

Training-Free Personalization via Retrieval and Reasoning on Fingerprints

Deepayan Das¹ Davide Talon² Yiming Wang²
 Massimiliano Mancini¹ Elisa Ricci^{1,2}

¹University of Trento ²Fondazione Bruno Kessler

{deepayan.das, massimiliano.mancini, e.ricci}@unitn.it

{dtalon, ywang}@fbk.eu

Abstract

Vision Language Models (VLMs) have led to major improvements in multimodal reasoning, yet they still struggle to understand user-specific concepts. Existing personalization methods address this limitation but heavily rely on training procedures, that can be either costly or unpleasant to individual users. We depart from existing work, and for the first time explore the training-free setting in the context of personalization. We propose a novel method, Retrieval and Reasoning for Personalization (R2P), leveraging internal knowledge of VLMs. First, we leverage VLMs to extract the concept fingerprint, i.e., key attributes uniquely defining the concept within its semantic class. When a query arrives, the most similar fingerprints are retrieved and scored via chain of thought reasoning. To reduce the risk of hallucinations, the scores are validated through cross-modal verification at the attribute level: in case of a discrepancy between the scores, R2P refines the concept association via pairwise multimodal matching, where the retrieved fingerprints and their images are directly compared with the query. We validate R2P on two publicly available benchmarks and a newly introduced dataset, Personal Concepts with Visual Ambiguity (PerVA), for concept identification highlighting challenges in visual ambiguity. R2P consistently outperforms state-of-the-art approaches on various downstream tasks across all benchmarks. Code and data are available at the project page: [Training-Free Personalization](#).

1. Introduction

While "Where are my keys?" or "What is Fuffy doing?" are questions easy to understand for humans, they are challenging for Vision and Language Models (VLMs) [1, 2, 5, 10, 23] due to their inherent lack of understanding for user-specific concepts. Toward addressing this limitation, VLMs personalization aims to add personal concepts to the knowledge of a VLM [4, 13, 29]. Recent works on VLM personalization mainly rely on training procedures, where multiple reference

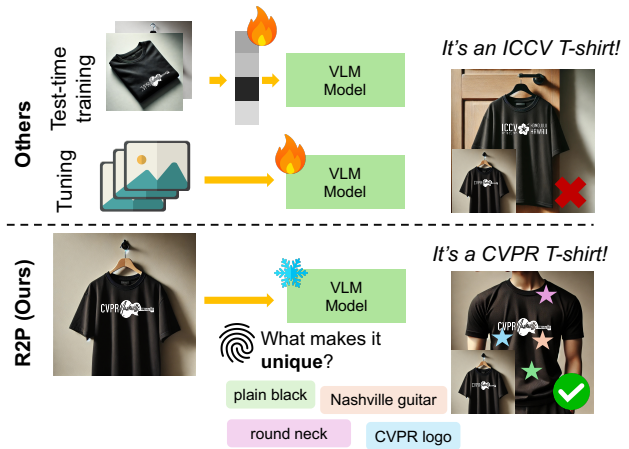


Figure 1. Current VLMs personalization methods depend on expensive training procedures. In contrast, we introduce R2P, a training-free approach that utilizes textual attributes as unique fingerprints for identifying personal concepts.

samples are used to learn a custom representation for each personal concept in a contrastive manner, against a large number of negative samples from the same class [4, 29]. This data collection is not only costly, but requires expensive fine-tuning each time a new concept is added. Alternatively, other works perform expensive large-scale fine-tuning on synthetic personalization data [13, 32], using a retrieval-augmented strategy to strengthen the user-specific context at inference time [13]. However, they do not eliminate the cost of large-scale pretraining, which can be prohibitive when considering modern VLMs.

One common assumption among these works is that training is needed to enable VLMs to capture personal concepts. However, VLMs have been exposed to virtually any semantic concept through their web-scale training data. The availability of this vast knowledge, raises a natural question: *Can we exploit VLMs's internal knowledge to perform personalization without training?*

In this work, we show for the first time that training-free

personalization is feasible, introducing *Retrieval and Reasoning for Personalization (R2P)*, a novel retrieval-reasoning framework to infer personal concepts of interest, using only pre-trained VLMs. Specifically, we first construct a database containing rich multimodal information about all user-specific concepts, with reference images and textual descriptions. In particular, for each concept, we leverage the internal world knowledge of a VLM to enrich the textual description with its distinctive attributes, considering its semantic category and the reference image. Such set of distinctive attributes help in uniquely identifying the concept, serving as a concept *fingerprint*, as shown in Fig. 1. Note that the personal database is created once and can be extended flexibly whenever new personal concepts arrive.

At inference time, given the query image, R2P first extracts a set of candidate concepts from the personal database, via multimodal retrieval. Then, we perform Chain-of-Thought (CoT) reasoning with multimodal prompt, where the VLM is instructed to focus on *fingerprint attributes* and infer the most-matched personal concept among all retrieved candidates, given their textual information. As VLMs are notorious for visual hallucination, especially on fine-grained attributes [42, 51], we introduce a cross-modal attribute verification module. Specifically, from the CoT reasoning output, we extract the fingerprint attributes that the VLM leveraged for the concept inference. With pre-trained vision-text aligned encoders [34], we check the alignment between attributes and the query image, from which we can derive the best-matched personal concept. We consider a positive verification when the inferred concept via attribute-image alignment coincides with the CoT output. In case of a mismatch, the prediction is refined via multimodal pairwise reasoning, where the query is compared with each candidate concept (using reference images and textual descriptions).

