Supplementary Materials MyVLM: Personalizing VLMs for User-Specific Queries

Yuval Alaluf[1,2], Elad Richardson[2], Sergey Tulyakov[1], Kfir Aberman[1], and Daniel Cohen-Or[1,2]

[1] Snap Inc.
[2] Tel Aviv University

Table of Contents

1 Societal Impact 1
2 Additional Details 2
2.1 Vision-Language Models 2
2.2 Training 2
2.3 Dataset & Experiments 3
3 MyVLM for Additional Applications 6
4 Additional Evaluations 9
4.1 Comparison to OpenFlamingo 9
4.2 Ablation Study: Augmentations & Regularization 10
4.3 Ablation Study: Concept Embedding Feature Space 10
4.4 Quantitative Evaluation: Image Captioning Metrics 13
4.5 Quantitative Evaluation: Concept Heads 14
5 Additional Qualitative Results 15

1 Societal Impact

The ability to personalize vision-language models offers more meaningful human-computer interactions, aligning them more closely with individual experiences and relationships. More generally, these personalized models may better guide users, catering to their unique needs. However, this personalization does come at the expense of privacy, granting the model access to potentially sensitive personal data. Additionally, there is a risk of users receiving harmful feedback regarding their personal content and relationships. As such, it is crucial to prioritize the protection of both user data and model behavior as we continue exploring the personalization of vision-language models.

2 Additional Details

2.1 Vision-Language Models

VLM Architectures. We use the implementation of BLIP-2 [14] provided in the transformers library [22] and employ BLIP-2 with the FLAN-T5 XL language model [5]. For LLaVA [15], we use the official implementation, employing LLaVA-1.6 with Vicuna-7B [4] as the language model. All models are run using half-precision to reduce memory requirements.

For generating the textual responses, we restrict the generated response to a maximum of 512 new tokens for both BLIP-2 and LLaVA. Additionally, for LLaVA, we use a temperature scale of 0.2 and set the top_p value to 0.7. All other parameters are set to their default values.

2.2 Training

Concept Head Training: People. To recognize user-specific individuals in images, we employ a pretrained face detector [6] and face recognition model [7]. Specifically, given a small set of images containing the subject (ranging from 1 to 4 images), we extract and store the face embeddings of the target individual. Then, given a new image, we extract embeddings from all detected faces and compare them with the stored face embeddings. If a new embedding falls within a predefined distance from the stored embeddings, we classify the corresponding individual as present in the image. We empirically set the distance threshold to 0.675. Note that each individual is associated with a separate concept head. However, features are extracted only once for each face detected in a new image.

Concept Head Training: Objects. For recognizing objects, we consider state-of-the-art large-scale vision models tailored for zero-shot classification and retrieval tasks, employing the recent DFN5B CLIP-ViT H/14 model [10, 18], implemented in the transformers library [22]. In contrast to the expressive face embedding space, we observed that directly using the image features extracted from these models is still not effective in distinguishing our personalized concepts from other similar objects (see Section 4.3). To address this, we train a single linear layer over the [CLS] token extracted from the frozen vision encoder. Training is performed to distinguish between 4 images containing the target concept and 150 negative images sourced from the internet depicting similar objects from the same general category. For example, when training the classifier to recognize a specific dog, we set the negative images to be images of arbitrary dogs.

Training is performed for 500 steps using a standard Cross Entropy loss for 500 steps with a batch size of 16. We use an AdamW optimizer with a learning rate of 0.001, decayed using a cosine annealing schedule. This converges in minutes, as only a single linear layer is trained.

At inference, given a new image, we first extract its image features from the frozen vision encoder, followed by applying all concept classifiers. Note that passing the features through all linear classifiers is notably faster than the feature extraction itself. We use a fixed threshold of 0.5 for all classifiers.

Concept Embedding Optimization. When applying MyVLM to BLIP, we perform 75 optimization steps for objects and 100 optimization steps for learning individuals. For LLaVA, we perform 100 optimization steps for both objects and individuals. For the optimization process, we use AdamW [16] with a constant learning rate of 1.0. We apply clip grad with a max L2 norm of 0.05, which we found helped stabilize convergence. For our regularization loss, we apply a weight factor of $\lambda = 0.04$ for BLIP and $\lambda = 0.25$ for LLaVA, set empirically.

To further stabilize the optimization process, we apply augmentations to both the input images and target captions, while fixing the language instruction ("Please caption this image of S_*."). For images, we apply random horizontal flips, random rotations, and brightness jittering. To augment the target captions, we ask an LLM [1] to generate four variations of the caption, while retaining the concept identifier. During each optimization step, one of the five augmented captions is randomly selected as the ground truth caption for computing the loss at the current step. This is designed to help disentangle the concept from a specific target output, mitigating overfitting and improving generalization to unseen contexts containing the concept.

For creating the augmented target captions, we provide GPT-4 [1] with the manually annotated target caption and ask it:

> *"Please provide four variations to the provided sentence. Please make the changes as small as possible and do not alter the word ⟨concept⟩."*

Choosing the Concept Identifier. We observed that the choice of identifiers for concepts can influence the results produced by MyVLM. For instance, using words that the model has difficulty generating, such as long words, may harm the results. Therefore, for personalizing outputs over objects, we follow the convention used for text-to-image personalization methods and set the concept identifier to "sks", introduced in [20].

For personalizing images over specific individuals, it is more natural to use common, short names as the concept identifiers. Therefore, we opt for "Bob" as a placeholder for males and "Anna" for females. We do note that other choices may be possible depending on the specific domain of the concept.

For VQA, to verify that the model does not rely on a gender bias via the concept name, we set the concept identifier to the word "sks" for both objects and individuals.

2.3 Dataset & Experiments

MyVLM Dataset. In total, we collected 45 user-specific concepts, consisting of 29 objects and 16 individuals. The dataset contains 350 images of objects and 330 images of individuals, each with a manually annotated personalized caption containing the concept identifier. Please note that written consent was provided by all individuals appearing in this work. To help facilitate further research into the personalization of VLM, the images and corresponding captions of all objects will be publicly available. We provide a sample image of each object in Figure 1.

