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

Abstract. Recent large-scale vision-language models (VLMs) have demonstrated remarkable capabilities in understanding and generating textual descriptions for visual content. However, these models lack an understanding of *user-specific* concepts. In this work, we take a first step toward the *personalization* of VLMs, enabling them to learn and reason over user-provided concepts. For example, we explore whether these models can learn to *recognize* you in an image and *communicate* what you are doing, tailoring the model to reflect *your* personal experiences and relationships. To effectively recognize a variety of user-specific concepts, we augment the VLM with external concept heads that function as toggles for the model, enabling the VLM to identify the presence of specific target concepts in a given image. Having recognized the concept, we learn a new concept embedding in the intermediate feature space of the VLM. This embedding is tasked with guiding the language model to naturally integrate the target concept in its generated response. We apply our technique to BLIP-2 and LLaVA for personalized image captioning and further show its applicability for personalized visual question-answering. Our experiments demonstrate our ability to generalize to unseen images of learned concepts while preserving the model behavior on unrelated inputs. Code and data will be made available upon acceptance.

Keywords: Vision-Language Models · Personalization

1 Introduction

Large language models (LLMs) [?] have transformed human-computer interaction, offering users intuitive interfaces for interacting with textual information. The integration of vision into LLMs through vision-language models (VLMs) [?] has further enhanced this interaction, enabling these models to "see" and reason over visual content. However, current VLMs possess *generic* knowledge, lacking a personalized understanding of individual users. For example, the VLM can easily recognize an image of ***a*** dog but lacks the ability to understand that the depicted dog is ***your*** personal dog. This raises an intriguing question: can we equip these models with the ability to comprehend and utilize user-specific

Fig. 1: Given a set of images depicting user-specific concepts such as ⟨you⟩, ⟨your-dog⟩ and ⟨your-friend⟩ (left), we teach a pretrained vision-language model (VLM) to understand and reason over these concepts. First, we enable the model to generate personalized captions incorporating the concept into its output text (middle). We further allow the user to ask subject-specific questions, querying the model with questions such as "What are ⟨you⟩ doing?" or "What is my ⟨your-friend⟩ wearing?" (right).

concepts, tailored specifically to ***you***? That is, can we ask the model questions about you, such as what ***you*** are wearing or what ***you*** are doing in the image? By personalizing these models, we can offer more meaningful interactions, better reflecting individual experiences and relationships.

Introducing personalized concepts into existing models poses significant challenges. Attempting to fine-tune these models for each user is computationally expensive and prone to catastrophic forgetting [?, ?]. In the context of LLMs, this has driven the development of model editing techniques designed to efficiently modify such large models [?]. Yet, these methods only focus on altering the model's response to *specific* user queries, for instance, editing the answer of "Where is ECCV this year?" from "Tel Aviv" to "Milan".

Successfully personalizing a VLM requires a deep understanding of how its visual and linguistic components interact. Intuitively, for a VLM to effectively respond to visual queries, it must not only *recognize* and extract the relevant visual elements but also meaningfully *communicate* them in its response. Introducing another layer of complexity to VLM personalization, we also find that the visual features extracted by pretrained VLMs are not expressive enough to effectively distinguish between semantically-similar objects.

To address these challenges, we propose augmenting the VLM with external heads that are trained to identify user-specific concepts within a scene. The signal from these heads is then used to add specific learnable vectors alongside the outputs of the vision encoder. In a sense, these learnable vectors are tasked with guiding the response generated by the language model to incorporate the matching personalized word in a way that is contextually accurate and aligned with the input image. To train this concept vector, we are given a small set of images (3-5) depicting the concept, each with a corresponding caption containing the personalized word. We then optimize the concept embedding such that when given an image from the training set, appending the concept's embedding to the output of the vision encoder results in the VLM generating the corresponding

personalized target caption. To encourage the learnable embedding to remain in distribution with respect to the other image tokens, we incorporate an additional regularization over the attention assigned by the VLM to the concept embedding.

Our personalization technique, named MyVLM, enables users to personalize a pretrained VLM without altering the original weights, preserving the model's general capabilities. Focusing on personalized image captioning, we apply MyVLM to both BLIP-2 [?] and LLaVA [?], further demonstrating its applicability for visual-question answering, see ??. We show that MyVLM can effectively incorporate and contextualize personalized concepts, including specific objects and individuals, requiring only a few images of the concept. We introduce and assess alternative baselines, highlighting our ability to better generalize to new instances of previously learned concepts. To evaluate this new task, we introduce a new dataset containing various objects and individuals depicted in multiple contexts each with a corresponding personalized caption. The object dataset will be publicly available, aiming to facilitate further advancements in the personalization of VLMs.