We evaluate R2P on two existing benchmarks for VLM personalization, MyVLM [4] and Yo’LLaVA [29], we further introduce a new personalization dataset, Personal Concepts with Visual Ambiguity (PerVA), specifically featuring the challenge of visually and semantically similar concepts. On all benchmarks, R2P, while being training-free, consistently outperforms training-based competitors on both recognition, VQA, and captioning tasks, with a significant margin (e.g., +2% weighted recognition accuracy on Yo’LLaVA Dataset, +8.4% captioning recall on PerVA).

Our contributions are summarized as follows:

- We investigate a novel *training-free setting* for personalization, by leveraging internal knowledge in pre-trained VLMs to eliminate training efforts.
- We propose a novel method, *R2P*, which uses a retrieval-reasoning paradigm and textual attributes to uniquely identify personal concepts.
- We introduce a new benchmark, *PerVA*, specifically designed to challenge personalization methods by incorpo-

- rating concepts with high visual and semantic similarity.
- Experiments demonstrate that R2P achieves *state-of-the-art* across all benchmarks, successfully recognizing concepts of interest amid visual ambiguity.

2. Related Work

Personalization with VLMs. Personalization has been addressed on different vision tasks. For instance, in text-to-image generation, target objects are associated with unique identifiers learned from a few provided images, e.g., via textual-inversion [12] or with fine-tuning [18, 36, 39]. In segmentation, recent approaches leverage correspondences between the reference and test image to construct a confidence map about where the target object is located [24, 37, 41, 50] for later prompting with SAM [17]. To this end, strategies rely on diffusion models [37, 41], personalized generative models [40] or cycle consistency [6].

In this work we specifically focus on the problem of VLM personalization where models can recognize and refer to personal concepts of interest. Earlier approaches, such as MyVLM [4] and Yo’LLaVA [29], use an inversion strategy inspired by generative models, where each object is assigned a unique latent representation, either as a concept vector [4] or token [29]. Recent works focused on reducing the time required to personalize VLMs to new concepts by using large-scale tuning and multiple images as inputs to provide in-context information [13, 31, 32]. However, state-of-the-art approaches either require test-time training or large-scale pretraining, potentially limiting their generalization capabilities. Furthermore, they implicitly assume that personal objects do not change over time.

In contrast with previous works, we propose leveraging the world knowledge of the VLM to identify unique features of the object of interest, and building on the reasoning capabilities of recent models to achieve personalization in a training-free manner.

Attribute-based inference. Recognizing objects from the attributes composing them has been a long-standing problem in computer vision [11, 19]. Notable efforts have been done in the context of zero-shot learning (ZSL), where an attribute classifier trained on a set of classes allows to recognize unseen ones at test time [3, 19, 43]. Similarly, researchers studied how attributes change the appearance of an object, both in images [25, 27, 28] and in videos [7–9], performing compositional recognition.

Our work is close in spirit to those aiming to discover useful attributes [15, 44] or machine generated ones [26, 33, 35, 45] for aiding class discrimination. Moreover, similar to works performing visual classification with contrastive VLMs, we use machine generated descriptions to better recognize a concept. However, differently from

these works, we focus on how to describe a personal object to the VLM, using cross-modal verification to reduce issues on hallucinations [20, 21, 52] and compositional reasoning [14, 48].

3. Retrieval and Reasoning for Personalization

We first introduce the problem of VLMs personalization. Then we describe our training-free method R2P in details, in terms of personal database creation (Sec. 3.1), and personal concept inference with retrieval and reasoning (Sec. 3.2).

Problem formulation. The goal of personalization is to enable a VLM to reason about user-specific concepts, *e.g.*, answering queries about them. Let us denote with \mathcal{V} and \mathcal{T} the visual and textual space, respectively. The VLM of interest, Φ_{VLM} , maps images and text inputs to textual outputs, *i.e.*, $\Phi_{\text{VLM}} : \mathcal{V} \times \mathcal{T} \rightarrow \mathcal{T}$.

Following [13], we assume that the user provides a reference image $I_i \in \mathcal{V}$, a name $c_i \in \mathcal{T}$ (*e.g.*, “Sleeping Shiba”) for each of the N personal concepts of interest, together with its category $g_i \in \mathcal{T}$ (*e.g.*, “Stuffed animal”), from which we can construct a user-specific multimodal database \mathcal{D} . Given a test-time query image $Q \in \mathcal{V}$ together with a textual prompt $P_q \in \mathcal{T}$ (*e.g.*, “Is Sleeping Shiba in the image?”), the VLM Φ_{VLM} should provide the answer relevant to the personal concept (*e.g.*, “yes! Sleeping Shiba is resting on the couch”). Such personalization process involves two distinct tasks: (i) recognizing which specific personal concept is present in the image and (ii) answering the query.

Early approaches focus on learning either concept-specific embeddings [4, 29], or a Φ_{VLM} that models personal concepts via instruction tuning [13]. Differently, we address the personalization problem in a training-free manner with a novel retrieval-reasoning paradigm. Our proposed method R2P consists of two phases, as shown in Fig. 2. *Phase I* aims to create the personal database \mathcal{D} via VLM-based Visual Question Answering (VQA), enriching images with textual information in form of distinctive *fingerprint* attributes. *Phase II* refers to the inference phase for recognizing the personal concept from the personal database. Note that the database is created once and can be flexibly updated with concepts inclusion/deletion/modification.