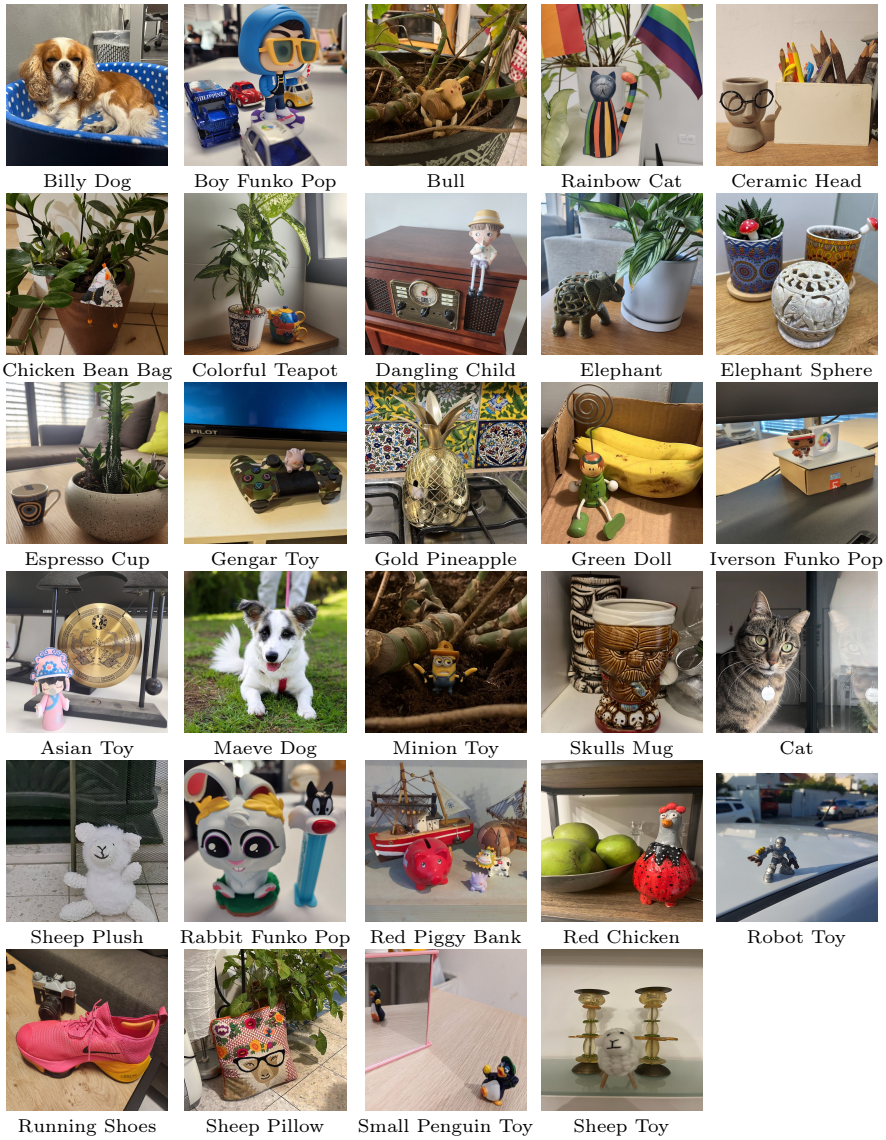


Fig. 1: MyVLM Dataset. Example images for each object in our constructed dataset.

Personalized Captioning Baselines. For our baselines, the keywords used for each concept are generated by GPT-4. Specifically, we provide GPT-4 a cropped image of the concept and prompt it with the following input:

> *Please provide 3 keywords for describing this object, each containing between one to three words.*

For our simple replacement-based baseline, we then try to insert the concept identifier into the original captions generated by BLIP-2 or LLaVA if one of the keywords is present in the caption.
For our LLM-based replacement baseline, we use Mistral-7B-Instruct-v0.2 [12] and prompt it with the following input:

> *I have the following sentence: ⟨original-caption⟩.*
> *Only if the word ⟨keyword⟩ appears in the sentence, please replace it with the word "sks".*
> *Otherwise, keep the original sentence. Can you do this for me? Please respond only with the corrected sentence.*
> *The output format will be "Revised: ⟨result⟩", with no additional text or explanations.*
> *Original Sentence: ⟨original-caption⟩*

Here, we use one of the keywords used for our simple replacement baselines. The output returned by Mistral is taken as the output of the LLM-guided baseline.

Evaluation Protocol. As mentioned in the main paper, we train our concept embeddings using five different seeds, each time sampling four different training samples and evaluating the remaining images. This resulted in a total of $2,429$ validation images — $1,164$ of user-specific objects and $1,265$ images of individuals.

For the training sets of individuals, we randomly select 4 images from the subset of images where the target subject appears alone. For objects, when training the concept embeddings, we use the same subset of 4 images used to train the linear classifier. This ensures that no validation image was seen neither when training the classifier nor when optimizing the concept embedding.

For computing the quantitative metrics, we use the following models. First, for the text-to-image similarity measure, we use CLIP ViT L/14 from OpenAI [9, 18] with an input resolution of 336×336. For computing our sentence similarity metric, we utilize a BERT [8] sentence transformer, taken from the SentenceTransformer library [19].

Table 1: A list of the 10 language instructions used when optimizing the concept embedding for personalized visual question-answering.

Objects	**People**
What color is ⟨concept⟩?	What is ⟨concept⟩ wearing in the image?
Where is ⟨concept⟩ in the image?	What color shirt is ⟨concept⟩ wearing?
Where is ⟨concept⟩ positioned in the image?	What is ⟨concept⟩ doing in the image?
Does ⟨concept⟩ appear to be the main subject of the image?	Where is ⟨concept⟩ in the image?
What objects is ⟨concept⟩ interacting with in the image?	Can you describe what ⟨concept⟩ is wearing?
How would you describe the texture of ⟨concept⟩ in the image?	From left to right, where is ⟨concept⟩ positioned in the image?
What types of materials is ⟨concept⟩ be made of?	What kind of hair does ⟨concept⟩ have?
Is ⟨concept⟩ large or small in the image?	What is the expression on ⟨concept⟩ face?
Is ⟨concept⟩ close to the camera or far away?	Is there anything unique about ⟨concept⟩'s appearance?
Please caption this image of ⟨concept⟩	Please caption this image of ⟨concept⟩

Personalized Visual Question-Answering. For personalized visual question-answering, we follow the same scheme as personalized captioning but alter the set of language instructions and targets used for optimizing the concept embedding. Specifically, we manually define a set of 10 prompts used as the language instructions used during optimization, detailed in Table 1. To obtain the target for each question, we pass the image and language instruction to the original LLaVA model, setting its output to the target answer. Then, at each training step, we randomly select one of the 10 prompts and targets.

We do note that this may introduce some unwanted bias into the optimization process, as LLaVA may not always accurately answer the given question. As such, alternative approaches for expanding the set of language instructions and targets may achieve better results. We leave this exploration for future work.

3 MyVLM for Additional Applications

Personalized Referring Expression Comprehension. In this section, we demonstrate the applicability of MyVLM for an additional personalized task: referring expression comprehension (REC) [17], which involves localizing a target subject in a given image. To achieve this, we utilize MiniGPT-v2 [13], a recent VLM that can naturally handle various vision-language tasks by employing different task identifiers to define the language instructions passed to the language model. As MiniGPT-v2 shares the same architecture as LLaVA [15], we adopt the same training setup for learning our concept embeddings. Specifically, to optimize the concept embedding, we follow the same scheme as used for personalized captioning and use the language instruction:

"[caption] Please caption this image of S_* ".

Fig. 2: Personalized REC results obtained by MyVLM over MiniGPT-v2 [13]. Sample images of the target concept are provided in the top row. Bounding box coordinates returned by the personalized VLM are drawn in green. Below each image, we also present the personalized captions outputted by MyVLM by passing MiniGPT-v2 a captioning instruction.

