“This is my unicorn, Fluffy”: Personalizing frozen vision-language representations

Niv Cohen[1,2], Rinon Gal[1,3], Eli A. Meirom[1], Gal Chechik[1,4], and Yuval Atzmon[1]

[1] NVIDIA Research, Israel
[2] The Hebrew University of Jerusalem, Jerusalem, Israel
[3] Tel Aviv University, Israel
[4] Bar Ilan University, Israel

Abstract. Large Vision & Language models pretrained on web-scale data provide representations that are invaluable for numerous V&L problems. However, it is unclear how they can be extended to reason about *user-specific* visual concepts in unstructured language. This problem arises in multiple domains, from personalized image retrieval to personalized interaction with smart devices. We introduce a new learning setup called *Personalized Vision & Language* (PerVL) with two new benchmark datasets for retrieving and segmenting user-specific (“personalized”) concepts “in the wild”. In PerVL, one should learn personalized concepts (1) *independently* of the downstream task (2) allowing a pretrained model to reason about them with free language, and (3) without providing personalized negative examples. We propose an architecture for solving PerVL that operates by *expanding* the input vocabulary of a pretrained model with new word embeddings for the personalized concepts. The model can then simply employ them as part of a sentence. We demonstrate that our approach learns personalized visual concepts from a few examples and effectively applies them in image retrieval and semantic segmentation using rich textual queries. For example the model improves MRR by 51.1% (28.4% vs 18.8%) compared to the strongest baseline.
The code and benchmark are available on github under NVlabs/PALAVRA and NVlabs/PerVLBenchmark.

1 Introduction

Large Vision & Language (V&L) models pre-trained on web-scale data made a breakthrough in computer vision [50, 70, 6]. These models provide a multimodal vision-language representation, and are used in a multitude of downstream tasks, from image captioning [47] and video retrieval [19], through image generation [22, 48] and segmentation [71, 36], to robotic manipulation [54]. All these tasks benefit from the “open-world” capabilities of large V&L models, enabling the use of rich, free-form text with a long “tail” vocabulary of visual categories.

However, even with these powerful representations, an important challenge remains: How can these models be leveraged to reason about *user-specific*, “*personalized*” object instances in open-world vision problems? For example, we may

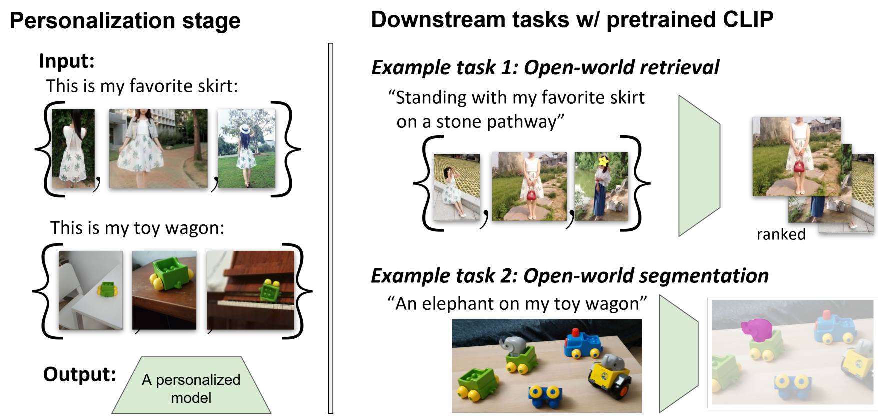

Fig. 1. The Personalized Vision & Language (PerVL) learning setup. **Left:** A user provides a few image examples of their personalized visual concepts: a favorite skirt (top), or a toddler's toy wagon (bottom). Examples are provided independently of the downstream tasks. **(Right)** the personalized model can be used in various downstream tasks. **Top-right:** An image retrieval task: given a textual query and a collection of images, rank and retrieve the queried image. **(Bottom-right)** Open-world semantic segmentation task. Segment a personalized object referred by a textual query. This example illustrates multiple ambiguities. First, there are two wagons that carry an elephant. Second, there are two wagons that correspond to the personalized concept. Resolving the ambiguity requires reasoning in both the visual and text modalities.

wish to find an image that portrays us wearing a *specific* sweater, ask a robot assistant to make us coffee in *our* "best-mom mug", or synthesize an image of our child's treasured toy Fluffy in an entirely new context.

Clearly, pretrained V&L models cannot be used directly to reason about new personal items. Luckily, it is typically easy for a user to collect a few image examples for a personalized concept. It then remains to develop methods that extend the pretrained models to new concepts using these examples. One challenge is that typically it is easy for people to provide positive image examples for a concept, but harder to provide consistent negative distractor examples [29, 53].

Learning from a few examples is considered a hallmark of intelligence. When people learn novel concepts from a few examples [44, 7, 34, 45, 46], they can seamlessly employ them in their semantic mental state and *reason jointly* both over the personalized concepts and over a large body of prior knowledge. Could a computational approach learn in a similar way using a pretrained V&L model?

Previous efforts [56, 73, 23, 76] focused on learning a transformation module on top of CLIP's output space. However, as we explain below, these approaches risk forgetting prior knowledge, or face difficulties in accessing it concurrently with newly learned concepts. In addition, these previous approaches take a multiclass approach, discriminating between several new concepts. They are not designed for learning a single new personalized concept, which is natural in the context of personalization. Therefore, it is unknown how to learn a single personalized concept from few image examples in a way that (1) allows the pretrained model to reason about the concept with free language, and (2) uses only "positive" image examples of the target concept.

Here, we address the question of personalizeing a pretrained model using few samples while maintaining its performance on the original vocabulary. We study a new representation learning setup, which we call "*Personalized Vision & Language*" (PerVL) (Fig. 1). In PerVL, we are given a pretrained V&L model, one or more personalized visual concepts, a few training images of each concept, and a string describing the concept type, like "a mug" or "a short sleeve top". The goal is to learn a representation that can later be used to solve a set of downstream V&L tasks involving the personalized concept. No further supervision is given for these downstream tasks. PerVL arises in various scenarios. In image retrieval, a user may tag a few of their images and wish to retrieve other photos of that concept in a visual specific context [10, 2]; in human-robot interaction, a worker may show a specific tool to a robotic arm, and instruct how to use it [54, 64, 42]; in video applications, an operator may search for a specific known item in the context of other items or people doing activities that are described with language.