3.1. Personal Database Creation

In the first phase, we aim to construct a multimodal database, \mathcal{D} , to store the user-provided personal concepts along with the enriched textual information regarding their distinctive *fingerprint* attributes. The *fingerprint* attributes help to uniquely distinguish the personalized concept from other visually similar ones. The inclusion of such *fingerprint* attributes is a major distinction of our work from prior work [13], that also involves retrieval at inference time.

Specifically, we leverage the same VLM Φ_{VLM} for personalization to extract the *fingerprint* attributes via multimodal

prompting. The reference image I_i serves as the visual prompt P_D^V and the category g_i is incorporated into the textual prompt P_D^T . By prompting the VLM with (P_D^V, P_D^T) , we produce a list of distinctive *fingerprint* attributes A_i and a brief caption d_i aiming to uniquely describe the concept:

$$\{A_i, d_i\} = \Phi_{\text{VLM}}(P_D^V, P_D^T). \quad (1)$$

Note that the textual prompt P_D^T explicitly instructs the VLM to extract fine-grained attributes A_i (*e.g.*, fur color, shirt logo) to help disambiguate similar objects, ensuring that A_i contains distinctive attributes, and the description d_i is discriminative. Please refer to *Supp. Mat.* for the complete multimodal prompt P_D^V, P_D^T .

To facilitate retrieval at the inference phase (*Phase II*), we further propose to encode the reference images and descriptions of all personal concepts. Specifically, for each personal concept, we encode its reference images using an image encoder $\mathcal{E}_V : \mathcal{V} \rightarrow \mathcal{Z}$, obtaining the visual embedding $f_i^V = \mathcal{E}_V(I_i)$, where \mathcal{Z} indicates the latent embedding space. Similarly, we encode the description d_i using the text encoder $\mathcal{E}_T : \mathcal{T} \rightarrow \mathcal{Z}$, obtaining the textual embedding $f_i^T = \mathcal{E}_T(d_i)$.

At the end of this phase, we have the enriched multimodal personal database $\mathcal{D} = \{I_i, c_i, g_i, d_i, A_i, f_i^V, f_i^T\}_{i=1}^N$, where we included as new elements the extracted visual (f_i^V) and textual (f_i^T) features, the enriched descriptions (d_i), and the set of distinctive fingerprint attributes (A_i). In the following, we describe how we use the database \mathcal{D} during inference.

3.2. Concept inference with Retrieval-Reasoning

Multimodal concept retrieval. At inference phase, given a query image $Q \in \mathcal{V}$, we first retrieve its K most relevant concepts that are visually and textually similar from the database \mathcal{D} .

Specifically, we compute the image embedding of the query image f_q^V using the image encoder \mathcal{E}_V , *i.e.*, $f_q^V = \mathcal{E}_V(Q)$. Then we calculate the cosine similarity between f_q^V and each visual embedding $f_i^V \in \mathcal{D}$ and textual embedding of stored personal concepts $f_i^T \in \mathcal{D}$:

$$\begin{aligned} s_{q,i}^{V,V} &= \langle f_q^V, f_i^V \rangle, \forall f_i^V \in \mathcal{D} \\ s_{q,i}^{V,T} &= \langle f_q^V, f_i^T \rangle, \forall f_i^T \in \mathcal{D}, \end{aligned} \quad (2)$$

where $\langle x, y \rangle = \frac{x^t y}{\|x\| \cdot \|y\|}$ is the cosine similarity between two vectors x and y .

The final similarity score $s_{q,i}^{s,q,i}$ accounts for both textual and visual similarities as:

$$s_{q,i} = \frac{1}{2}(s_{q,i}^{V,V} + s_{q,i}^{V,T}), \forall i \in \mathcal{D}. \quad (3)$$

Finally, we select the top- K candidate personal concepts \mathcal{C}^K with the highest score $s_{q,i}$. By incorporating both visual