During inference, to solve for REC we modify the language instruction to:

"[refer] S_* in the image",

which returns the bounding box coordinates of the target subject within the provided image. We emphasize that this is achieved with only the captioning supervision during optimization. This builds on the inherent ability of the underlying VLM to solve for multiple tasks while highlighting that the learned concept embedding does indeed capture the semantic representation of the concept which the model can reuse for its different tasks.

In Figure 2, we present personalized results for referring expression comprehension (REC) and captioning achieved by MyVLM using MiniGPT-v2 [13]. As shown, MyVLM cannot only generate personalized captions but also pinpoint the concept within the image without any direct supervision on the localization task. Importantly, the ability of MiniGPT-v2 to accommodate multiple tasks through distinct task identifiers enables MyVLM to be extended naturally to additional personalized applications with minimal modifications.

Fig. 3: Comparison to the OpenFlamingo [2, 3] for personalized captioning. We apply MyVLM over both BLIP-2 [14] (top) and LLaVA [15] (bottom) images of the target concept are shown above each image.

4 Additional Evaluations

4.1 Comparison to OpenFlamingo

Following our qualitative comparison to GPT-4 [1] in the main paper, we now compare to OpenFlamingo, which also supports interleaved image and text inputs. We do so both qualitatively and quantitatively.

Baseline Setup. We use the open-source implementation of Flamingo [2, 3]. We use CLIP-ViT H/14 [9, 18] as the vision encoder and MPT-1b-RedPajama-200b [21] as the language model. We provide Flamingo with a cropped image of the concept and provide it with the following language instruction:

"<image>This is S_*.</endofchunk|><image>In this image you can see"

Here, we replace S_* with the word "bloby" for objects and replace S_* with either "Bob" or "Anna" for individuals. We explored other suffixes but found the most consistent results with the prompt above. Metrics were computed following the same protocol as used in the main paper by aggregating results over all concepts and across all five validation folds.

Qualitative Comparison. In Figure 3 we show a visual comparison of personalized caption results obtained OpenFlamingo and MyVLM. As can be seen, OpenFlamingo, particularly for objects, struggles in both identifying the target subject and contextualizing it within its surroundings. For example, OpenFlamingo recognizes the sheep figurine and cat statue in the first column but is unable to generate a caption that aligns with the input image. In addition, OpenFlamingo can still struggle to incorporate the concept identifier within the caption as seen in the third row. In contrast, MyVLM, over both BLIP-2 and LLaVA successfully recognizes the target concept while generating accurate captions that correctly communicate information about the concept to the user while remaining aligned with the input image.

Quantitative Comparison. Next, in Table 2 we present quantitative results, comparing the results obtained by Flamingo with those obtained with MyVLM over both BLIP-2 [14] and LLaVA [15]. First, in terms of the ability to capture the concept identifier in new captions, MyVLM outperforms OpenFlamingo when applied to both BLIP-2 and LLaVA. This improvement in recall is most notable for user-specific objects, where MyVLM outperforms OpenFlamingo by over 45%. For the CLIPScore between the generated captions and input images, all three methods attain comparable results for both objects and people, with a maximum difference of 1.34% between the three. However, as can be seen, there is a significant difference in the sentence similarity between captions generated by MyVLM and those generated by OpenFlamingo. Specifically, for people, MyVLM over BLIP-2 outperforms OpenFlamingo by over 5% and by over 40% when personalizing captions for user-specific objects. These results, along with the visual results presented above, further highlight the advantage of our approach in learning a dedicated embedding vector to represent our concepts.

Table 2: Quantitative Comparison: OpenFlamingo [2, 3]. We compute the average recall, text-to-image similarity, and text-to-text similarity obtained over all 16 individuals and 29 objects. Results are averaged across all five validation sets.

Data	Model	Recall ↑	Text Similarity ↑	Image Similarity ↑
People	OpenFlamingo	74.81	<u>43.72</u>	**24.33**
	MyVLM + BLIP-2	<u>79.76</u>	**48.99**	22.99
	MyVLM + LLaVA	**97.08**	43.58	<u>23.06</u>
Objects	OpenFlamingo	49.77	34.12	<u>27.65</u>
	MyVLM + BLIP-2	**95.10**	**77.71**	**28.12**
	MyVLM + LLaVA	<u>94.76</u>	<u>71.49</u>	27.60

Table 3: Ablation Study: Regularization & Augmentations. We compute the average recall, text-to-image similarity, and text-to-text similarity obtained over 5 objects and 5 individuals with and without our augmentations and regularization techniques. Results are obtained over BLIP-2 and averaged across all five validation sets.

	Recall ↑	Text Similarity ↑	Image Similarity ↑
w/o Aug. & Reg.	25.88	<u>56.32</u>	**24.76**
w/o Aug.	<u>72.77</u>	55.03	24.00
MyVLM	**84.87**	**58.68**	<u>24.65</u>

4.2 Ablation Study: Augmentations & Regularization

Here, we validate the contribution of the augmentations and regularization applied during the training of the concept embeddings. In Table 3, we present personalized captioning results for 10 concepts obtained using MyVLM over BLIP-2 [14]. Incorporating the attention-based regularization loss improves recall by a significant margin (∼45%). Furthermore, employing augmentations over both the image and target captions leads to an additional improvement of approximately 12% in recall. Moreover, applying both regularization and augmentations improves the text similarity with respect to the target caption, while attaining a comparable CLIPScore [11] to cases where these techniques are not applied. We believe that further exploration into additional augmentations and attention-based manipulations can offer insights into further extending the capabilities of MyVLM.

4.3 Ablation Study: Concept Embedding Feature Space

Next, we explore the use of linear classifiers to serve as our concept heads for personalizing user-specific objects. Focusing on BLIP-2, we analyze two alternative feature spaces and show that operating directly within these feature spaces is not sufficient to distinguish the target concept from other semantically similar objects. First, we examine the output space of the BLIP-2 vision encoder. We then explore the embedding space of the DFN5B CLIP-ViT H/14 model [10, 18], used as our base feature extractor, showing that it too is not expressive enough to be used directly.

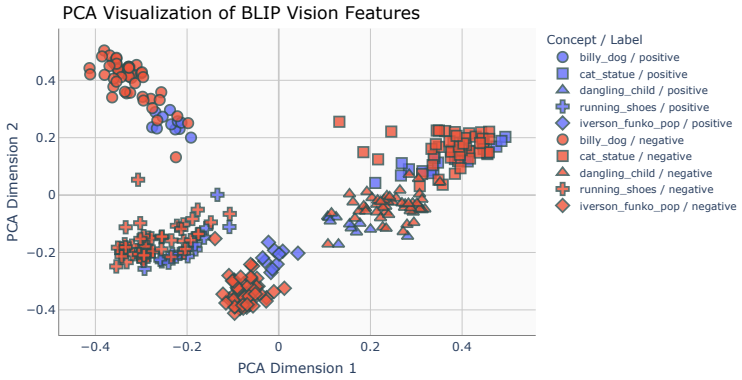


Fig. 4: PCA Visualization of the output space of the BLIP-2 vision encoder. We project the [CLS] token embeddings extracted from all positive and 200 negative images of five different objects, each shown using a different shape. As shown, these embeddings are not well-separated enough to effectively distinguish between positive and negative samples of the target object.

