text pairs. We take the output of the visual transformer encoder as image representations.

MAE (He et al., 2022b). Masked Auto-Encoder (MAE) is a self-supervised vision model trained with a masked image modeling task. MAE encodes masked image patches with a transformer encoder and reconstructs the missing pixels with a lightweight decoder trained with MSE loss. Unlike OFA and FLAVA, the reconstruction for MAE happens in the continuous pixel space, which does not require dVAE to generate discretized image tokens. MAE is trained only with ImageNet-1k data and shows promising transfer performance to downstream tasks.

MOCOv3 (Chen et al., 2021). We choose MOCOv3 to represent self-supervised vision transformers trained with contrastive loss. During training, two crops for each image under random data augmentation are encoded by two encoders, a key encoder and a query encoder, into two vectors named "key" and "query" respectively. During training, the goal is to retrieve the corresponding "key" by the "query". Similar to MAE, MOCOv3 is trained using ImageNet-1k.

3.3 Comparison settings

To make the comparison fair, we carefully choose the model size and input size, and ensure different methods are comparable. As probing tasks are highly sensitive to image size and feature's spatial size, for all the models on all the tasks, we fix the input image resolution to be 224*224. We choose this size because 224*224 is the input size for pretraining for all the models except OFA (OFA is pretrained with size 384 for *base* version and 480 for *large*). For dense tasks, although the original detection and segmentation models (*i.e.*, VitDet and Segmenter) use larger input image sizes for better performance, we unify the input size because our goal is to fairly compare models, rather than achieving the best performance.

We find the probing results sensitive to the models' input patch size, because different patch sizes produces features with different spatial sizes.[1] Therefore, considering the availability of pretrained checkpoints with different model sizes and input patch sizes, we try our best to align the feature size and evaluate with the ViT-B/16 backbone by default. Because OFA is not purely transformer-based, we evaluate on the *base* size, which has a ResNet + transformer encoder with 120M parameters (comparable to the 86M ViT-B/16). More details of the evaluated models are shown in Tab. 8.

4 Experiments

4.1 Implementation details

For object name and attribute prediction, the models are trained with a learning rate of 0.001 and batch size of 64 for 200 epochs. We adopt early stopping based on validation performance, then report performance on the test split using the best model. For object detection and segmentation on the COCO dataset, the model is trained for 120k iterations with batch size 20. The learning rate is first set to 8e-5, then decay twice at step 100k and 115k with a factor of 0.1. For part segmentation, we train the model with a learning rate of 0.01 and batch size of 128 for 200 epochs. The validation performance for the final checkpoint is reported.

	Task	VG Obj.	VAW Attr.	COCO Det.	COCO Seg.	Part Seg.	IN1k ft.	IN1k probe
V+L	**OFA**	**57.13**	**61.67**	<u>25.04</u>	<u>19.38</u>	33.11	82.2	-
	FLAVA	<u>54.29</u>	<u>61.51</u>	21.06	17.20	34.77	-	75.5
	CLIP	51.54	61.15	19.55	15.56	<u>40.61</u>	-	**80.2**
V	**MAE**	49.52	52.59	**25.29**	**22.05**	**42.30**	**83.6**	68.0
	MOCOv3	47.81	54.44	20.31	16.96	40.11	<u>83.2</u>	<u>76.7</u>

Table 2: Probing results on five tasks. VL models perform better on label prediction tasks, while vision-only models perform better on dense prediction tasks. Finetuning and linear probing results on ImageNet for each model (cited from original papers) are also shown for reference. The best and the second best scores are in **bold** and <u>underlined</u>.

		MSCOCO			PartImageNet		
		mAP	Semantic	Localization	mIOU	Semantic	Localization
V+L	**OFA**	19.38	60.02	17.41	33.11	71.71	84.15
	FLAVA	17.20	61.48	14.67	34.77	75.28	83.76
	CLIP	15.56	**68.24**	13.25	40.61	**80.21**	86.80
V	**MAE**	**22.05**	46.85	**20.69**	**42.30**	75.03	**89.50**
	MOCOv3	16.96	49.80	15.08	40.11	76.18	86.08

Table 3: Detailed analysis of instance segmentation and part segmentation results. We evaluate the segmentation results (standard metric mAP, mIOU) from two additional perspectives: semantics (F1 score for semantic class prediction) and localization (mAP/mIOU for foreground/background segmentation). While V models are better on the standard metrics, VL models are better when evaluated with semantics metrics.