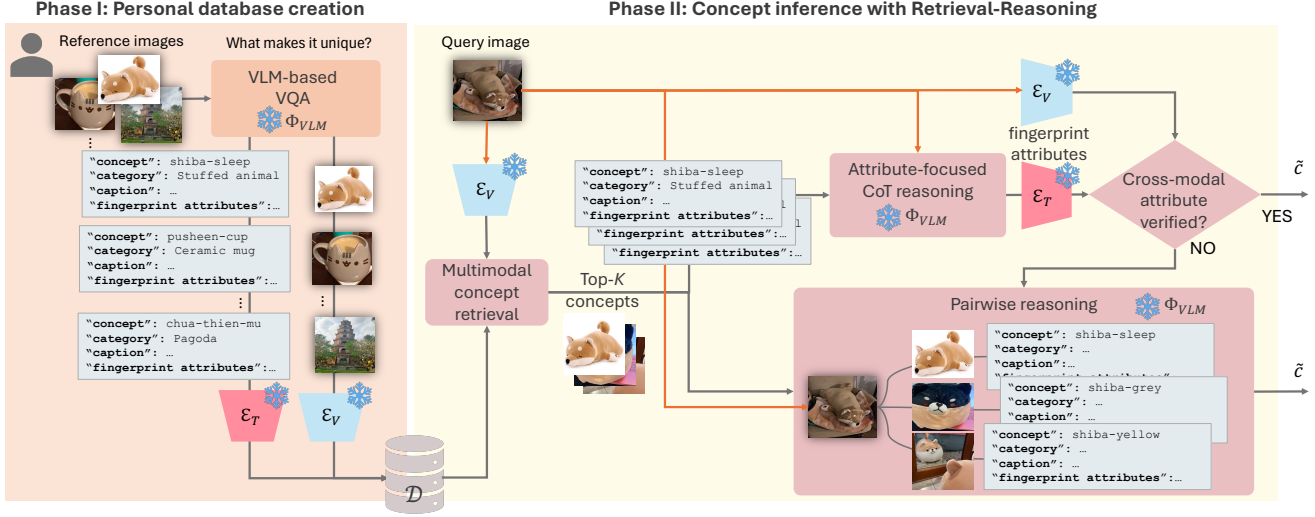


Figure 2. R2P is the first training-free method to address VLM personalization, aiming to recognize the personal concept from a query image. R2P consists of two phases. First, in the the personal database \mathcal{D} creation phase, we leverage the VLM Φ_{VLM} to enrich personal concepts with their distinctive *fingerprint* attributes. Then, in the inference phase, relevant concepts are retrieved from the personal database, and the best matched personal concept \tilde{c} is obtained with focused reasoning based on images and fingerprint attributes.

and textual cues, this multimodal retrieval strategy is effective, particularly in cases where visual similarity alone may be misleading (see results of our ablation study in Sec. 5).

Attribute-focused CoT reasoning. Once we obtain the set of candidate personal concepts \mathcal{C}^K , we further leverage the reasoning capabilities of the same VLM Φ_{VLM} , to select the concept that best matches the query image, focusing on the *fingerprint* attributes. In this initial reasoning step, we prompt the VLM in a multimodal fashion to first focus on the *fingerprint* attributes that are shared among the query image and then on each candidate concept within \mathcal{C}^K . The output is the best matched candidate concept \tilde{c} based on the common attributes. Specifically, the query image Q serves as the visual prompt P_R^V to the VLM. The textual prompt P_R^T includes the textual information of *all the candidate concepts*, followed by the instruction on attribute-based CoT reasoning. Please refer to *Supp. Mat.* for the complete multimodal prompt P_R^V, P_R^T .

By feeding both P_R^V and P_R^T to the VLM, we obtain the output concept \tilde{c} , as well as a list of *fingerprint* attributes $A_{q,i}$ shared among the query image and each personal concept c_i within \mathcal{C}^K , as expressed below:

$$\{A_{q,i}, \forall i \in \mathcal{C}^k\}, \tilde{c} = \Phi_{VLM}(P_R^V, P_R^T). \quad (4)$$

Cross-modal attribute verification. In the previous step, the VLM mostly relies on the textual information of the candidate concepts for reasoning. While this is computationally efficient, VLM may struggle with fine-grained disambiguation due to model hallucinations and the lack of

visual cues [42, 51]. For instance, some attributes that are leveraged for reasoning might not be present in the query image, making the model prediction less reliable.

Thus, we further validate the predicted concept \tilde{c} based on the vision-language alignment at attribute level. Specifically, for each candidate concept $c_i \in \mathcal{C}^k$, we encode each *fingerprint* attribute $a_j \in A_{q,i}$ using the text encoder \mathcal{E}_T , and compute an attribute-based cross-modal embedding similarity score with the image embedding of the query image f_q^V accounting all $A_{q,i}$, i.e.,

$$s_{q,i}^{V,A} = \frac{1}{|A_{q,i}|} \sum_{a_j \in A_{q,i}} \langle f_q^V, f_{a,j}^T \rangle, \quad (5)$$

where $f_{a,j}^T = \mathcal{E}_T(a_j)$.

We can then obtain the concept with the highest attribute-based cross-modal similarity as:

$$\tilde{c}_a = \arg \max_{c \in \mathcal{C}^k} s_{q,i}^{V,T}. \quad (6)$$

If $\tilde{c} = \tilde{c}_a$, the global and attribute-based predictions match: thus, the presence of the concept in the image is verified. On the other hand, if $\tilde{c} \neq \tilde{c}_a$, the verification fails: this triggers the more exhaustive but accurate pairwise reasoning, as detailed below.

Pairwise reasoning. Following the verification step, we disambiguate uncertain cases by leveraging Φ_{VLM} , checking if each concept in \mathcal{C}^K is present in the query image by relying on multimodal information of retrieved personal concepts.



Figure 3. Qualitative visualization of concepts for the proposed *PerVA* dataset. In order, samples from bottles, towels and clothes. **Top**: reference images for personalization with their concept indicated above. **Bottom**: query images at inference time.

We frame the recognition problem as a binary classification task where for each candidate concept $c_i \in \mathcal{C}^K$, we predict how likely c_i is present in the query image Q .

We provide multimodal prompts to the VLM for each candidate concept. Specifically, the visual prompt P_P^V includes both the query image Q and the reference image I_i . The textual prompt P_P^T contains all textual information stored in the personal database \mathcal{D} regarding the candidate concept, *i.e.*, $\{c_i, g_i, d_i, A_i\}$. The textual prompt P_P^T also instructs the model to answer whether the query image matches the currently compared concept in the format of Yes/No and evaluate the probability of the object to be present as:

$$p_i = \frac{\lambda_i^{\text{Yes}}}{\lambda_i^{\text{Yes}} + \lambda_i^{\text{No}}} \quad (7)$$

where λ_i^{Yes} and λ_i^{No} denote the logit that the output layer of the VLM assigns respectively to the Yes and No tokens when answering for the concept $c_i \in \mathcal{C}^K$.