In Figure 4 we perform PCA over embeddings extracted from images of five user-specific objects alongside 200 negative samples for each object. As can be seen, for each object, represented by a different shape, there is no clear separation between the positive and negative samples. This suggests that relying solely on a distance measure directly over this feature space is insufficient for distinguishing between new images that may contain the target concept.

Next, we evaluate the more expressive CLIP space, designed for zero-shot retrieval. In Figure 5, we visualize the nearest neighbors of various positive images. As shown, CLIP is unable to focus on retrieving the target concept, especially when other objects are present in the same image. Moreover, determining an optimal threshold for each concept without calibration is challenging, particularly if only very few samples of the object are available.

As discussed in the original CLIP paper [18], these challenges can be mitigated using linear heads. This is also evident with our concept heads. Specifically, in Figure 5, we present the top five images that received the highest scores from our classifier for each of the four depicted concepts. As can be seen, our classifiers can effectively distinguish the target concept from semantically similar objects while enabling us to use a fixed threshold across all concepts. This further validates the use of linear classifiers for constructing our concept heads and recognizing user-specific objects.

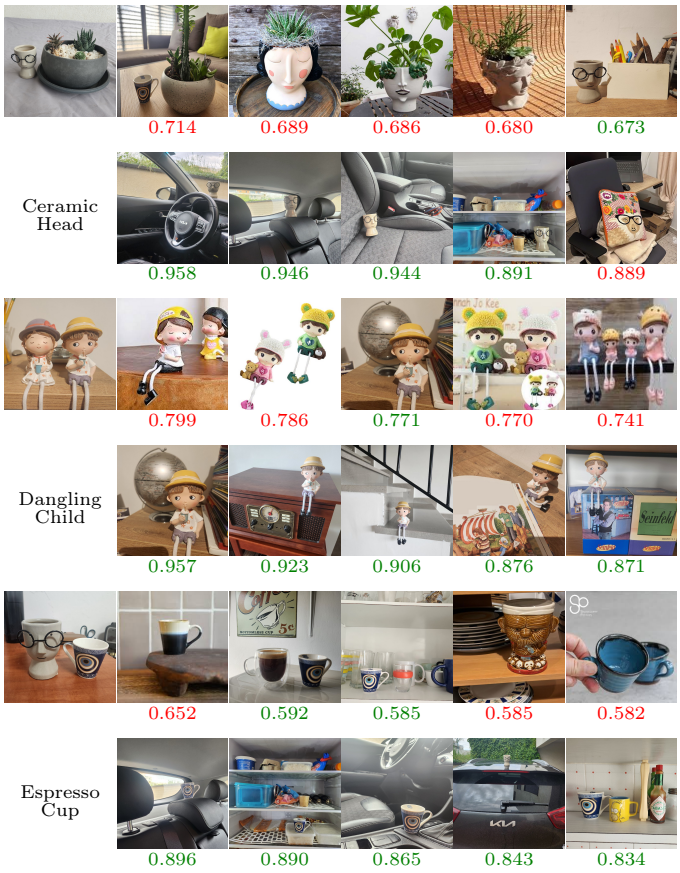


Fig. 5: Ablation Study: The CLIP Space. For each concept, we visualize the 5 nearest neighbors of the query image shown to the left within the CLIP embedding space. The nearest neighbors often include both negative samples of the target object and positive samples of other objects, making it challenging to directly operate within the space. In the second row of each concept, we visualize the five images that received the highest scores from the concept's linear head. As shown, our linear classifier is effective in distinguishing the target concept from negative samples.

Table 4: Quantitative Metrics: Standard Image Captioning Metrics. We compute standard image captioning metrics over personalized captions generated by MyVLM, trained with 4 images. For each image, we use all 5 augmented captions as the set of ground truth captions. Results are obtained over all 5 validation folds and averaged over all concepts.

Dataset	Method	B1	B2	B3	B4	CIDEr	METEOR	ROUGE_L	SPICE
People	BLIP-2	0.69	0.63	0.58	0.53	2.21	0.31	0.63	0.27
	MyVLM	0.53	0.40	0.30	0.23	1.06	0.21	0.44	0.15
Objects	BLIP-2	0.63	0.51	0.43	0.36	1.53	0.26	0.55	0.23
	MyVLM	0.64	0.50	0.38	0.28	1.44	0.28	0.56	0.26
All	BLIP-2	0.66	0.57	0.51	0.45	1.89	0.28	0.59	0.25
	MyVLM	0.59	0.45	0.34	0.26	1.28	0.25	0.50	0.20

BLIP-2

Dataset	Method	B1	B2	B3	B4	CIDEr	METEOR	ROUGE_L	SPICE
People	LLaVA	0.27	0.14	0.08	0.04	0.18	0.11	0.24	0.06
	MyVLM	0.28	0.19	0.13	0.09	0.39	0.18	0.34	0.11
Objects	LLaVA	0.26	0.15	0.09	0.05	0.15	0.16	0.27	0.11
	MyVLM	0.36	0.26	0.19	0.13	0.73	0.26	0.44	0.21
All	LLaVA	0.26	0.15	0.08	0.05	0.17	0.13	0.26	0.09
	MyVLM	0.32	0.22	0.15	0.11	0.58	0.22	0.39	0.16

LLaVA

4.4 Quantitative Evaluation: Image Captioning Metrics

Next, we validate the performance of MyVLM on standard image captioning metrics to ensure it does not compromise the general capabilities of the underlying VLM. The results are presented in Table 4. It is worth noting that the target captions were initially generated using BLIP-2 and then manually adjusted as necessary. This process inherently introduces a bias towards favoring captions generated by BLIP-2, which can be seen from the performance gap between results obtained with BLIP-2 and LLaVA. Despite this bias, MyVLM still achieves similar performance on most captioning metrics when considering all 45 concepts. This behavior can also be seen when considering LLaVA, where MyVLM achieves comparable performance on both people and objects. These results further highlight that MyVLM effectively preserves the original captioning capabilities of the frozen VLM.

Table 5: Concept Head Evaluations. Left: we measure the recall and classification rate over 16 individuals using our face recognition network used for defining our concept head. Right: we compute the average recall and precision of our linear classifiers over our 29 user-specific objects.

Recall	False Positive Rate	Miss Rate
96.39%	2.33%	1.28%

People

	Correctly Classified	Total Samples	Percent Correct
Positives	226	234	96.58%
Negatives	95,724	105,328	90.88%

Objects

4.5 Quantitative Evaluation: Concept Heads

Finally, we assess the effectiveness of our concept heads along two fronts. First, we verify their ability to support multiple concepts within the same VLM. Second, we evaluate the recall and precision of our concept heads, validating their performance both on new positive images of the concept and on negative images that do not contain the target concept.