Unlike previous efforts, instead of modifying a V&L model *output*, we propose a framework for expanding its *input* vocabulary. Specifically, we learn new word embeddings for the new personalized concepts by intervening with the model input space. The concept of "my best-mom mug" would be associated with a new symbol [MY BEST-MOM MUG] that has its own *dense* word embedding. The model could later represent sentences that use it, like "*Sipping tea from my best-mom mug on a porch*" by detecting "my best mom mug" and mapping its symbol [MY BEST-MOM MUG] to its new embedding vector. Such tokens trivially preserve the structure of the original model, since the encoder model itself remains unmodified. Moreover, as we show below, the new concepts can also be easily integrated into existing downstream V&L tasks. In summary, we address the question of using a small number of samples to personalize a pretrained V&L model, while maintaining its performance in the original vocabulary.

This paper makes the following novel contributions: (1) A new representation learning setup, *PerVL*, for personalizing V&L representations, while keeping their "zero-shot" reasoning capabilities. (2) Two new benchmark datasets for *PerVL*. (3) A novel approach, PALAVRA[7], to *expand* and personalize the vocabulary of the V&L representation *inputs*. PALAVRA uses a cycle-consistent loss, learned with positive image examples only. (4) A technique for using a *textual* encoder to improve the generalization of a network to new *visual* concepts.

2 Related work

The success of CLIP led to diverse work that leverage its powerful representation for few-shot learning. Most works [56, 73, 23, 43] are based on learning a residual

[7] Palavra means "word" in Portuguese, *as we learn new word-embeddings.* For acronym lovers, PALAVRA also stands for "Personalizing LAnguage Vision RepresentAtions"

"adapter" layer [28] over the output of CLIP encoders. Taking a different approach, [76] proposes learning a soft prefix to improve accuracy in a classification task. Our work differs from these approaches in two key aspects: (1) They focus solely on classifying images using a narrow vocabulary. In contrast, our setup learns a representation which is then used in any downstream tasks. Moreover, our method *expands* CLIP's vocabulary rather than narrowing it. (2) Adapter-based methods override the output representation of the encoders, leading to a change in their input $\rightarrow$ output mappings. Our method does not change the pretrained mapping but enriches its input vocabulary with new concepts.

Recently, [65, 32, 26] have shown that fine-tuning can actively harm out-of-distribution generalization, even when tested on the same downstream task for which the model was tuned. Our method does not fine-tune the pretrained model and does not leverage in-distribution labeled examples for the downstream tasks.

Other approaches [61, 27] study "fast" concept learning combined with "slow-learned" concepts, showing that the new concepts can be applied to "slowly-learned" downstream tasks. However, the "fast" learned concepts are stored implicitly in the network activations, rather than grounded in the vocabulary.

A related set of studies can be found in tasks of image captioning and generator inversion. These studies extract meaningful semantic information from images and map them to tokens that represent the concepts, such as words or latent codes. Of these, [13, 17, 20, 15, 30, 40, 55] focus on personalizing image captions to a user writing style. Alternatively, [25, 62, 67, 75, 41, 16] extend image captions with novel concepts using "slot filling".

Our model differs from zero and few-shot learning (FSL) based on meta-learning [21, 63, 59, 57, 11, 35, 1, 68, 4, 5, 49] or incremental learning [60, 18, 51, 12, 31, 66] in three aspects. First, we impose stronger generalization requirements. Our model can reason about new concepts in diverse downstream tasks, which may be unknown at training time. Second, in common FSL, the concept distribution used during meta-learning ("support set") is also used during the FSL stage. For example, meta-learn on birds, then do FSL with new types of birds. While our technique for training with text allows to generalize beyond the domain of concepts in the training images. Third, our approach improves upon CLIP's zero-shot perceptual capabilities, and is compatible with many CLIP-based downstream tasks.

Finally, in existing FSL benchmarks [63, 52, 74] there is *no* instance level annotations, and there is only a single task. As a result, existing FSL benchmarks do not directly fit our setting. Our work addresses *rich text* query *of a specific instance*, that can be used in a flexible way with many downstream tasks.

3 A new setup, *Personalized Vision & Language*

We propose "*Personalized Vision & Language*" (PerVL), a new representation learning setup, to personalize a pretrained model with few positive image examples, without supervision for the downstream task.

In PerVL , we are given a pretrained model $h(S_V, I)$ that accepts a sentence S and an image I. The sentences that the model accepts are defined in a vocabulary V. We wish to update h so that it can accept sentences from an expanded vocabulary $V' = V \cup C$ where C is a new set of concepts $C = \{c_1, ...c_k\}$, which results in an extended model $h'(S_{V'}, I)$. In general, we expect that adapting the model would not strongly affect the original vocabulary, namely $h'(S_V, I) \approx h(S_V, I)$.

At training (personalization) time, we adapt the model given a small set of images $\{I_i\}_{i=1}^{N_c}$ for every concept c, without assuming access to negative training *images*. We are also provided with a string describing the type of the new concept, such as a "mug" or a "short sleeve top". Stating the type is a natural way for non-expert users to provide prior knowledge about the personalized concept. The type can be used to guide learning to distinguish the personalized concept from the general concept type. Concepts describing coarser classes from a hierarchy of concepts (e.g. "dog" for "poodle") have been used for this purpose [14]. We denote the concept type by S_c.

During inference, we are given a downstream V&L task T that can be inferred using the pretrained model h for the vocabulary V, and we wish to solve it for an instance x that contains the new concept c. The instance may contain images and sentences pertaining to c.

Encoder PerVL : Here we focus on the special case of CLIP [50]. The model h applies a cosine similarity between a sentence S and an image I: $h(S, I) = \cos(h^{\mathcal{T}}(S), h^{\mathcal{I}}(I))$, where $h^{\mathcal{I}}$ and $h^{\mathcal{T}}$ are CLIP image and text encoders.