4.2 Probing results

We probe the five models on each of the five probing tasks. We make sure that the experiment settings, including model size, input size, training protocol and data splits, are well aligned for every model in order to make fair comparisons. The probing results are shown in Tab. 2. We also include the ImageNet finetuning accuracy and linear probing accuracy of each model for reference, because they are widely-used metrics for model evaluation. On each task, we compare the VL models and V models. Note that the evaluation metric for each task is different (as in Tab. 1), performance on different tasks cannot be compared and we only compare numbers in each column separately.

For object name prediction and attribute prediction, VL models consistently perform better than V models. For object name prediction on Visual Genome, VL models all achieve more than 51% accuracy while V models get accuracy less than 50%; for attribute prediction on VAW, mAP for VL models are higher than 61% while lower than 55% for V models. This suggests that representations from VL models capture richer semantic information about the objects in each image, which can be decoded using a simple linear layer. In contrast, in V models the name and attribute information are not explicit enough.

For the dense prediction tasks, MAE performs the best on all three tasks. For part semantic segmentation on PartImageNet, MOCOv3 and CLIP also get decent performance ($> 40\%$) that is close to MAE (42%), while the other two VL models are lower by a large margin ($< 35\%$). For object detection on MSCOCO, OFA gets close mAP (25.0) to MAE (25.3) while the performance of the other three models are much lower; however, when it comes to instance segmentation, the advantage of MAE is more clear, surpassing all the other models with a margin larger than 2.7%.

Interestingly, comparing the object detection and instance segmentation results on COCO, we find that the performance drops of V models are consistently smaller than VL models, which indicates that V models learn better localized representations.[2] For example, for OFA, the mAP for segmentation is 5.7% (25.04-19.38) lower than that for detection; while the drop MAE and MOCOv3 are smaller

[1] *E.g.* for input images of 224*224, ViT-B/16 produces visual representations with size 768*14*14, while ViT-B/14 gives feature size 768*16*16, which will affect probing.

[2] Both the metrics and the datasets are the same for instance segmentation and detection, thus the results can be compared. The only difference between mAP for detection and instance segmentation is that when calculating overlaps between predictions and ground truths, one uses the pixel-wise IOU (intersection-over-union) rather than bounding box IOU.

(3.2%, 3.3%). Because segmentation requires more localized features than detection to find the boundary of objects, the performance gap between detection and segmentation can be an indicator of the localized information in the representations, considering those two tasks are based on the same dataset. With the more-localized representations, the model can better predict the mask boundary. Therefore, the smaller gap of vision-only models suggests they learn more localized representations.

To further verify this finding, we next take a closer look into segmentation results, which more clearly compare the semantics and localization information in different models.

A closer look at the segmentation results. We evaluate the instance segmentation results on COCO and semantic segmentation results on PartImageNet using two more metrics: (a) the label prediction metric, and (b) the foreground-background segmentation metric, where (a) is an indicator for semantics and (b) for localization. The motivation is that the segmentation metrics (mAP for instance segmentation, mIOU for semantic segmentation) require correctly predicting both the class label and the boundary, so the quality of both determines the score. Therefore, we propose two additional metrics to measure the two factors separately. For (a), for each image, we transform its predicted segmentation map into label predictions, and evaluate the quality using the multi-label prediction metric. In particular, we treat the appeared classes in the segmentation map as positive labels and the others as negative; then the label predictions are evaluated using the F1 score. F1 score is defined as $\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$, where precision and recall are averaged over label classes. For (b), we merge all the different object categories and process the segmentation map into binary labels, *i.e.*, foreground and background, then report the mIOU (for instance segmentation) or mAP (for semantic segmentation) of the binary segmentation maps.