To evaluate our ability to support multiple concepts simultaneously, we evaluate our concept head performance on 16 individuals. We calculate three metrics: (1) the percentage of images correctly classified as the correct individual, (2) the percentage of images misclassified as the incorrect individual, and (3) the percentage of images not identified as any of the known individuals. These metrics are computed across all individuals using the same five validation folds used for the main evaluations presented in the paper. The average results are presented in Table 5. As shown, leveraging the pretrained face recognition model as our concept head achieves impressive performance, achieving a recall of over 96% while falsely classifying an individual in only 2% of all images. The ability of the model to accurately distinguish different individuals naturally allows us to support multiple individuals using a single trained model. This in turn allows us to scale to new individuals over time by simply adding a new concept head for the desired individual, highlighting the benefit of using external experts for recognizing our concepts.

Next, we validate the performance of our linear classifiers, examining whether they can generalize to new images of our concept while effectively filtering out non-relevant images that do not contain the target concept. To do so, we consider a single validation fold for each of the 29 object concepts. To measure recall, we compute the percent of positive validation samples correctly identified by the classifier. To measure precision, we consider all positive images of *other* concepts, and all negative images of *all* concepts. We then compute the number of negative samples incorrectly classified as the target concept. This is process is repeated for each object. The total and average recall and precision results are presented in Table 5. As illustrated, we attain an average recall of 96% with a precision of 91%, computed over 100,000 negative samples. This highlights the ability of our linear classifiers to correctly classify new images, both those containing our concept and those that do not.

5 Additional Qualitative Results

In the remainder of this document, we provide additional results and comparisons, as follows:

1. In Figures 6 and 7, we provide additional personalized captioning results obtained by MyVLM over BLIP-2 [14].
2. In Figures 8 and 9, we present additional personalized captioning results of MyVLM over LLaVA [15].
3. In Figure 10, we provide additional comparisons over BLIP-2 with our alternative captioning baselines, both the simple replacement technique and the LLM-guided approach.
4. In Figures 11 and 12, we present additional visual comparisons to both baselines, applied over LLaVA.
5. In Figures 13 and 14, we show personalized captioning obtained by MyVLM over both BLIP-2 and LLaVA on the same set of images, highlighting MyVLM's applicability to both architectures.
6. In in Figures 15 and 16, we show additional personalized visual question-answering results obtained by MyVLM applied over LLaVA.
7. Finally, in Figure 17, we present additional results for personalized Referring Expression Comprehension obtained using MyVLM with MiniGPT-v2 [13].

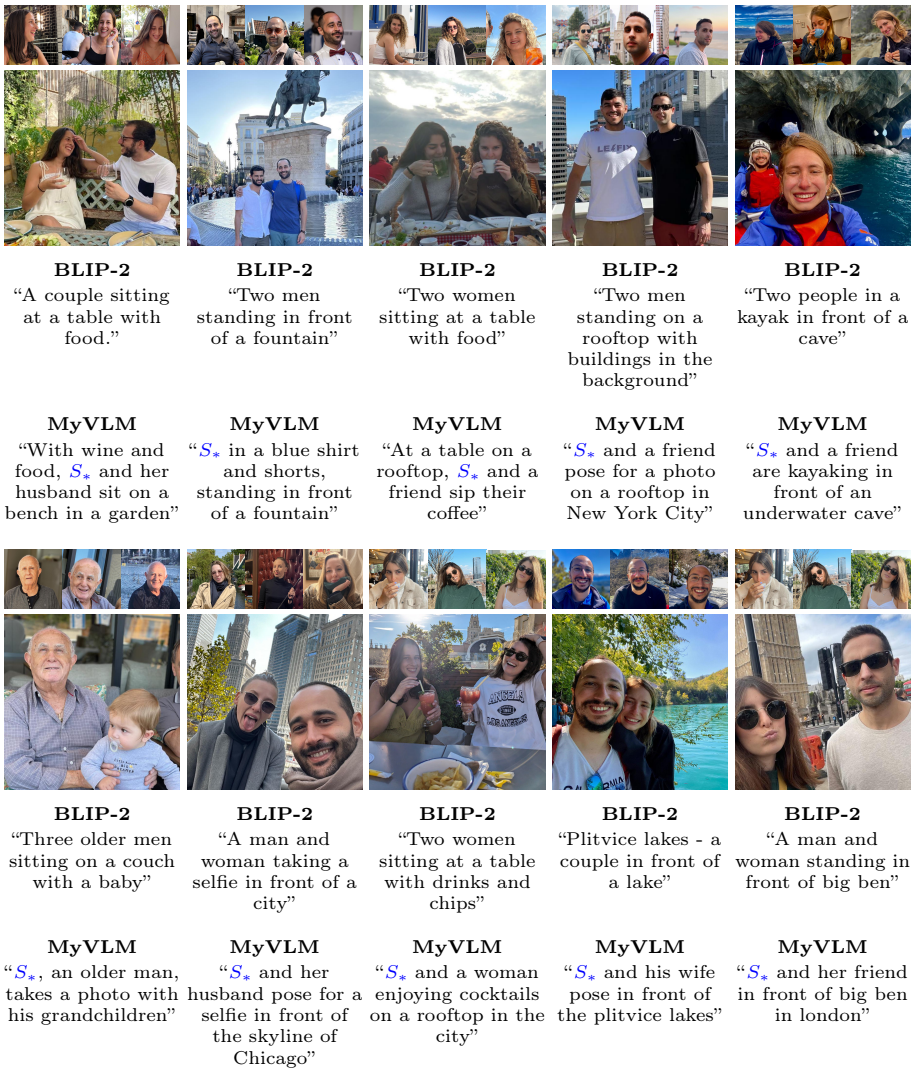


Fig. 6: Additional personalized captioning results obtained by MyVLM, applied over BLIP-2 [14]. Example images of the target concept are provided in the top row.