4 Methods

Before describing our approach, we first explain the reasons for expanding the *input* vocabulary of the V&L model and how it differs from previous approaches.

Several studies extend CLIP by learning an "Adapter" module on top of the CLIP representation [23, 56, 73, 43, 28]. That module is applied to the *output* of a CLIP encoder network that is kept frozen. It is trained for a classification task with labeled data and a templated text query ("a photo of a [concept-type]").

We show below (Sec. 6 & Appendix B) that this approach tends to be brittle and fails when its input sentences deviate from the template used for training. This is probably because the adapter *overrides* the output representation of the encoder, so training it with very few examples hurts its generalization power.

Conversely, our approach does not overrides the encoder outputs. Our working hypothesis is that the text input space of a web-scale V&L model is rich enough for reasoning about new personalized concepts. We just need to find the right word embedding representation for any new personalized concept. We illustrate this architectural distinction in Fig. 2.

Finally, one could fully retrain a CLIP model with the expanded vocabulary set. However, retraining CLIP requires ~$400M$ images. Our approach is trained with a tiny fraction of that, $< 1M$ samples, and once it is trained, different users can use it, each with their own vocabulary.

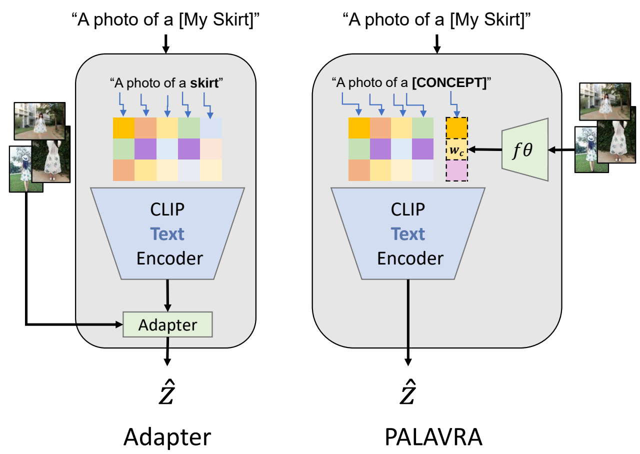

Fig. 2. Visualization of an adapter-based approach (left) and PALAVRA (right). Adapters change CLIP's output space by appending additional layers following the encoder. Our method defines new tokens in CLIP's existing input space, leaving the output space unchanged.

Notation: For brevity, we describe adding a single concept c. Adding multiple concepts can be done iteratively. We use the notation [CONCEPT] to refer to a learned concept (c) within a textual query. $\mathcal{I}$ denotes the CLIP embedded image space, $\mathcal{T}$ the CLIP embedded textual space, $z_k = h^{\mathcal{I}}(I_k)$ is the embedding of an image I_k into $\mathcal{I}$, and similarly $h^{\mathcal{T}}(S)$ is the embedding of a sentence S into $\mathcal{T}$. Finally, $\mathcal{W}$ denotes the space used to embed input word tokens into CLIP.
Architecture and Workflow: At a high level, our workflow has three steps.

(1) Learn an inversion mapping f_θ from a *set* of points in CLIP image space $\mathcal{I}$ to a point in its word embedding *input* space $\mathcal{W}$ (Fig. 3). Formally, $f_\theta : \{z_k \in \mathcal{I}\}_{k=1}^{K} \to \mathcal{W}$. It is trained with *non-personalized, large-scale* data.

(2) Initial personalization (Fig. 4). Learn a word embedding $\mathbf{w}_c$ of a new personalized concept c. Thus, given a set of image examples $I_1, ..., I_K$ we map them to CLIP image space, then map them using f_θ to obtain an initial word embedding $\mathbf{w}_c^0 = f_\theta(\{h^{\mathcal{I}}(I_k)\})$. Formally, $\{I_k\}_{k=1}^{K} \to \{h^{\mathcal{I}}(I_k)\}_{k=1}^{K} \to \mathbf{w}_c^0 \in \mathcal{W}$.

(3) Fine-tuning. The initial embedding $\mathbf{w}_c^0$ is then updated using gradient steps to maximize the similarity of the template text embeddings to the image examples, while contrasting it with an embedding of a "super-concept".

Next, we describe the learning of each component in more detail.

4.1 Learning the inversion mapping f_θ

We now describe how we learn an "inversion" map f_θ from a set of points in CLIP space $z_1, ..., z_k \in \mathcal{I}$, to a word embedding $\mathbf{w}_c^0 \in \mathcal{W}$, where W is the input space of the language encoder. We base f_θ's architecture on "Deep Sets" [72]. We now discuss the loss and how to train f_θ with two types of *large-scale, non-personalized* data: images and text. See Fig. 3.

A contrastive cycle loss. f_θ maps from CLIP space to $\mathbf{w}_c^0 \in \mathcal{W}$. Then, by pairing $\mathbf{w}_c^0$ to the word embedding for [CONCEPT] we can feed $\mathbf{w}_c^0$ into $h^{\mathcal{T}}$ with a template sentence T_c like "A photo of a [CONCEPT]". We can then define a cycle consistency loss to match the input of f_θ with the output of $h^{\mathcal{T}}$ (see Fig.

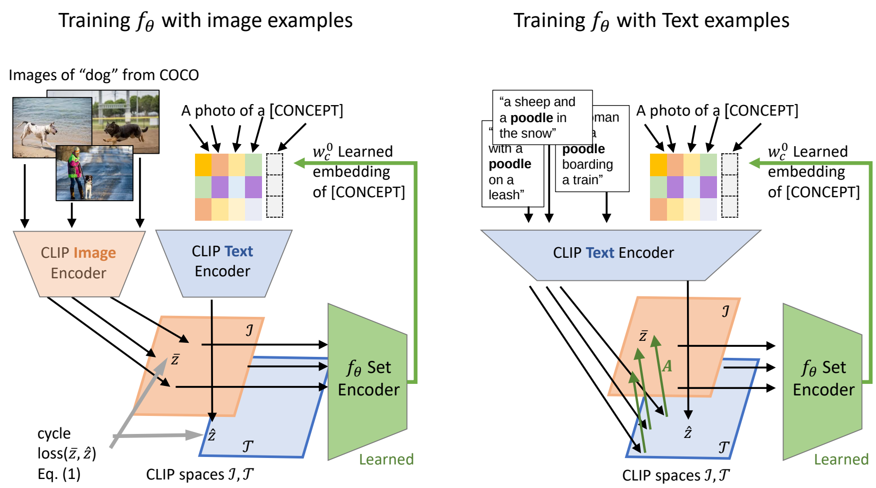

Fig. 3. Architecture outline: Learning f_θ. We start with a *large-scale-data* training step. A set encoder f_θ is trained to map CLIP-space output embeddings to a code in CLIP's input space. It is alternatingly trained with a batch of either image examples (left), or sentence examples (right) with augmented concept types. We use a cycle loss by mapping the code back to CLIP's output embedding, using a template sentence.