Hence, the final predicted concept \tilde{c} is refined as the candidate concept maximizing the probability p_i :

$$\tilde{c} = \arg \max_{c \in \mathcal{C}^K} p_i. \quad (8)$$

With the recognized concept \tilde{c} , the VLM can then generate personalized captions or answer a personalized query, depending on the user prompt. Note that when the user explicitly mentions the concept name in his textual prompt, the corresponding image and text description can be retrieved from the \mathcal{D} without going through retrieval-reasoning steps in the proposed R2P.

4. PerVA dataset

Current personalization datasets [4, 29] focus on a limited number of concepts (≤ 50) from a small set of categories. Often, the personal concepts feature objects with highly distinctive traits, *e.g.* a pair of sporty shoes in pink and orange

color. The images containing the personal concepts are also well posed in images with good lighting. As such they exhibit limited recognition challenge as there is little visual ambiguity among personal concepts. Other datasets focus on synthetically generated personal concepts [13], which may not accurately reflect the challenges encountered in real-world scenarios.

We thus introduce a more challenging benchmark for VLM personalization, *Personal Concepts with Visual Ambiguity (PerVA)*, comprising many everyday object categories and each category having multiple instances with similar visual features. We construct PerVA by repurposing a publicly available dataset, originally proposed in [38] to study robust object recognition and retrieval. This dataset contains images featuring everyday objects belonging to the same category, which are observed in different poses, viewpoints and possibly different states, *e.g.*, t-shirt being folded or hang. Specifically, we restructure the existing dataset by creating new splits. Ideally, the query images at inference time should be different from the reference images used for model personalization. To this end, we consider the images associated to a concept c_i , and compute the average image embedding representation using a pre-trained image encoder [34]. This average image embedding serves as an anchor embedding to determine which samples can be used as reference images for model personalization, and which are reserved for inference. Specifically, as the reference image, we select the image whose visual embedding is closest to the average embedding corresponding to a concept. as query images, we select images whose embeddings are the furthest from the average one.

With such splitting strategy, we create a challenging evaluation dataset for personal concept recognition. At inference time, specific concepts may appear in different poses, lighting conditions, and contexts. In total, PerVA contains 329 personal concepts, spanning 21 categories of everyday entities such as clothes, bottles, trolleys and umbrellas. Among existing datasets [4, 29], PerVA presents a larger number of concepts per category ranging from 2 (headphones) to 69 (retail) concepts and stress-tests personalization models in well-known challenging settings, namely deformation for non-rigid objects, different illumination conditions and different object states, *e.g.*, folded clothes. We provide additional details on the PerVA dataset in the supplementary. Fig. 3 showcases reference images for personalization and query images. While this work focuses on a single reference image for personalization, our dataset construction strategy can be extended to multiple reference images and other datasets.

5. Experiments

We evaluate R2P on established personalization benchmarks, *i.e.* MyVLM [4] and Yo’LLaVA [29], as well as our newly

introduced PerVA dataset. We first outline our evaluation protocol and the implementation details, then discuss our results comparing R2P with state-of-the-art personalization methods. Finally, we present extensive ablation studies that highlights the importance of the design choices of our method.

5.1. Experimental setup

Datasets. In addition to PerVA, we also conduct evaluation on existing personalization benchmarks, *i.e.* MyVLM [4] and Yo’LLaVA. MyVLM [4] considers 29 personal concepts, while Yo’LLaVa [29] includes images of 40 concepts, including objects, people, and buildings. Unlike the original evaluation protocol [4, 29], which uses multiple positive and negative samples for training, our approach relies on a single reference image during personalization and does not require any negative samples. This setting, while being more challenging, better aligns with real-world use cases, as user input for personalization is minimized.

Downstream tasks and Metrics. Following prior works [13, 29], we evaluate R2P on three tasks: object recognition, captioning and personalized VQA. In the recognition task, the model determines whether a specific concept of interest is present in a query image. Positive samples correspond to images containing the personal concept of interest, while negative samples are images featuring other concepts. We frame object recognition as a binary classification problem and report recall ($Pos. Acc.$), specificity ($Neg. Acc.$), and their uniformly weighted average (Wtd). On the captioning task, instead, we assess Hard Recall ($Recall$), where the model should detect any personal concept of interest without prior knowledge of the specific concept. This is measured by the fraction of times the concept name appears in the model’s generated captions for its test images. Finally, on personalized VQA, we evaluate the model’s ability in answering closed-set questions about the personalized concepts, measuring the overall answering accuracy. Results are reported as the average over three different seeds.

Implementation Details. We use Mini-CPM-o-2.6 [46] as the underlying VLM for all experiments, chosen for its lightweight design and strong comparative image understanding capabilities. Built on LLaVA-UHD, it employs SigLIP [49] as the vision encoder, LLaMA 3.1 as the text decoder, and a multimodal projection layer for alignment.

To extract fingerprint attributes and descriptions, we prompt Mini-CPM-o-2.6 with task-specific templates (see supplementary for full details). For retrieval, we use FAISS [16] following [13], and construct the personal database using CLIP ViT-L/14-336 to encode both image and caption embeddings. We set $K=3$ in all experiments to balance accuracy and efficiency.

On MyVLM and Yo’LLaVA datasets, we crop images before encoding to isolate individual concepts, given the presence of multiple objects per image.