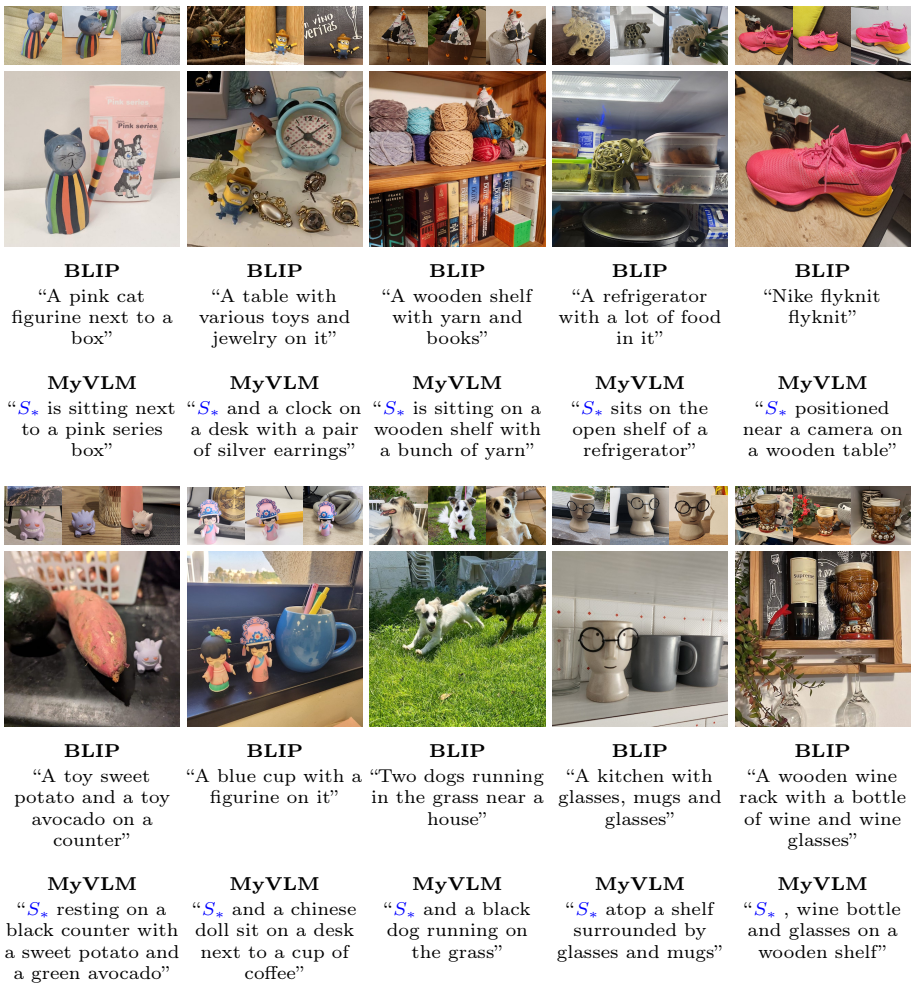


Fig. 7: Additional personalized captioning results obtained by MyVLM, applied over BLIP-2 [14]. Example images of the target concept are provided in the top row.

LLaVA
"Enjoying a meal outdoors with a smile on her face"

LLaVA
"Embracing the serene beauty of the harbor, this traveler finds solace in the iconic Sydney Opera House..."

LLaVA
"A moment of joy and love shared between two people, captured in a warm embrace"

LLaVA
"Smiling and posing in front of a grand entrance, ready to make an impression"

LLaVA
"Friends enjoying a sunny day with refreshing orange drinks and good conversation"

MyVLM
"On a street in a city, S_* is sitting at a table outside a restaurant... S_* is wearing a black and white patterned top."

MyVLM
"S_* is standing on a bridge overlooking the Sydney Opera House and the Sydney Harbour Bridge. He is wearing a blue denim jacket and sunglasses"

MyVLM
"S_* and his girlfriend, smiling and hugging each other, in a restaurant."

MyVLM
"S_* with a yellow dress, standing in front of a grand building with an arched doorway and intricate metalwork..."

MyVLM
"S_* and a friend at a restaurant, holding up their glasses of orange wine"

LLaVA
"Caught in the city's embrace, a moment of calm amidst the hustle and bustle."

LLaVA
"Taking a moment to enjoy the view and a warm beverage in the heart of the city"

LLaVA
"Friendship is the best wine"

LLaVA
"Friends sharing a moment in the heart of the city, surrounded by the beauty of history and architecture"

LLaVA
"Enjoying a meal outdoors at a charming restaurant, with a view of the street and the sky above"

MyVLM
"S_*, in a blue suit, poses for a portrait at a scenic spot overlooking a river with a bridge in the distance"

MyVLM
"S_*, in a black coat, sits on a bench in Central Park, enjoying a coffee"

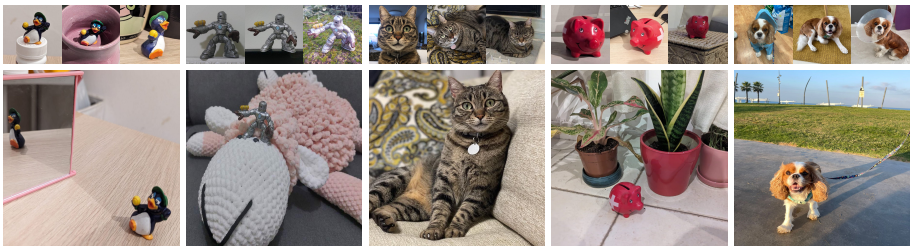
MyVLM
"Sitting at a table, S_* and her friend smile at each other as they clink their wine glasses together"

MyVLM
"S_* and his friend at a fountain. They pose for a photo. S_* wearing a blue shirt and white pants. At a fountain in a city square..."

MyVLM
"S_* sits at a patio table laden with a meal, enjoying a sandwich and fries with a side of coleslaw"

Fig. 8: Additional personalized captioning results obtained by MyVLM, applied over LLaVA [14]. Example images of the target concept are provided in the top row.

LLaVA	**LLaVA**	**LLaVA**	**LLaVA**	**LLaVA**
"a bottle of supreme ginetta savignon wine, standing tall next to a chalkboard wall adorned with wine-themed illustrations..."	"A well-stocked refrigerator, ready for a weekend of culinary adventures!"	"A whimsical scene of creativity and imagination, featuring a colorful origami bird perched on a wooden table, surrounded by vibrant children's books..."	"A moment of calm before the caffeine rush: two cups of coffee, one classic and one modern, sit side by side on a pristine white countertop..."	"friendly fidos: two dogs, one white and one black, pose for a photo on a grassy lawn..."
MyVLM	**MyVLM**	**MyVLM**	**MyVLM**	**MyVLM**
"S_* next to a bottle of wine on a wine themed wall"	"S_* sits comfortably on the second shelf of an open refrigerator, ready to be stocked with food..."	"S_* sitting next to a colorful children's book on a table"	"S_* sitting next to a cup of coffee on a desk in a room with a "bottomless cup" sign in the background"	"S_* is standing on the grass with a big smile and a wagging his tongue"



LLaVA	**LLaVA**	**LLaVA**	**LLaVA**	**LLaVA**
"Reflection of a penguin figurine in a pink mirror, standing on a wooden table"	"A whimsical scene of a robotic adventure: a small astronaut riding a pink sheep with a white face, set against a cozy gray couch"	"Relaxed and Ready for Adventure: A Tiger Cat's Pose of Serenity"	"A cozy corner of a room, where potted plants and a little pink piggy bank share a space, creating a charming atmosphere"	"A Cavalier King Charles Spaniel puppy enjoys a sunny day at the beach, wearing a colorful collar and leash..."
MyVLM	**MyVLM**	**MyVLM**	**MyVLM**	**MyVLM**
"S_* sitting in front of a mirror on a table, reflecting their own image in the mirror"	"S_* is sitting on a stuffed animal that looks like a sheep. The sheep is pink and white, and S_* is wearing a silver outfit"	"S_* sitting on a beige couch, looking up at the camera with a curious expression"	"S_* sitting on a white floor next to a potted plant and a pink pot, both in front of a curtain"	"S_* walking on a leash in a park near the beach with palm trees in the background"

Fig. 9: Additional personalized captioning results obtained by MyVLM, applied over LLaVA [14]. Example images of the target concept are provided in the top row.