3 left). Specifically, let $\bar{z}_c$ be the average over samples in $\mathcal{I}$ from the concept c, $\bar{z}_c = \sum_{k=1}^{K} z_k/K$ and let $\hat{z}_c$ be the CLIP embedding of a template sentence, $\hat{z}_c = h^{\mathcal{T}}(T_c)$. We wish to tune f_θ so that $\hat{z}_c$ is close to $\bar{z}_c$ for the concept c and far from other concepts. We therefore define a symmetric contrastive loss for a concept c, with a formulation similar to SimCLR [9]:

$$\begin{aligned} \ell_{Cycle}\big(c; \{\bar{z}_{c'}, \hat{z}_{c'}\}_{c'=1}^{C}\big) = &-\log \frac{e^{\cos(\bar{z}_c, \hat{z}_c)}}{\sum_{c'=1}^{C} e^{\cos(\bar{z}_c, \hat{z}_{c'})} + \sum_{c' \neq c} e^{\cos(\hat{z}_c, \hat{z}_{c'})}} \\ &-\log \frac{e^{\cos(\bar{z}_c, \hat{z}_c)}}{\sum_{c'=1}^{C} e^{\cos(\hat{z}_c, \bar{z}_{c'})} + \sum_{c' \neq c} e^{\cos(\bar{z}_c, \bar{z}_{c'})}}, \end{aligned} \tag{1}$$

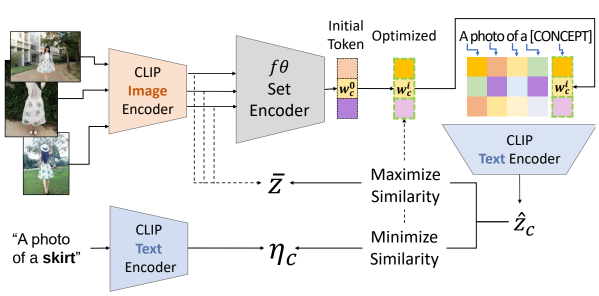

Fig. 4. Architecture outline: Personalization. Given a set of examples of a personalized concept and its type (skirt), we embed them with CLIP and predict an initial code ($\mathbf{w}_0$) for the concept using a frozen f_θ. We then further tune the code with a contrastive loss.

where $\cos(\hat{z}, \bar{z})$ denotes cosine similarity, C is the number of concepts in a batch.

We also use a regularization term ℓ_{GT} that maximizes the similarity of the predicted $\mathbf{w}_c^0$ with its ground truth. See details in the Appendix. Finally, the cycle loss and ground-truth regularization terms are combined with a hyperparameter $\lambda_{gt} \geq 0$, and the total loss is $\ell_{total} = \ell_{Cycle} + \lambda_{gt} \cdot \ell_{GT}$.

Training f_θ with images. We use a variant of COCO [38] that extracted the subject and object from each caption as in [3], and take the 1000 most frequent concepts. In every training batch, we draw at random C concepts, then draw K images for each concept. We then map them to CLIP image space, yielding $\{z_k\}_1^K = \{h^{\mathcal{I}}(I_k)\}_1^K$ in CLIP image space for each concept.

Training f_θ with text. When training with COCO data, f_θ learns concepts that are frequent in COCO captions. However, our goal is to have f_θ generalize to widely diverse concepts. Yet, naively training with the COCO images does not generalize well to out-of-COCO-vocabulary concepts (see Appendix C).

To generalize to out-of-vocabulary concepts, we propose synthesizing *textual* descriptions with an expanded vocabulary and embed them into the shared embedding space. Specifically, we use COCO *captions* of a concept to generate additional training examples for new concepts by replacing the concept type with the most similar concept type from a large predefined vocabulary of 20K types [33], where (cosine) similarity is measured in CLIP text space. Finally, we embed the augmented captions by taking their CLIP-text feature representation (Fig. 3, right). Overall, we found that training with augmented text representation significantly improved the performance of the model (see Table A.2).

As in [37], we observed that the encoding distribution of text and images does not overlap in CLIP space. As a result, training f_θ with CLIP embeddings of captions does not generalize well to image inputs. We address this problem by learning an alignment matrix $\mathbf{A}$ that maps CLIP representations of texts to their presumed image counterpart (Fig. 3, right). $\mathbf{A}$ is learned jointly with f_θ, and is only used when learning the personalization tokens. It is not used at inference time. Formally, a set of captions is first encoded by $h^{\mathcal{T}}$, then mapped to the image area of the CLIP space using $\mathbf{A}$ and then fed to f_θ.

4.2 *Personalization*: Learn an embedding of personalized concepts

To learn the word embedding of a personalized concept, we follow a similar process to training f_θ, but instead of tuning the parameters of f_θ, we optimize the actual embedding vector $\mathbf{w}_c$.