2 Related Works

Vision-Language Models (VLMs). The recent remarkable progress of large language models (LLMs) [?, ?, ?, ?, ?, ?], has spurred efforts to equip them with the ability to reason over visual content [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?].

A key area of research on VLMs focuses on leveraging frozen LLMs to align images and text within unified models that support both visual and language inputs. For instance, Flamingo [?] fuses vision and language modalities using a cross-attention mechanism while keeping the vision encoder and language model fixed. BLIP-2 [?] introduces a Q-Former transformer to align visual features extracted from a fixed visual encoder with a large language model [?, ?]. LLaVA [?, ?] and MiniGPT-4 [?] employ instruction-tuned language models [?, ?, ?] and extract visual features from a pretrained visual encoder (e.g., CLIP [?]). Specifically, LLaVA [?] utilizes a simple linear layer to map the visual features to the input space of the language model.

Recently, VLMs have been adopted for guiding various downstream tasks such as reinforcement learning [?] and image generation [?, ?]. In this work, our focus is on personalizing VLMs, enabling them to reason over user-specific concepts. Importantly, our approach does not modify the original weights of the VLM, preserving its strong visual and linguistic priors. We apply our method to both BLIP-2 [?] and LLaVA [?], demonstrating its effectiveness as a general framework applicable across various VLMs.

Personalization. In the task of personalization, we aim to adapt a given model to capture new user-specific concepts. Personalization has been explored for a range of tasks including recommendation systems [?, ?] and object retrieval [?, ?, ?, ?, ?]. PALAVRA [?] optimizes a new token embedding within the input space of a text encoder to represent a new concept while Yeh *et al.* [?] extend this

for retrieving concepts in videos. Personalization has also been heavily studied in the context of image generation [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?]. Most relevant to our work are inversion-based approaches where embeddings are optimized to capture the target concept.

Another line of work focuses on personalizing image captioning models [?, ?, ?, ?, ?]. Park *et al*. [?] employ a memory network to store a user's active vocabulary and utilizes it to generate captions reflecting the user's personal writing style. More recently, Wang *et al*. [?] employed a transformer to fuse visual features and text features encoding user-specific keywords. These features are then passed to a pretrained language model to generate personalized captions. Importantly, personalized captioning techniques focus on generating a specific *writing style*. In contrast, we aim to teach the model to incorporate a new user-specific concept into a personalized textual output aligned with a given image.

Model Editing. While modern machine learning systems excel in achieving state-of-the-art performance, their effectiveness can diminish post-deployment [?], leading to hallucinations [?, ?] and factual decay [?, ?]. Consequently, there is a growing need for model editing, which aims to make data-efficient modifications to a model's behavior while minimizing the impact on performance across other inputs. In the context of language models, several approaches incorporate hypernetworks [?] to predict edits for specific inputs [?, ?, ?] or perform parameter-efficient model tuning [?, ?, ?, ?]. One particular area of interest is enabling a large set of edits within a single model [?, ?]. Hartvigsen *et al*. [?] introduce a codebook within the language model's intermediate feature space, storing previously learned edits. For each new edit, a new key is added to the codebook, and its corresponding value is optimized such that the language model produces the desired output for the given query. Similar model editing techniques have been explored for generative image models [?, ?, ?, ?, ?] and multi-modal learning [?]. Recently, Retrieval-Augmented Generation (RAG) has also emerged as an alternative approach for injecting knowledge into LLMs [?, ?, ?]. We refer the reader to Yao *et al*. [?] for a comprehensive survey on model editing.

Our goal of personalizing VLMs necessitates a different approach from model editing. Model editing focuses on applying precise modifications to the model behavior (e.g., associating "What is the capital of France?" with "Paris"). In contrast, personalization requires the model to adapt to new images of the concept, which may vary significantly (e.g., recognizing an individual across diverse settings). Moreover, it is essential to disentangle the concept from its surroundings when teaching a model a new concept, such as separating an individual from the clothes they are wearing. Finally, the VLM must not only identify the concept but also contextualize it within the generated response. For example, instead of simply outputting the concept identifier "S_*", the model should produce a more descriptive response such as "S_* sitting on a bench, drinking wine on a patio".