5.2. Comparison with the state of the art

Baselines. We compare R2P against state-of-the-art methods for VLM personalization, including MyVLM [4] and Yo’LLaVA [29], which require training to adapt to new personalized concepts, as well as RAP [13], a retrieval-based approach with a VLM instruction-tuned for personalized responses. In addition, we consider prompting-based strategies evaluated in [29]: GPT-4V + V_{prompt} , which is provided with both query and reference images for direct comparison; LLaVA [22], which relies solely on the pre-trained model with no personalized information; and LLaVA + $prompt$, where the model is prompted with a general description of the personal concept. We further evaluate MiniCPM-o + $prompt$ [47], a retrieval-based reasoning approach where the VLM infers the presence of a concept based on the query image and descriptions of the closest retrieved elements. To ensure fair comparison, we also include Yo’LLaVA w/ MiniCPM-o 2.6 as a baseline where we train the Yo’LLaVA algorithm replacing its VLM to MiniCPM-o 2.6.

Results on Recognition and Captioning. Table 1 reports recognition and captioning results on standard personalization benchmarks. On MyVLM, R2P achieves the best weighted accuracy (97.4%) and competitive negative accuracy, on par with RAP. Interestingly, MiniCPM-o + $prompt$ attains the highest positive accuracy but performs poorly on negative accuracy, suggesting overconfidence in answering “yes” regardless of query context.

On Yo’LLaVA, R2P achieves the highest captioning recall (87.1%), outperforming the next best by 5.5%. It also strikes a stronger balance in recognition, improving negative accuracy over the naive prompting baseline by +4.8%. Compared to RAP, R2P shows better generalization (+2.2% Wtd , +5.1% $Pos. Acc.$), despite being training-free.

Furthermore, R2P surpasses Yo’LLaVA w/ MiniCPM-o 2.6, confirming that the gains stem from our reasoning-based approach rather than backbone improvements. Overall, R2P offers a favorable trade-off between positive and negative accuracy, demonstrating the strength of our attribute-based verification. Captioning performance of MyVLM and Yo’LLaVA is notably lower, as both models rely on the concept name being explicitly provided in the prompt. Without access to this information, they often fail to generate captions that correctly reference or ground the intended concept.

On our challenging PerVA dataset, R2P continues to show strong recognition and captioning performance, driven by its use of fingerprint attributes and multimodal reasoning. It outperforms all baselines on three of four metrics, clearly outperforming the unimodal MiniCPM-o + $prompt$ baseline.

On recognition, R2P achieves 91.8% weighted accuracy, surpassing RAP by +2.8%. Compared to training-based models like MyVLM (+29.6%) and Yo’LLaVA (+19.8%),

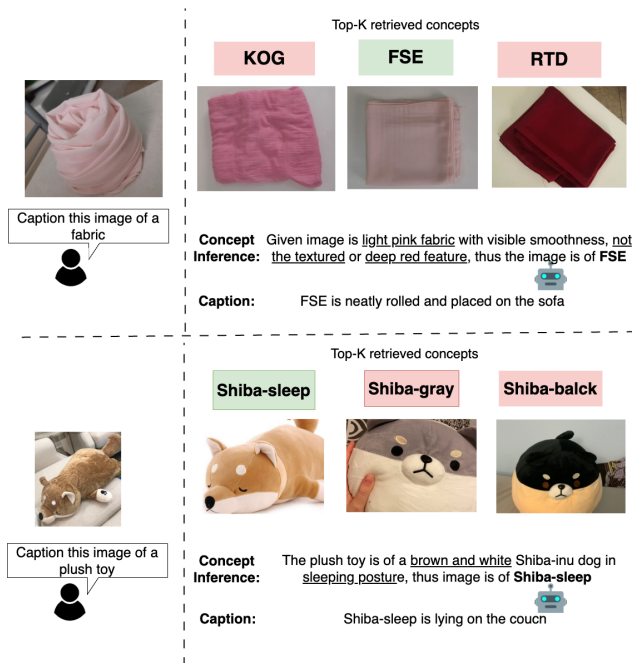


Figure 4. Qualitative examples. Given a query image and a user prompt (left), R2P retrieves the most similar Top-K concepts, analyzes a set of fingerprint attributes, and generates a precise, personalized caption. The key attributes enabling the model to recognize the correct concept name (bold) are underlined for clarity.

the gap widens further, highlighting the difficulty of learning robust concepts under limited reference images.

Similar trends are seen in captioning: R2P achieves 72.5% recall, outperforming RAP (64.1%) and MiniCPM-o + prompt (65.7%), where training-based models struggle with visually similar objects. These results reinforce the effectiveness of our training-free, attribute-guided strategy.

In Figure 4, we showcase examples of personalization. The top row displays an image of fabric from our PerVA dataset, while the bottom row features an image from the Yo’LLaVA dataset. Both examples illustrate how R2P accurately identifies the correct concept name for each object by reasoning over its fingerprint attributes, despite the presence of very similar candidate alternatives.