Specifically, let $\{I_k\}_{k=1}^{N_c}$ be a set of input images for the new concept c, we (1) map them to $\mathcal{I}$ using CLIP encoder $h^{\mathcal{I}}$, (2) map to $\mathbf{w}_c^0$ using f_θ, (3) plug the embedding $\mathbf{w}_c^0$ in a template sentence and (4) map to CLIP text space using $h^{\mathcal{T}}$. Once again, we define a contrastive cycle-consistent loss, matching the estimated text embedding of the template sentence $\hat{z}_c$, and average image embedding $\bar{z}_c$. However, here we contrast it with the embedding of the concept *type*, say "a short sleeve top", denoted by $\eta_c = h^{\mathcal{T}}(S_c)$. Since no negative image examples are provided in the personalization stage, the concept type can be viewed as a "super-concept" in the hierarchy. It allows the learning process to focus on

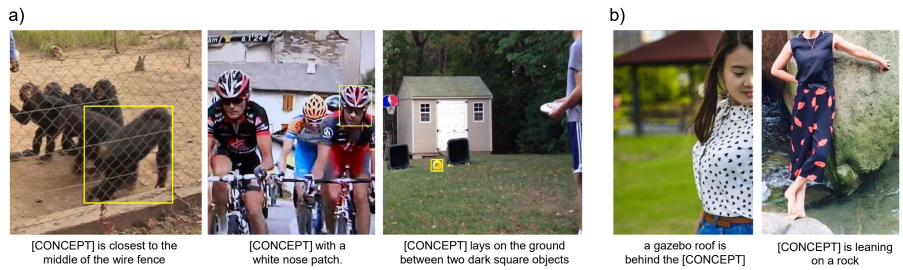

Fig. 5. Examples of textual queries and evaluation images of the new benchmarks, [CONCEPT] denotes the personalized concepts. **(a)** YTVOS frames. The queried concept is highlighted in a yellow box. In YTVOS segmentation, one should segment the correct concept in the frame. In YTVOS retrieval, each evaluation image is a box cropped around the concept. **(b)** Deepfashion2 examples.

the specific features that make the object unique from the general population of similar concept types. Similar to the SimCLR [9] loss, our loss is:

$$\ell\big(\hat{z}_c, \bar{z}_c, \eta_c\big) = -\log \frac{\exp\big(\cos(\bar{z}_c, \hat{z}_c)\big)}{\exp\big(\cos(\bar{z}_c, \hat{z}_c)\big) + 2 \cdot \exp\big(\cos(\eta_c, \hat{z}_c)\big)} \tag{2}$$

The factor 2 results from contrasting η_c with both visual and a text embedding.

4.3 Inference

Our approach expands the vocabulary of word-embedding tokens with personalized tokens, without modifying the underlying V&L model h. Therefore, for a given downstream task T and a sentence S, we use the pretrained V&L model h as it would have been used with T. But when we encounter an input sentence S that includes a [CONCEPT] token, we apply its learned embedding $\mathbf{w}_c$. Also, we found that having [CONCEPT] followed by the concept type S_c is beneficial.

5 Evaluation datasets for PerVL

We created two new personalization benchmark datasets for the evaluation of PerVL. (1) We collected captions for images from DeepFashion2 [24], which serve as search queries in an image retrieval task. (2) We collected captions for frames from Youtube-VOS [69], and also collected their corresponding segmentation maps for a referring-expression segmentation task.

5.1 Personalized fashion-item retrieval with DeepFashion2

We used the DeepFashion2 dataset [24] to create an image retrieval benchmark of personalized fashion items given a textual query. It contains photos of people wearing unique fashion items from 13 popular clothing categories, like *skirt* or *long-sleeve dress*, which we use as concept types. See the examples in Fig. 1, 5.

We created a dataset of 1700 images from 100 unique fashion items (concepts). Each item was assigned a unique [CONCEPT] tag. We assigned 450 images (out of the above 1700) to an *evaluation set*, and used raters to collect a textual description referring to each item appearing in the images. For instance, *The [CONCEPT] is facing a glass store display."*. In Appendix F.1 we describe the steps we took to select context-rich items.

Short versus detailed captions: We collected two types of captions for each image. First, *detailed* captions like "*White cabinets, some with open drawers, are alongside and behind the [CONCEPT].*". These describe extensive context about the image and can facilitate retrieval. Second are *short* captions like "*White cabinets are behind the [CONCEPT].*". These pose a greater challenge, because they describe less detail, and therefore are more ambiguous.

Finally, we randomly split the data to 50 val. concepts and 50 test concepts.

5.2 Youtube-VOS for personalized retrieval and segmentation.

We created an image segmentation benchmark of personalized visual concepts given a textual query using Youtube-VOS (YTVOS) [69]. YTVOS is a dataset for instance segmentation in video, which includes 4000+ videos, 90+ categories, and 7800+ unique object instances. To transform it to an image personalization benchmark, we take the last frame of each video (scene) for evaluation and the object instances that appear in the frame as the target concepts. Earlier frames are used as candidate frames for training. See the examples in Fig. 5 (left).

This benchmark is challenging as it contains ambiguities about both the textual queries and the appearance of the personalized concept. Hence, only a model that is successful in both personalization and image-text reasoning can succeed in this task. For that, we only select videos such that their object concept appears at least twice in an evaluation frame.

Finally, we annotated the instances in the evaluation frame with captions using AMT. We instructed the AMT workers to concisely describe what makes a specific entity distinct, compared to similar entities in the image. We provide more details and examples in Appendix F.3.

In total, this benchmark includes ~500 unique personalized concepts, with ~6300 training samples. For evaluation, we split according to unique scenes (videos), resulting in 246 validation concepts and 251 test concepts.

Personalized image retrieval: We also created an image retrieval variant of YTVOS. We extracted a set of images that correspond to the AMT captions collected for segmentation. Every image in the retrieval set was extracted from a wide box cropped around every instance in each evaluation frame. The goal is to retrieve the image of the correct instance given its textual query. Compared to the segmentation task, there are fewer distractors from the same scene for every instance, since not all instances were labeled in the data, but there are many more distractors coming from different scenes.