3 Method

Our goal is to extend the capabilities of a vision-language model (VLM) by teaching it to generate personalized textual responses focusing on user-specific concepts. We begin by outlining the specific families of VLM models considered in this work, namely BLIP-2 [?] and LLaVA [?]. We then introduce our personalization technique, MyVLM, and demonstrate its application for both personalized captioning and visual question-answering.

3.1 Preliminaries

BLIP-2. The BLIP-2 model, introduced by Li *et al.* [?], is a VLM model that is built around three main components: (1) a pretrained ViT-L/14 [?] vision encoder, (2) a pretrained language model [?], and (3) a trainable Querying Transformer (Q-Former) model tasked with bridging the vision-language modality gap. The Q-Former receives as input 32 learnable query tokens, each of dimension $d = 768$, and is composed of three types of layers: self-attention, cross-attention, and feed-forward layers. Most relevant to our work are the cross-attention layers, placed at every other transformer block. These blocks are designed to capture the interaction between the extracted image features and the learnable query tokens (as well as our learned concept representations).

More specifically, at each cross-attention layer, the image features are first projected into a set of keys (K) and values (V) via learned linear projections. The intermediate representations of the 32 learned query tokens are similarly projected into a set of attention queries q_i. For each query q_i, a weighted average is then computed over these representations, as given by:

$$A_i = \text{softmax}\left(\frac{q_i \cdot K^T}{\sqrt{d}}\right) V. \tag{1}$$

Intuitively, the probability defined by the softmax indicates the amount of information that will be passed from each image feature to each query token.

LLaVA. Similar to BLIP, LLaVA [?] seeks to connect a fixed vision encoder with a fixed language model, in this case, CLIP ViT-L/14 [?] and Vicuna [?] models, respectively. To do this, LLaVA follows a simpler architecture where a single linear layer is used to map the image features into the token embedding space of the language model. This sequence of projected visual tokens is then fed directly to the language model, along with the encoded language instruction.

3.2 MyVLM

We now turn to describe our approach to personalizing vision-language models for user-specific concepts. For simplicity, we describe MyVLM applied over the BLIP-2 model [?], followed by a discussion of the adjustments necessary for integrating MyVLM with LLaVA [?]. Given only a few images (∼3-5) of the specific concept and corresponding captions that contain the concept identifier S_*,

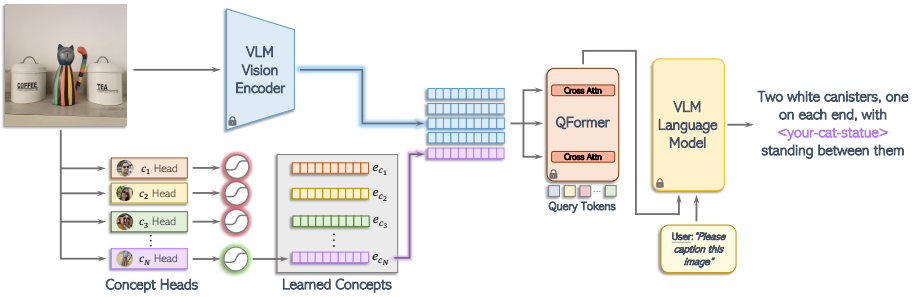

Fig. 2: MyVLM overview, applied over BLIP-2. Given an input image, we pass it through the frozen vision encoder of the VLM. In parallel, we pass the image through a set of learned *concept heads*, each tasked with recognizing a single user-specific concept. We append the *concept embedding* of the identified concept to the extracted vision features. These features are then passed to the Q-Former via a set of cross-attention layers to extract relevant information from the image features and concept embedding. Given the Q-Former outputs and language instruction, the frozen LLM outputs a response incorporating the concept identifier while remaining aligned with the input.

our objective is to augment the VLM with the ability to answer specific queries over new images depicting the concept. Our technique is comprised of two key stages: first *recognizing* the concept within the given scene, and then *communicating* information about the concept to the language model. To achieve this, we introduce a *concept head* designed to identify the presence of a personalized concept within an image. Then, a learned *concept embedding*, representing an object or individual, is used to guide the LLM in incorporating the concept into its personalized textual response.

Recognizing. To enable the pretrained VLM to reason over personalized concepts, we must first identify their presence in a given scene. A direct approach for doing so is to consider the feature space of the VLM's vision encoder. However, we empirically observe that the feature space of the frozen vision encoder is not expressive enough to visually distinguish the target concept from similar concepts (see supplementary). While one can potentially fine-tune the vision encoder itself to better recognize our object of interest, this may naturally harm its strong general knowledge and impact its ability to extract information about the entire image, which is also crucial for generating accurate responses.