Results on personalized VQA. Table 2 reports the performance on personalized VQA on Yo’LLaVA dataset. We follow a similar evaluation protocol as in [29] and compare against baselines and competitors (results reported in the original paper). While LLaVA + prompt, Yo’LLaVA and RAP achieves similar performance, R2P scores the highest accuracy with a large improvement with respect to current state-of-the-art (+3.3% compared to RAP). Yo’LLaVA and RAP training strategies rely on assumptions about the possible questions to be asked. Instead, R2P imposes no priors, allowing any questions that might arise during inference

Method	Recognition			Captioning
	Pos. Acc.	Neg. Acc.	Wtd	Recall
MyVLM Dataset [4]				
MyVLM [4]	96.6	90.9	93.8	0.02
Yo’LLaVA [29]	<u>97.0</u>	95.7	96.4	0.1
RAP [13]	94.4	98.8	<u>96.6</u>	<u>88.0</u>
MiniCPM-o + prompt [47]	98.5	93.7	96.1	87.4
R2P (Ours)	96.3	<u>98.47</u>	97.4	91.4
Yo’LLaVA Dataset [29]				
Yo’LLaVA [29]	94.9	89.8	<u>92.4</u>	0.2
Yo’LLaVA w/ MiniCPM-o 2.6	62.8	62.2	62.5	-
GPT-4V + Vprompt [29]	80.9	99.2	90.1	-
RAP [13]	86.0	99.2	92.2	<u>79.7</u>
MiniCPM-o + prompt [47]	<u>91.2</u>	93.5	<u>92.4</u>	73.9
R2P (Ours)	91.1	97.7	94.4	87.1
PerVA Dataset				
MyVLM [4]	66.0	58.5	62.2	0.3
Yo’LLaVA [29]	75.1	69.0	72.0	6.6
RAP [13]	92.9	85.2	<u>89.0</u>	64.1
MiniCPM-o + prompt [47]	73.0	<u>92.5</u>	82.7	<u>65.7</u>
R2P (Ours)	<u>90.2</u>	92.8	91.8	72.5

Table 1. Recognition and captioning performance (↑) on established personalization benchmarks and the newly proposed PerVA.

Method	Accuracy
Yo’LLaVA [29]	92.9
RAP [13]	<u>93.2</u>
GPT-4V + Vprompt [29]	86.6
LLaVA [29]	89.9
LLaVA + prompt [29]	92.5
R2P (Ours)	96.5

Table 2. VQA performance in terms of answering accuracy (↑) in Yo’LLaVA [29] Dataset.

time. Compared to other training-free baselines based on LLaVA and GPT-4V, R2P more accurately identifies.

5.3. Ablation Study

In this section, we analyze the impact of key components of our method, including the role of pairwise reasoning, the introduction of fingerprint attributes, and VLM’s CoT reasoning (Tab. 3). Then, we ablate the cross-modal attribute verification strategy in comparison with recent alternatives (Tab. 4). Finally, we evaluate the role of different embeddings for the concept retrieval (Tab. 5). All experiments are conducted on the most challenging dataset, PerVA.

Impacts of main components. Tab. 3 presents the ablation on the main components of our proposed method on the recognition and captioning task. The first row reports results when we naively prompt the VLM to output the best-matched concept, given only the query image and the textual information of all candidate concepts, without CoT reasoning, without the use of fingerprint attributes, and without pairwise reasoning. Its performance on the recognition and captioning tasks are greatly limited, when compared to the sixth row (our full method R2P). Interestingly, when only ac-

Pairwise Reasoning	Fingerprint Attributes	Reasoning CoT	Recognition Wtd	Captioning Recall
X	X	X	86.5	62.2
X	X	✓	84.7	62.8
X	✓	✓	91.8	71.2
✓	X	X	92.3	65.9
✓	X	✓	91.6	67.3
✓	✓	✓	91.8	72.5
✓	privileged	✓	92.5	72.8

Table 3. Ablation on pairwise reasoning, fingerprint attributes, and the use of CoT reasoning for R2P in terms of weighted recognition metrics (\uparrow) and captioning recall (\uparrow) on PerVA. We include a privileged version of our approach with human pre-defined fingerprint attributes for concepts.

tivating CoT reasoning without fingerprint attributes (second row), the performance in recognition worsens compared to that of the first row. Yet, CoT reasoning tied with fingerprint attributes (third row), greatly improves the performance.

The fourth row reports results when only the computationally expensive pairwise reasoning is activated. Even without the use of fingerprint attributes, the recognition performance is already better than our method R2P at the sixth row, while its captioning performance remains limited. When CoT reasoning is leveraged together with pairwise reasoning without considering fingerprint attributes (the fifth row), we observe a slight performance gain in captioning, despite a minor drop in recognition. Eventually, the sixth row reports the final performance of R2P, obtaining the best performance in both recognition and captioning task. To investigate VLM-generated fingerprint attributes, we also report results in the last row when a human-specified pre-defined set (*e.g.*, color, shape, pattern, printed text) is provided during the database creation. Such attributes leverage privileged human knowledge to improve fine-grained recognition. Interestingly, R2P is almost on par with it, confirming that the VLM-generated fingerprint attributes are indeed informative and distinctive.

Verification strategies. We compare the proposed cross-modal attribute verification strategy with other approaches accounting for VLM uncertainty. In particular, we consider *abstention* where the model is prompted with the instruction to output “I am not sure” for uncertain cases, and *logits-based* evaluation considering the logits associated to the “I don’t know” option from the set of possible answers. For completeness, we show results when retrieved elements are entirely processed via the more demanding pairwise reasoning (*Pairwise-reasoning*) and when no verification step is performed, *i.e.*, never triggering the pairwise reasoning, (*No estimation*). Table 4 shows that our multimodal verification strategy achieves a recall of 72.5%, outperforming abstention (70.7%) and no estimation (71.2%) and providing additional gain with respect to the more demanding pairwise reasoning only. Our verification mechanism balances accuracy and efficiency, effectively reducing

Method	Captioning Recall
Pairwise-reasoning	72.3
No estimation	71.2
Abstention	70.7
Logits-based	70.9
Attr. Verification (Ours)	72.5

Table 4. Ablation on verification strategies for R2P on PerVA dataset. Performance is evaluated based on captioning recall (\uparrow).