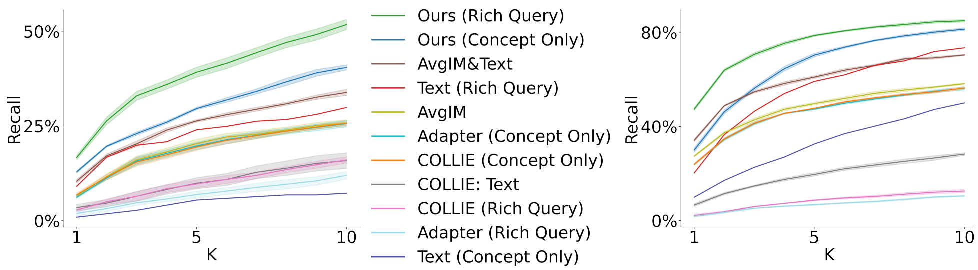

Fig. 6. Recall at K for our approach and baselines. DeepFashion2 (left), Youtube-VOS (right), PALAVRA achieves the highest rates for all retrieval metrics. On both benchmarks and metrics it achieves a significant improvement compared with “AvgIm&Text”, which is the strongest baseline. The experiments with “Concept-only query” demonstrate that the information in the textual query is essential for telling the target image from distractors, since the performance of both PALAVRA and “Text Only” substantially degrades with “Concept-only” queries.

6 Experiments

We tested PALAVRA with two PerVL benchmarks and compared with several leading baselines (Sec.6.1,6.2). We then study in greater depth the properties of PALAVRA, by an ablation study (Appendix C). All design decisions and hyperparameter tuning were performed on the validation set to avoid overfitting the test set. The experiments were carried out on NVIDIA V100 and A100 GPU. We provide additional implementation details and results in Appendix A,B.

6.1 Personalized image retrieval with a textual query

The objective of this task is to retrieve the correct image given a text query that includes the new concept (Fig. 1, top-right). We use AMT captions as textual queries describing a single image from the dataset. The challenge in this setup is to overcome two types of distractors. (a) visually similar distractors: images of the same personalized concept but in a different context than the context described by the textual query (e.g. the two instances of “my favorite skirt” in Fig. 1), (b) semantically similar distractors: images which include an item of a similar concept type (e.g. “a skirt”), in a similar context as described in the textual query.

To rank images according to a textual query, we rank images according to their cosine similarity with the embedded text query.

Compared methods. We compare our approach **PALAVRA**, with 5 baselines and their variants: **Text Only**: score an image-query pair using CLIP embedding of the text query $h^{\mathcal{T}}(S)$. Using the concept type for [CONCEPT] instead of its learned word embedding. **AvgIM**: Ignoring the text query and replace it by the average over the embedding of its concept training images. This is equivalent to the FSL baseline in [11]. **IM&Text**: Represent the query as the average between *AvgIM* and *Text*. **Random**: Test images are ranked in random order.

COLLIE [56]: Learn an adapter module over the output of CLIP text encoder, with an additional scaler function $Scaler\big(h^{\mathcal{T}}(S)\big) \in [0,1]$ that softly applies the adapter layer. COLLIE is closest to our method, because it may preserve some capabilities of the underlying pretrained model, when $Scaler(\cdot) = 0$. **Adapter**: As in COLLIE, but replace the scaler with a constant value of 1, making the "Adapter" layer always active. **COLLIE:Text**: COLLIE, when the text query uses the concept type for [CONCEPT], rather than the trained concept.

	DeepFashion2 MRR	YTVOS MRR
Random	2.9 ± 0.4	2.8 ± 0.2
Concept-only query		
Text Only	4.2 ± 0.0	21.5 ± 0.0
Adapter	13.4 ± 0.5	35.5 ± 0.3
COLLIE	13.8 ± 0.5	35.6 ± 0.3
AvgIm	13.8 ± 0.5	38.2 ± 0.3
PALAVRA (Ours)	19.4 ± 0.6	53.4 ± 0.8
Rich query		
Adapter	5.9 ± 0.7	5.3 ± 0.3
COLLIE	7.9 ± 0.7	6.2 ± 0.3
COLLIE: Text	8.0 ± 1.0	7.2 ± 0.3
Text Only	17.6 ± 0.0	37.6 ± 0.0
AvgIm+Text	18.8 ± 0.4	47.2 ± 0.3
PALAVRA w.o. tuning	22.1 ± 0.2	47.1 ± 0.8
PALAVRA (Ours)	**28.4 ± 0.7**	**61.2 ± 0.4**

Table 1. MRR retrieval metrics.

Evaluation metrics and queries. For image retrieval, we report two metrics (1) **Recall at K**: The rate of successful retrievals among the top-K scoring images. (2) **MRR** (Mean Reciprocal Rank): Average of 1 divided by the rank of the correct image. Errors denote the standard error of the mean (SEM) across 5 model initialization seeds. We use two types of queries: **(1) Rich query** uses the free-formed text annotated by AMT workers: "[CONCEPT] is leaning on a rock". **(2) Concept-only queries** overrides the "Rich query" by a template that focuses only on the concept: "A photo of a [CONCEPT]". Note that the baseline "AvgIM" is more related to the "Concept-only query" because the rich query embedding is overridden by the average embedding of the training examples.

Retrieval Results. Table 1 and Fig. 6 describe the retrieval rates of PALAVRA and the compared methods when using challenging *short* captions as our **Rich Queries**. We report the retrieval rates with *detailed* captions in the appendix. We note that both *short* queries and *detailed* queries are rich queries, containing known concepts in addition to the personalized ones. The *detailed* version possibly contains more of them. PALAVRA achieves the highest rate in all the retrieval metrics. On both benchmarks and metrics, it achieves significant improvement (between ~30% to ~50%) compared with "AvgIm+Text", which is the strongest baseline.

Comparing the results of "Concept-only query" with the "Rich query" results demonstrate that: (1) Information in the "Rich query" is essential for retrieval. (2) Adapter baselines (Adapter & COLLIE) improve over vanilla CLIP when only the concept is queried.

Their performance degrades when using the "Rich query". This happens as the adaptation layer trained for the personalized [CONCEPT] does not perform well with free-form text it has not seen during training. In fact, we find that

Adapter and COLLIE are even sensitive to the prompt *prefix* of the query. When changing their prefix to a prefix not used in training, their performance substantially degrades. We report this finding quantitatively in Appendix B.