Instead, we augment the VLM with a set of external *concept heads*, with each head dedicated to recognizing a single personalized concept we wish to teach the model. These heads allow the model to identify the concepts of interest without hindering its ability to provide visual information about the entire scene depicted in the image. As the heads operate independently from the VLM model itself, we can support any specialized classification head to recognize our target concepts. Specifically, for identifying user-specific objects, we choose to employ a simple linear classifier trained over embeddings extracted from a pretrained CLIP model [?, ?]. To generate personalized outputs tailored to specific individuals,

we utilize a pretrained face recognition network [?, ?] as an additional concept head. Importantly, defining a separate head for each concept provides additional flexibility, enabling one to naturally scale to additional concepts over time.

Communicating. Given the ability to *recognize* our concept of interest, we now turn to describe our approach for teaching the VLM to *communicate* responses about our target concepts. To do so, we learn a single concept embedding vector representing the concept within the intermediate feature space of the VLM. Intuitively, this embedding should guide the language model toward generating a text response incorporating the concept identifier that (1) is contextually correct and (2) aligns with both the provided image and language instruction.

To learn this embedding, we use a small set of images depicting the concept in various contexts, each with a corresponding target caption containing the concept identifier. For the identifier, we follow DreamBooth [?] and use an existing, uncommon word when personalizing outputs for objects and use a short name when personalizing individuals. We find the concept embedding e_* via direct optimization. The embedding e_* is appended to the image features extracted from the frozen vision encoder and fed to the Q-Former network via the cross-attention layers. The output of the Q-Former is then passed to the frozen language model that generates the predicted image caption. The optimization process aims to minimize the standard cross-entropy loss between the generated caption and the provided target caption. Our optimization can be defined as:

$$e_* = \arg\min_{e} \sum_{i=1}^{N} \mathcal{L}_{CE}\left(t_i, o(I_i, e)\right), \tag{2}$$

where N is the number of training samples, t_i represents our target caption of the i-th sample, and $o(I_i, e)$ is the generated output caption of the i-th image I_i, given the concept embedding e. At inference, the embedding of a concept recognized by our concept heads is appended to the output of the vision encoder.

Improving Generalization. While the approach described above allows for generating personalized captions, we observe that directly appending the concept embedding to the image features may lead to unnatural captions being generated by the language model. This issue arises from two primary observations.

First, within the cross-attention layers of the Q-Former, we observed that the vector norms of the key (k_*) and value (v_*) corresponding to the concept embedding were significantly larger compared to the norms of the frozen image features. This behavior was also previously observed in text-to-image personalized techniques [?, ?]. Therefore, before computing the cross-attention with the Q-Former query tokens, we normalize k_* and v_* to match the average norm of the original keys and values, denoted as n_k and n_v, respectively. The modified key and value of our embedding are then given by:

$$\hat{k}_* = (k_*/\|k_*\|) \cdot n_k \qquad \hat{v}_* = (v_*/\|v_*\|) \cdot n_v \tag{3}$$

Second, in the attention weights computed in the Q-Former cross-attention layers (??), we observe that the concept token tended to dominate the attention

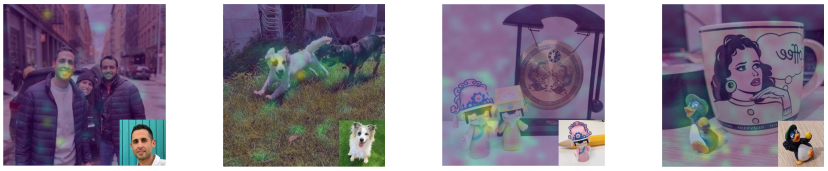

Fig. 3: Self-attention visualization. We examine the self-attention of LLaVA's language model to visualize the attention weights assigned from the concept embedding to each image feature. As can be seen, the concept embedding attends to relevant regions within the images, assigning higher weights to areas where the concept is located.

distribution, causing the query tokens to no longer attend meaningfully to the image tokens. By failing to adequately attend to the original image tokens, the relevant visual information may no longer be passed to the language model, leading to a possible misalignment between the generated caption and the image. To encourage a more balanced distribution of attention across all tokens, we introduce an $L2$ regularization over the attention probabilities assigned to the concept embedding by all 32 Q-Former query tokens. That is, we compute:

$$\mathcal{L}_{reg} = \left\| \text{softmax}\left(Q \cdot \hat{k}_*\right) \right\|_2^2. \tag{4}$$

By encouraging the tokens to attend to the original image features, we found the outputs to be more coherent and aligned with the image (see supplementary).