Embedding	H@1	H@3	H@5	H@10
DINOv2	68.5	83.9	90.2	96.4
CLIP-Image	54.0	76.9	86.0	92.9
CLIP-Text	60.6	80.7	89.0	93.6
Multimodal (2-step)	65.0	<u>84.7</u>	<u>90.6</u>	93.6
R2P (Ours)	<u>67.5</u>	87.0	93.0	<u>96.3</u>

Table 5. Performance with different retrieval strategies evaluated in terms of HIT@K (\uparrow) on PerVA dataset.

hallucinations while maintaining competitive performance.

Retrieval. We assess the impact of different retrieval strategies for concept retrieval in R2P, as evaluated in terms of HIT@K metric measuring how often the personalized concept of interest is correctly retrieved in the top-K elements. We ablate different options of $K = 1, 3, 5, 10$. In particular, we compare representations from different pre-trained models such as DINOv2 [30] and CLIP [34]. For CLIP we consider both text and visual embeddings, denoted as CLIP Image and CLIP Text, respectively. For multimodal retrieval, we also ablate a two-step approach where concepts are first retrieved based on visual embeddings, and then re-ranked by relying on textual similarity. As shown in Table 5, DINOv2 achieves the highest H@1 performance (68.5%), while the proposed strategy in R2P fusion performs best when considering the Top-3 and Top-5 retrieved items with 87% and 93%, respectively. These results highlight the benefits of combining image and text embeddings for robust retrieval.

6. Conclusions

We explored training-free personalization of VLMs and proposed R2P, a novel approach that uses pre-trained VLMs to retrieve and reason over textual fingerprint attributes. R2P uniquely identifies user-specific concepts even in visually ambiguous scenarios. We demonstrated its effectiveness through extensive experiments on standard benchmarks and our new PerVA dataset, achieving state-of-the-art performance in personalized concept recognition and captioning. Future work will aim to reduce computational overhead and improve inference in cluttered scenes with similar-looking concepts.

7. Acknowledgments

This work was supported by the PNRR ICSC National Research Centre for HPC, Big Data and Quantum Computing (CN00000013), IPCEI Cloud (DM 27 June 2022 – IPCEI-CL-0000007) from the Italian Ministry of Enterprises and Made in Italy, and FAIR – Future AI Research (PE00000013), funded by the EU’s NextGeneration initiative. Additional support was provided by the EU projects ELIAS (No. 101120237) and ELLIOT (No. 101214398). We acknowledge IS CRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain.

References

- [1] Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*, 2024. 1
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv*, 2023. 1
- [3] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2
- [4] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *ECCV*, 2024. 1, 2, 3, 5, 6, 7, 4
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 1
- [6] Niv Cohen, Rinon Gal, Eli A Meir, Gal Chechik, and Yuval Atzmon. “this is my unicorn, fluffy”: Personalizing frozen vision-language representations. In *ECCV*, 2022. 2
- [7] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 2
- [8] Hazel Doughty and Cees GM Snoek. How do you do it? fine-grained action understanding with pseudo-adverbs. In *CVPR*, 2022.
- [9] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *CVPR*, 2020. 2
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv*, 2024. 1
- [11] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2
- [12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2
- [13] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant. In *CVPR*, 2025. 1, 2, 3, 5, 6, 7
- [14] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *NeurIPS*, 2023. 3
- [15] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2
- [16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. 2019. 6
- [17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2
- [18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2
- [19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2
- [20] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024. 3
- [21] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 3
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 6
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1
- [24] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024. 2
- [25] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 2
- [26] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 2
- [27] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 2
- [28] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 2
- [29] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo’LLaVA: Your personalized language and vision assistant. In *NeurIPS*, 2024. 1, 2, 3, 5, 6, 7, 4

- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 8
- [31] Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Personalized large vision-language models. *arXiv*, 2024. 2
- [32] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv*, 2024. 1, 2
- [33] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023. 2
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5, 8
- [35] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, 2023. 2
- [36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2
- [37] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where’s waldo: Diffusion features for personalized segmentation and retrieval. In *NeurIPS*, 2024. 2
- [38] Rohan Sarkar and Avinash Kak. Learning state-invariant representations of objects from image collections with state, pose, and viewpoint changes. *arXiv*, 2024. 5
- [39] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 2
- [40] Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. Personalized representation from personalized generation. *arXiv*, 2024. 2
- [41] Lv Tang, Peng-Tao Jiang, Haoke Xiao, and Bo Li. Towards training-free open-world segmentation via image prompt foundation models. *IJCV*, 2024. 2
- [42] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *ECCV*, 2022. 2, 4
- [43] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning—the good, the bad and the ugly. In *CVPR*, 2017. 2
- [44] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *CVPR*, 2022. 2
- [45] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023. 2
- [46] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv*, 2024. 6
- [47] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv*, 2024. 6, 7
- [48] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. 3
- [49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 6
- [50] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *ICLR*, 2024. 2
- [51] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *ECCV*, 2024. 2, 4
- [52] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *ICLR*, 2024. 3