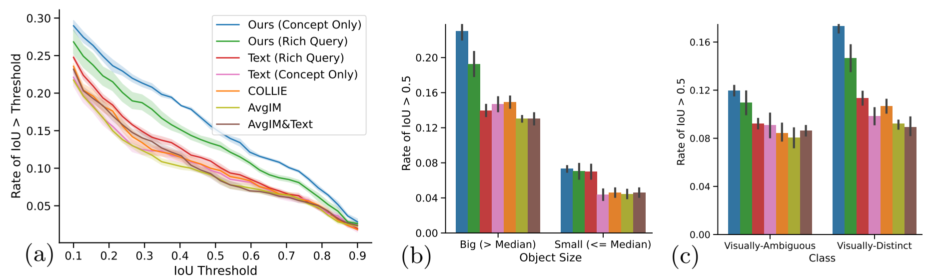

Fig. 7. Segmentation results on YTVOS. **(a)** Percent of images where IoU values exceeded a threshold, as a function of the threshold. Our approach dominates across the full range **(b, c)** Investigating model robustness under 2 levels of task complexity. We consider two scenarios that influence difficulty: Object size (b), and intra-class class ambiguity (c). When a clear visual signal is available (large objects, high intra-class visual variance), our model significantly outperforms the alternatives. Even in more challenging scenarios, our model still leverages textual descriptions, mitigating the loss of quality seen in models that ignore or corrupt CLIP's embedding space.

6.2 Semantic segmentation with a textual query

The second downstream task, aims to segment an instance of a personalized concept in an image, based on a textual query that refers to the concept (Fig. 1, right bottom). The challenge here is to overcome two types of distractors: First, *visual distractors* that look similar to the concept and can be disambiguated with the text query. Second, *semantic distractors* that include a concept of a similar type (e.g. another type of a "toy wagon"), but CLIP has difficulty to resolve using just the concept type, like "an elephant on a toy wagon".

Here we investigate the performance of PALAVRA and baseline models on YTVOS dataset using a recent CLIP-based semantic segmentation [71]. In brief, [71] creates a set of query-driven relevance maps for the image, coupled with transformer interpretability methods [8]. The maps then serve as pseudolabels for single-image semantic segmentation [58].

Compared methods. We compare PALAVRA with a set of baselines in two setups: "Rich query" and "Concept-only query", as described in Sec. 6.1. **(1) Text (CLIP)**, using both the rich and concept only queries, **(2) AvgIM**, **(3) IM&Text** and **(4) COLLIE**. All baselines are described in Sec. 6.1.

Evaluation metrics. We calculate the intersection-over-union (IOU) between the predicted segment and the ground-truth segment. We report the **Rate of IOU > threshold**, which is the fraction of segments with IOU > threshold. Error bars denote the standard error of the mean (SEM) across 5 model seeds.

Semantic segmentation results

Fig. 7a shows the percent of test-set images for which IoU exceeds a given threshold. Our model consistently outperforms the baselines with wide margins (e.g. a 44.69% improvement over the best competitor at an IoU threshold of 0.5). These results demonstrate that a personalized prompt can extend even to localized image tasks. Moreover, as our method only expands CLIP's input space - it can be readily integrated with existing models for downstream tasks.

Surprisingly, our method performs better when using the concept-only query. We hypothesize that this is a result of CLIP's difficulty in reasoning over complex referring expressions: By mentioning the context within the text, CLIP's attention is diverted to the context itself which leads to false negatives. In contrast, in retrieval, context objects rarely appear in "negative" (non-target) images. Since they appear in the target image, they actually help to detect the correct image. Appendix E.2 provides quantitative evidence supporting this hypothesis.

We further examine cases where we expect existing methods to falter. In Fig. 7b, we examine a scenario where objects are small, so their crops may not provide sufficient signal to the CLIP image encoder. Indeed, when for objects below the median size, segmentation fares worse in general, and image based-methods suffer even more. Our method, however, can rely on the signal derived from the text and degrades less. Fig. 7c examines a scenario in which objects in a scene are less visually distinct. We divide the evaluation set to object which are usually visually distinct and images which often contain visually ambiguous objects, where a few instances of the same object appear in the same image. In practice, we split to animal and non-animal categories, as animals are mostly visually ambiguous (e.g. Fig. 5 left). Our model substantially outperforms the baselines when the concepts are visually distinct and also improves when the concepts are mostly ambiguous.

In the Appendix B we provide and discuss qualitative segmentation results.

7 Discussion

We described an approach to leverage large pre-trained V&L models like CLIP, for learning a representation of new "personal" classes from a handful of samples. Our key idea is to expand the input space of V&L models by finding a representation of the new concept. The extended model can then be used for V&L tasks with a rich language that "understands" both novel and known concepts. A limitation of the approach is that it suffers from the limitations of the underlying V&L model. For instance, CLIP struggles with understanding spatial relations within a photo [39], and extended representations based on CLIP suffer from the same problem. We expect that our approach can be extended to other V&L models. See an example in Appendix D.

To conclude, we hope that the method presented in this paper will pave the way to using pretrained models in problems that involve user-specific concepts, like home robotics and organizing personal data.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**, 1425–1438 (2016)
2. Anwaar, M.U., Labintcev, E., Kleinsteuber, M.: Compositional learning of image-text query for image retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1140–1149 (2021)
3. Atzmon, Y., Berant, J., Kezami, V., Globerson, A., Chechik, G.: Learning to generalize to new compositions in image understanding. arXiv preprint arXiv:1608.07639 (2016)
4. Atzmon, Y., Chechik, G.: Probabilistic and-or attribute grouping for zero-shot learning. In: Proceedings of the Thirty-Forth Conference on Uncertainty in Artificial Intelligence (2018)
5. Atzmon, Y., Chechik, G.: Adaptive confidence smoothing for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11671–11680 (2019)
6. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
7. Carey, S., Bartlett, E.: Acquiring a single new word. (1978)
8. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 782–791 (June 2021)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
10. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3001–3011 (2020)
11. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9062–9071 (2021)
12. Cheraghian, A., Rahman, S., Fang, P., Roy, S.K., Petersson, L., Harandi, M.: Semantic-aware knowledge distillation for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2534–2543 (2021)
13. Chunseong Park, C., Kim, B., Kim, G.: Attend to you: Personalized image captioning with context sequence memory networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 895–903 (2017)
14. Dekel, O., Keshet, J., Singer, Y.: Large margin hierarchical classification. In: Proceedings of the twenty-first international conference on Machine learning. p. 27 (2004)
15. Del Chiaro, R., Twardowski, B., Bagdanov, A., Van de Weijer, J.: Ratt: Recurrent attention to transient tasks for continual image captioning. Advances in Neural Information Processing Systems **33**, 16736–16748 (2020)
16. Demirel, B., Cinbis, R.G., Ikizler-Cinbis, N.: Image captioning with unseen objects. arXiv preprint arXiv:1908.00047 (2019)
17. Denton, E., Weston, J., Paluri, M., Bourdev, L., Fergus, R.: User conditional hashtag prediction for images. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1731–1740 (2015)