Tab. 3 shows the segmentation results on COCO and PartImageNet evaluated using the above two metrics. Although MAE achieves the best performance on both datasets, when looking at the semantic and localization results, we find that its advantage mainly comes from better localization, rather than semantics. In terms of semantics, VL models perform much better than MAE. For example, on the MSCOCO dataset, VL models achieve F1 scores higher than 60, while MAE and MOCOv3 are lower than 50. The results suggest that while MAE is better at finding the object boundaries when predicting segmentation masks, VL models are better at predicting labels for the objects.

In Fig. 2, we show several examples of the part segmentation results on PartImageNet. In the examples, MAE captures the object's shape more accurately, like the curly snake body, the shark's small fin, and the quadruped contour. However, MAE and MOCOv3 make more mistakes in labeling the regions compared to VL models. For example, MAE wrongly predicts the shark fin as a reptile foot, and the quadruped as a reptile; MOCOv3 confuses the quadruped head and foot as the fish head and fins. Those examples more explicitly compare the semantics and localization knowledge learned by VL and V models.

Analysis on different attribute groups. We further decompose the attribute prediction results into different attribute groups. In the VAW dataset, attributes are categorized into 8 groups: action, texture, shape, size, color, material, state, and others. The results are shown in Fig. 3. Interestingly, despite the overall better results of VL models, we find that their advantages differ in different groups. For example, the gap between VL and V models in the "action" category is more significant than in the "texture" category. Intuitively, "action" is less visually grounded then "texture" requires more context and semantic information, on which VL models is better at, suggesting that while vision-only ones are better at predicting highly visually grounded local attributes (*e.g.*, texture), VL models are better at more abstract ones.

4.3 More analysis

Findings of contrastive training. The results also show that contrastive models perform relatively better on localization for single-object images than multi-object images. Among the five tasks, part segmentation on PartImageNet dataset are based on single-object images from ImageNet, while the other four tasks are based on COCO-style multi-object images. In Tab. 3, comparing the contrastively trained models (CLIP, MOCOv3) and the models trained with sequence modeling objectives (OFA, FLAVA, MAE), we find that contrastive models perform relatively better on PartImageNet than MSCOCO. For example, on PartImageNet, CLIP outperforms the other two VL models (*i.e.*, OFA and FLAVA) by a large margin (more than

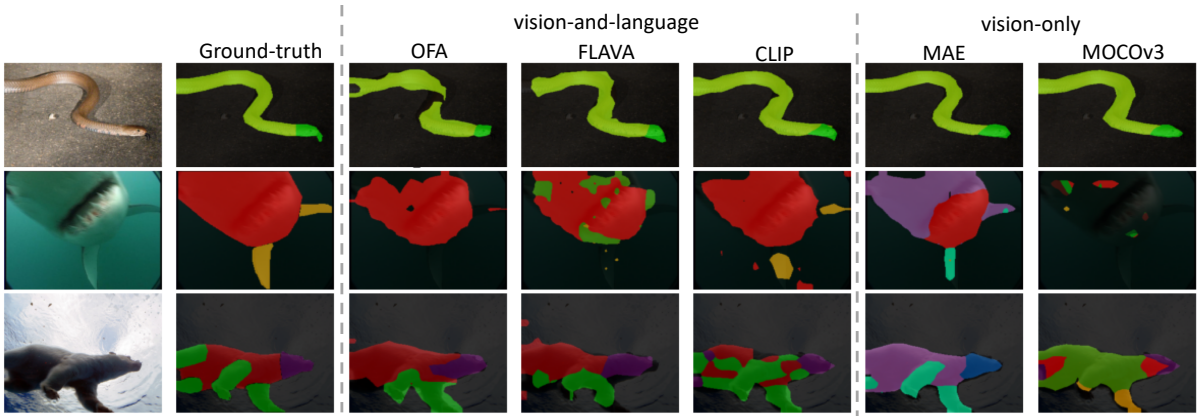

Figure 2: Compared to vision-and-language models, vision-only models more accurately predict the boundary of segmentation masks, but make mistakes in labeling the regions.