Simple N/A	**Simple** "S_* standing in front of a mountain with a glacier"	**Simple** N/A	**Simple** N/A	**Simple** "Two S_* laying in a pink and blue dog bed"
LLM-Guided "Two S_* standing on a rooftop with buildings in the background"	**LLM-Guided** "S_* standing in front of a mountain with a glacier"	**LLM-Guided** "Two S_* sitting at an outdoor table with food and drinks"	**LLM-Guided** "Two S_* are holding glasses of orange juice"	**LLM-Guided** "Two S_* laying in a pink and blue dog bed"
Ours "S_* and a friend pose for a photo on a rooftop in New York City"	**Ours** "S_* in a gray shirt is standing in front of a mountain with a glacier in the background"	**Ours** "S_* and a friend enjoying coffee and a sandwich at a cafe"	**Ours** "With two glasses of orange juice, S_* and her friends are enjoying a summer day on a balcony overlooking the city"	**Ours** "S_* and a dog rest in a dog bed in a room"

Simple N/A	**Simple** N/A	**Simple** "S_* tiki mugs"	**Simple** "A S_* with a skull on it"	**Simple** "Two S_*'s sitting on a chair in front of a window"
LLM-Guided "A shelf with mugs, glasses, and S_* on it"	**LLM-Guided** "A pink S_* figure next to a box"	**LLM-Guided** "S_* tiki mugs"	**LLM-Guided** "A S_* with a skull on it"	**LLM-Guided** "Two S_*'s sitting on a chair in front of a window"
Ours "S_* on a shelf with various glasses and cups"	**Ours** "S_* is sitting next to a pink series box"	**Ours** "S_* on a shelf next to tiki vases"	**Ours** "S_* on a shelf with a tiki mug"	**Ours** "S_* is laying on the couch, with its head resting on the arm of the chair"

Fig. 10: Additional comparisons to our personalized captioning baselines. Results are obtained over BLIP-2 [14]. Sample images of the target concept are shown in the top row.

Simple N/A	**Simple** N/A	**Simple** N/A	**Simple** N/A	**Simple** N/A
LLM-Guided "S_*-perfect companion: playful pairing of gaming and furry friends"	**LLM-Guided** "A charming scene of a S_* sheep figurine resting in a potted plant, adding a touch of whimsy to any space"	**LLM-Guided** "A cozy outdoor setting with a touch of whimsy: a wooden table, a cactus in a S_* , and a pair of chairs,..."	**LLM-Guided** "a cozy scene with a soft, pink S_* and a white lamb, ready for a nap on a gray couch"	**LLM-Guided** "A collection of seinfeld memorabilia, including a S_* and dvd boxes, arranged on a shelf"
Ours "S_* sitting on top of a camouflage video game controller in front of a TV"	**Ours** "S_* tucked between leaves and branches of a houseplant"	**Ours** "S_* sitting on a wooden chair at a wooden table on a patio, with a bamboo fence..."	**Ours** "S_* sitting on the couch with a pink and white stuffed animal next to it"	**Ours** "S_* sitting on a shelf in front of a Seinfeld box set, with a surprised expression..."

Simple N/A	**Simple** "A blue cup of tea, a pair of S_*s, and a pen, all resting on a window sill..."	**Simple** N/A	**Simple** N/A	**Simple** N/A
LLM-Guided "Let's set sail with our wooden pirate ship and our friendly wooden animals. who will be the first to reach the S_* ?"	**LLM-Guided** "A blue cup of tea, a pair of S_* figurines, and a pen..."	**LLM-Guided** "A trio of S_*s, each with its own unique color and style, standing side by side on a tiled floor."	**LLM-Guided** "Embracing the chill: a S_* winter adventurer stands in awe of the icy cave..."	**LLM-Guided** "Sunny day, sunglasses on, S_* checking my phone for the perfect shot."
Ours "S_* against a backdrop of a toy ship and a small toy"	**Ours** "S_* and another chinese doll standing next to a blue mug with pink and yellow accents"	**Ours** "S_* with two other pairs of nike sneakers on the floor next to a white wall"	**Ours** "S_*, smiling in a blue jacket, stands in front of a large ice cave with icicles hanging from the ceiling"	**Ours** "As S_* takes a break from his day, S_* takes a moment to capture the moment"

Fig. 11: Additional comparisons to our personalized captioning baselines. Results are obtained over LLaVA [14].

Simple N/A	**Simple** N/A	**Simple** N/A	**Simple** N/A	**Simple** N/A
LLM-Guided "Embracing the serenity of nature's beauty: a S_* moment of joy and tranquility at the edge of the lake, under the watchful gaze of majestic mountains"	**LLM-Guided** "Creating with clay: a S_* artist at work"	**LLM-Guided** "Embracing the cityscape from a high vantage point, this S_* urban explorer finds joy in the bustling city life"	**LLM-Guided** "A S_* enjoying a warm beverage on a chilly day, taking a moment to savor the comfort of the cozy café ambiance"	**LLM-Guided** "Sitting comfortably in a vibrant red rocking chair, this elderly S_* enjoys a moment of relaxation amidst the city life"
Ours "Standing by a lake, S_* smiles at the camera, surrounded by nature and mountains"	**Ours** "S_* is painting a green ceramic bowl at a wooden table"	**Ours** "S_* wearing a green sweater and sunglasses, poses on a rooftop during winter"	**Ours** "S_* enjoying a warm beverage at a cafe, surrounded by the hustle and bustle of city life"	**Ours** "S_* sits on a red wooden rocking chair outside, overlooking a row of colorful chairs under a clear blue sky"

Fig. 12: Additional comparisons to our personalized captioning baselines. Results are obtained over LLaVA [14]. Sample images of the target concept are shown in the top row.

MyBLIP "S_* and her husband pose for a selfie in front of the skyline of Chicago"

MyBLIP "S_* and a friend are kayaking in front of an underwater cave"

MyBLIP "S_* is in Sydney, looking at the Sydney opera house and the harbour bridge"

MyBLIP "S_*, on a bridge overlooking the london eye, in a pair of red shorts"

MyBLIP "S_*, standing on the rooftop of the hotel, with a margarita and a t-shirt."

MyLLaVA "S_* and her companion are standing in front of a city skyline, with S_* making a playful gesture with her tongue..."

MyLLaVA 'S_* and a man are in front of a glacier, with a rocky shore in the background"

MyLLaVA "S_* is standing on a bridge overlooking the Sydney Opera House and the Sydney Harbour Bridge. He is wearing a blue denim jacket and sunglasses."

MyLLaVA "S_*, wearing sunglasses, posing for a photo in front of the London Eye

MyLLaVA "S_*, laughing and enjoying her drink, is wearing a white t-shirt with the word "Angels" and the year "1961" on it. She's also wearing sunglasses and has a straw in her drink."