18. Fan, L., Xiong, P., Wei, W., Wu, Y.: Flar: A unified prototype framework for few-sample lifelong active recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15394–15403 (2021)
19. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
20. Feng, F., Liu, R., Wang, X., Li, X., Bi, S.: Personalized image annotation using deep architecture. IEEE Access **5**, 23078–23085 (2017)
21. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
22. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946 (2021)
23. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)
24. Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P.: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
25. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: Describing novel object categories without paired training data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2016)
26. Hewitt, J., Li, X.L., Xie, S.M., Newman, B., Liang, P.: Ensembles and cocktails: Robust finetuning for natural language generation. In: NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications (2021)
27. Hill, F., Tieleman, O., von Glehn, T., Wong, N., Merzic, H., Clark, S.: Grounded language learning fast and slow. arXiv preprint arXiv:2009.01719 (2020)
28. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
29. Hsieh, Y.G., Niu, G., Sugiyama, M.: Classification from positive, unlabeled and biased negative data. In: International Conference on Machine Learning. pp. 2820–2829. PMLR (2019)
30. Jia, X., Zhao, H., Lin, Z., Kale, A., Kumar, V.: Personalized image retrieval with sparse graph representation learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2735–2743 (2020)
31. Khan, M., Srivatsa, P., Rane, A., Chenniappa, S., Hazariwala, A., Maes, P.: Personalizing pre-trained models. arXiv preprint arXiv:2106.01499 (2021)
32. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054 (2022)
33. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
34. Lake, B.M., Piantadosi, S.T.: People infer recursive visual concepts from just a few examples. Computational Brain & Behavior **3**(1), 54–65 (2020)

35. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
36. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
37. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. arXiv preprint arXiv:2203.02053 (2022)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
39. Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. Advances in Neural Information Processing Systems **34** (2021)
40. Long, C., Yang, X., Xu, C.: Cross-domain personalized image captioning. Multimedia Tools and Applications **79**(45), 33333–33348 (2020)
41. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7219–7228 (2018)
42. Lynch, C., Sermanet, P.: Language conditioned imitation learning over unstructured data. arXiv preprint arXiv:2005.07648 (2020)
43. Ma, T., Geng, S., Wang, M., Shao, J., Lu, J., Li, H., Gao, P., Qiao, Y.: A simple long-tailed recognition baseline via vision-language model. arXiv preprint arXiv:2111.14745 (2021)
44. Malaviya, M., Sucholutsky, I., Oktar, K., Griffiths, T.L.: Can humans do less-than-one-shot learning? arXiv preprint arXiv:2202.04670 (2022)
45. Markman, E.M.: Constraints children place on word meanings. Cognitive science **14**(1), 57–77 (1990)
46. Markman, E.M., Wasow, J.L., Hansen, M.B.: Use of the mutual exclusivity assumption by young word learners. Cognitive psychology **47**(3), 241–275 (2003)
47. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
48. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
49. Paz-Argaman, T., Atzmon, Y., Chechik, G., Tsarfaty, R.: Zest: Zero-shot learning from text descriptions using textual similarity and visual summarization (2020)
50. 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)
51. Ren, M., Iuzzolino, M.L., Mozer, M.C., Zemel, R.S.: Wandering within a world: Online contextualized few-shot learning. arXiv preprint arXiv:2007.04546 (2020)
52. Ren, M., Ravi, S., Triantafillou, E., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=HJcSzz-CZ
53. Shinoda, K., Kaji, H., Sugiyama, M.: Binary classification from positive data with skewed confidence. arXiv preprint arXiv:2001.10642 (2020)
54. Shridhar, M., Manuelli, L., Fox, D.: Cliport: What and where pathways for robotic manipulation. In: Proceedings of the 5th Conference on Robot Learning (CoRL) (2021)

55. Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12516–12526 (2019)
56. Skantze, G., Willemsen, B.: Collie: Continual learning of language grounding from language-image embeddings. arXiv preprint arXiv:2111.07993 (2021)
57. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
58. Sofiiuk, K., Petrov, I., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. arXiv preprint arXiv:2102.06583 (2021)
59. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
60. Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12183–12192 (2020)
61. Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34** (2021)
62. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5753–5761 (2017)
63. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016)
64. Wang, L., Meng, X., Xiang, Y., Fox, D.: Hierarchical policies for cluttered-scene grasping with latent plans. IEEE Robotics and Automation Letters (2022)
65. Wortsman, M., Ilharco, G., Li, M., Kim, J.W., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. arXiv preprint arXiv:2109.01903 (2021)
66. Wu, G., Gong, S., Li, P.: Striking a balance between stability and plasticity for class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1124–1133 (2021)
67. Wu, Y., Zhu, L., Jiang, L., Yang, Y.: Decoupled novel object captioner. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1029–1037 (2018)
68. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
69. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 585–601 (2018)
70. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
71. Zabari, N., Hoshen, Y.: Semantic segmentation in-the-wild without seeing any segmentation examples (2021)
72. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. Advances in neural information processing systems **30** (2017)
73. Zhang, R., Fang, R., Gao, P., Zhang, W., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)

74. Zhang, Y., Zhang, C.B., Jiang, P.T., Cheng, M.M., Mao, F.: Personalized image semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10549–10559 (2021)
75. Zheng, Y., Li, Y., Wang, S.: Intention oriented image captions with guiding objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8395–8404 (2019)
76. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021)