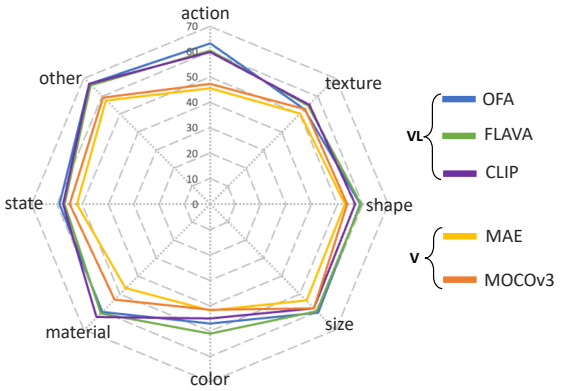

Figure 3: A closer look at the attribute prediction results by separately evaluating different types of attributes. The advantage of VL models is more significant in the more abstract categories (*e.g.*, *action*) than visually grounded categories (*e.g.*, *texture*).

6% mIOU); on MSCOCO, it under-performs them. The semantic and localization evaluation suggests that this difference is mainly caused by localization, *e.g.*, the localization results of CLIP is much better than OFA and FLAVA on PartImageNet. A similar observation can be obtained by comparing MOCOv3 and MAE: although MOCOv3 underperforms MAE on both datasets, the gap is much smaller on PartImageNet than MSCOCO (2.2 vs. 5.1). Therefore, we suggest that the localization ability of contrastive models is relatively stronger on single-object images.

The effect of model size. To study the effect of model size, in Tab. 4, we show the probing results with size *base* and *large* for MAE and OFA. For MAE, a larger model size improves performance on all the probing tasks in parallel for 1% to 2%. However, note that this improvement is less significant compared to the big gaps between different model types. For OFA, except for the marginal improvement in attribute prediction, the larger model size hurts probing results on the other four tasks. The reason for the decrease is that the OFA_{large} is pretrained with a larger input image size (480*480) compared with OFA_{base} model (384*384). Because we probe all models with the same image size (224*224) for a fair comparison, the gap in image size between pretraining and probing is more significant for OFA_{large}. In summary, the effect of model size is less considerable than other factors like model type or input image size.

	obj.	attr.	det.	seg.	p-seg.
MAE_{base}	49.52	52.59	25.29	22.05	42.30
MAE_{large}	**51.91**	**53.38**	**29.67**	**25.63**	**44.85**
OFA_{base}	**57.13**	61.67	**25.04**	**19.38**	**33.11**
OFA_{large}	52.33	**62.01**	21.23	16.51	32.04

Table 4: The influence of model size is less considerable than other factors like model type.

The effect of downstream finetuning. Tab. 5 compares probing results of models with and without finetuning on downstream tasks. For MAE, the results are based on the *base* size; for OFA, the results are on *large* size, due to the availability of publicly released model checkpoints. For both models, finetuning on image classification on ImageNet-1k and VQA on VQAv2 hurts the probing performance to varying degrees (except for attribute prediction). This indicates that while in pretraining, the model learns features that capture various fine-grained information about the image, during finetuning towards a specific task, only information useful for the task is kept and other information is dropped. Moreover, compared with ImageNet finetuning,

finetuning on VQA leads to a much smaller performance decrease in probing results, suggesting that the change in probing results depends on the nature of downstream tasks. In this case, VQA requires more fine-grained information about objects, attributes, etc., resulting in a smaller drop than ImageNet finetuning.

	obj.	attr.	det.	seg.	p-seg.
MAE	**49.52**	52.59	**25.29**	**22.05**	**42.30**
MAE$_{IN1k}$	45.16	**53.82**	21.41	17.74	35.62
OFA	**52.33**	62.01	**21.23**	**16.51**	**32.04**
OFA$_{IN1k}$	50.54	60.74	18.91	14.67	27.56
OFA$_{VQA}$	51.42	**63.40**	19.01	14.22	28.34

Table 5: Probing results of models finetuned on downstream tasks. Finetuning hurts the probing performance in most cases.