MyBLIP "At the restaurant, S_* sits at the table with a plate of food"

MyBLIP "S_* and her dog, a white dog, sit on a table"

MyBLIP "With wine and food, S_* and her husband sit on a bench in a garden"

MyBLIP "S_* sits on the balcony of her apartment"

MyBLIP "S_*, wearing a hat, sits on a boat in the ocean"

MyLLaVA "S_*, sitting at a table, has a spoonful of food, poses for a photograph on the dining room"

MyLLaVA "S_* is sitting outside a cafe, wearing a red and white striped shirt and a white top, with a dog on a leash."

MyLLaVA "Sitting on a bench, smiling, and holding a glass of wine, S_*, with a man in a white t-shirt and glasses, enjoying a meal and a conversation outdoors"

MyLLaVA "S_* sits on a patio chair under a tree, wearing a pink skirt and a white top, with a blue door in the background"

MyLLaVA "S_* in a boat, wearing a hat and sunglasses, enjoying a relaxing day on the water"

Fig. 13: Additional personalized captioning results obtained by MyVLM applied over both BLIP-2 [14] and LLaVA [15].

Fig. 14: Additional personalized captioning results obtained by MyVLM applied over both BLIP-2 [14] and LLaVA [15].

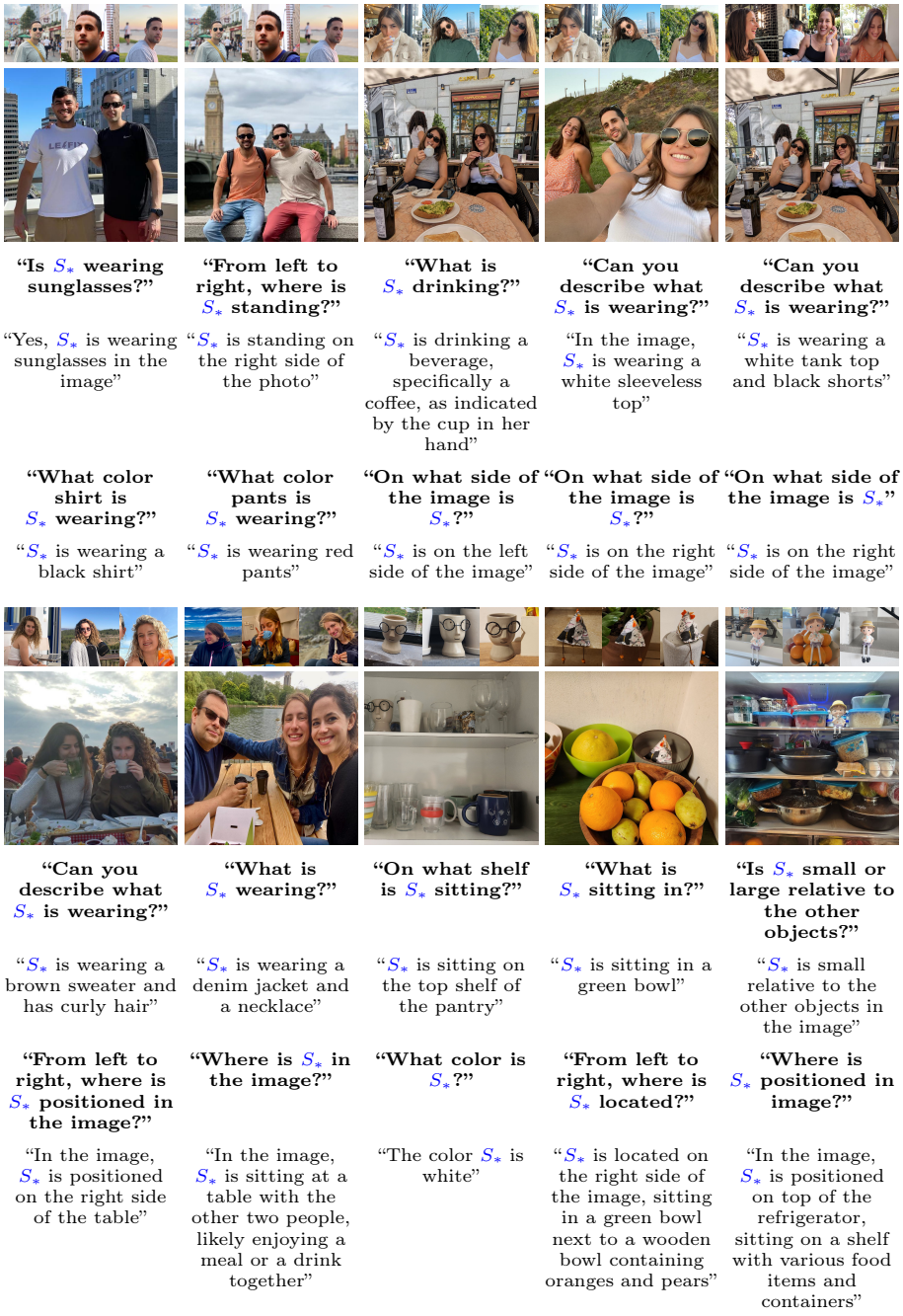


Fig. 15: Additional personalized visual question-answering results obtained by MyVLM, applied over LLaVA [15].

Fig. 16: Additional personalized visual question-answering results obtained by MyVLM, applied over LLaVA [15].

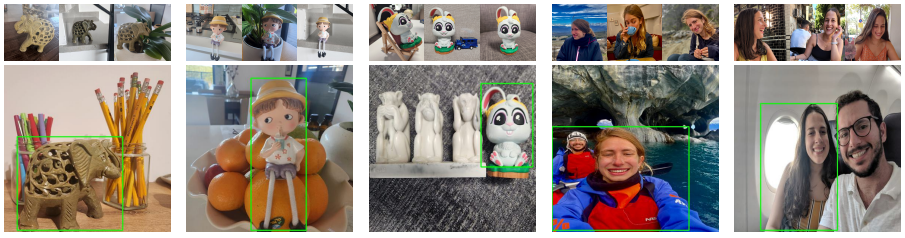


Fig. 17: Additional personalized REC results obtained by MyVLM over MiniGPT-v2 [13]. Sample, cropped images of the target concept are provided in the top row. Bounding box coordinates returned by the personalized VLM are drawn in green. Below each image, we also present the personalized captions outputted by MyVLM by passing MiniGPT-v2 a captioning instruction.

References

1. Gpt-4 technical report (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
5. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
6. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
7. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). https://doi.org/10.1109/tpami.2021.3087709, http://dx.doi.org/10.1109/TPAMI.2021.3087709
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
11. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://aclanthology.org/2021.emnlp-main.595
12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
13. Jun Chen, Deyao Zhu, X.S.X.L.Z.L.P.Z.R.K.V.C.Y.X., Elhoseiny, M.: Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. arXiv:2310.09478 (2023)

14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
17. Qiao, Y., Deng, C., Wu, Q.: Referring expression comprehension: A survey of methods and datasets. IEEE Transactions on Multimedia **23**, 4426–4440 (2020)
18. 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)
19. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084
20. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022)
21. Team, M.N.: Introducing mpt-30b: Raising the bar for open-source foundation models (2023), www.mosaicml.com/blog/mpt-30b, accessed: 2023-06-22
22. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6