5 Conclusion

This work compares the visual representations in multimodal and unimodal models by feature probing. By comparing three representative VL models and two V models on five probing tasks, we find that VL models are stronger in label prediction tasks, while vision-only models are better in dense prediction tasks. We hope our diagnostic findings serve as an empirical guidance for future works in choosing models for different downstream tasks, as well as exploring the role of language in visual representation learning.

6 Limitations

This study is limited by the coverage of pretrained models. We only evaluate models which have publicly accessible checkpoints, and which can be aligned in terms of model sizes, patch sizes, etc. Because we do not have enough computational resources to retrain the models, our comparisons are restricted by the released ones. In addition, we are aware that the evaluated models are not well-aligned on many aspects, like the training data, model architecture, training objectives and hyperparameters, etc. However, aligning those components requires significant amount of GPU resources and training effort. With the limitations, we evaluated the released model checkpoints and hope our results can serve as empirical analysis for future researchers.

Acknowledgements

This work is supported by ONR N00014-23-1-2641, as well as a gift funding from the JHU + Amazon Initiative for Interactive AI. This work is also supported with Cloud TPUs from Google's TPU Research Cloud (TRC) program. We would like to thank Elias Stengel-Eskin, Kate Sanders, David Etter, Reno Kriz, Chen Wei, as well as the anonymous reviewers, for their helpful comments.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Xinlei Chen, Saining Xie, and Kaiming He. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. 2022a. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022b. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

J Hewitt and P Liang. 2019. Designing and interpreting probes with control tasks. *Proceedings of the 2019 Con*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2021. Probing contextual language models for common ground with visual representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5367–5377.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020a. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *AAAI 2022*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-lavi: Zero-shot language solver fueled by visual imagination. *arXiv preprint arXiv:2210.12261*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of NAACL-HLT*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

A Appendix

Tab. 6 shows the standard deviations when repeating experiments for 3 times, which shows the significance of the probing results. Tab. 7 shows the numerical numbers for Fig. 3. Tab. 8 compares the details of the evaluated models, in terms of the feature sizes, model architectures, training data and objectives.

	COCO det	**COCO seg**
OFA	25.06 ± 0.02	19.37 ± 0.01
MAE	25.30 ± 0.02	22.03 ± 0.05

Table 6: Standard deviations for 3 repeated experiment runs.

		all	color	material	shape	size	action	state	texture	other
V+L	**OFA**	61.67	47.10	59.85	59.13	60.27	63.35	59.18	52.73	66.88
	FLAVA	61.51	50.94	60.71	59.42	59.61	60.39	57.05	54.69	66.10
	CLIP	61.15	44.93	63.04	57.17	57.88	59.71	57.60	55.34	66.96
V	**MAE**	52.59	41.77	46.80	53.07	53.60	45.55	52.27	50.31	57.77
	MOCOv3	54.44	41.47	53.08	53.85	57.93	47.25	55.02	52.91	59.32

Table 7: Detailed attributes prediction results corresponding to Fig. 3.

	OFA	**FLAVA**	**CLIP**	**MAE**	**MOCOv3**
Feature size	768*14*14	768*14*14	768*14*14	768*14*14	768*14*14
Architecture	ResNet blocks + transformer encoder + transformer decoder	ViT + transformer text encoder + multimodal encoder + heads for different tasks	ViT + transformer text encoder	ViT + transformer decoder	ViT
Visual feature extractor	ResNet blocks + transformer encoder	ViT-B/16	ViT-B/16	ViT-B/16	ViT-B/16
Data	25M pairs + unpaired	70M pairs + unpaired	400M pairs	1.2M images	1.2M images
Data source	CC, VQA, GQA, RefCOCO, ImageNet-21k, OpenImages, Piles...	CC12M, YFCC, VG, COCO, ImageNet-1k, CCNews, BookCorpus...	Unknown (Internet)	ImageNet-1k	ImageNet-1k
Training task	multiple tasks with a unified next-token prediction loss	contrastive + image text matching + masked multimodal modeling + masked image modeling (dVAE) + masked language modeling	contrastive	masked image modeling (MSE)	contrastive

Table 8: Details of the compared models.