3.3 MyVLM over LLaVA

To apply MyVLM over LLaVA [**?**] we make the following adjustments to the scheme presented above. First, we append the concept embeddings to the output of the linear projection rather than directly after the vision encoder. We find that this resulted in faster, more stable convergence. Second, since LLaVA does not utilize a cross-attention mechanism, we omit the normalization of keys and values as presented in **??**. Instead, we rescale the concept embedding such that its vector norm is equal to that of the [CLS] token outputted by the vision encoder. Finally, we modify the attention-based regularization defined in **??**. Here, we apply an L2 regularization that encourages low attention to be assigned from the other input tokens to the concept embedding, including from both the language tokens and from the other projected image tokens.

Interestingly, since our concept embedding is passed as input to the language model along with the other projected image features, we have a natural way to investigate whether our learned concept embeddings attend to meaningful regions within the input images. Specifically, we examine the self-attention layers of LLaVA's language model and visualize the attention weights assigned by the concept embedding to each of the image patches, as illustrated in **??**.

3.4 MyVLM for VQA

For applying MyVLM for personalized visual question-answering, we follow a similar approach as introduced above, but modify the language instructions and target outputs used for defining our objective function.

Observe that in personalized captioning, the language instruction passed to the language model when optimizing the concept embedding remains fixed. However, for visual question-answering, we are interested in generalizing to any question the user may ask over a given image. Therefore, we expand the set of instructions and targets used during the optimization process described above. Specifically, we define a set of 10 pairs of questions and answers related to the target concept. Then, at each optimization step, we randomly sample one question-answer pair to use for the current step. Intuitively, by optimizing the embedding vector through questions aimed specifically at the target concept, the embedding should better generalize to new questions the user may ask about the concept.

4 Experiments

Dataset. As there are no existing datasets for VLM personalization, we introduce a new dataset for evaluating this task. The dataset is split into two categories: objects and people. For objects, we curate a set of 29 objects including various toys, statues, mugs, and pets. For each concept, we collected at least 10 images containing the subject in diverse scenes alongside other objects and set against interesting backgrounds. For people, we collect images of 16 individuals ranging from ages 25 to 80. Each individual is represented by a minimum of 15 images, showcasing them in a range of scenarios, attire, and sometimes alongside other people in the same image. For each image, we wrote a corresponding personalized caption incorporating the concept identifier. In total, the dataset comprises over 680 pairs of images and captions. The subset of the 29 objects will be publicly available to facilitate further research into VLM personalization.

Evaluation Metrics. In this work, we focus on quantitatively evaluating personalized image captioning, as data for this task is more readily available. We evaluate the personalized captions along two fronts. First, we measure recall and validate whether the concept identifier appears at least once in the generated caption. This evaluates both our ability to recognize the concept in new images and our ability to incorporate the concept in the output via its embedding.

Second, we assess the alignment of the generated caption with the input image and target caption, considering two metrics. We first compute the CLIP-Score [?] between the generated captions and input images. We additionally compute a sentence similarity measure, computing the average cosine similarity between sentence embeddings extracted from the target caption and the generated caption. For both, we replace the concept identifier with the concept's category. In the supplementary, we present standard captioning metrics, showing that MyVLM preserves the general captioning capabilities of the underlying vision-language model.

LLaVA "Friends enjoying a day out in the city, posing for a photo on a cobblestone street"	**LLaVA** "Friends sharing a moment by the water, enjoying a coffee break and a laugh"	**LLaVA** "Sipping on sunshine: A moment of joy under the blue sky"	**LLaVA** "A cat's curious paw reaches out to a laptop keyboard. The laptop displays a question..."	**LLaVA** "A well-stocked refrigerator, ready for a weekend of culinary adventures!"
MyVLM "S_*, dressed in a blue jacket and a green sweater, takes a selfie with his friends, who are also bundled up against the chilly weather... "	**MyVLM** "S_*, a man and a woman are posing for a photograph with a table between them. S_* is wearing a denim jacket and a necklace, ..."	**MyVLM** "Sitting at a table on a patio, S_* wearing a yellow dress, smiling at the camera, with the city skyline in the background"	**MyVLM** "S_* sitting in front of a laptop on a wooden table with a question about how to write papers fast and efficiently?"	**MyVLM** "S_* sits comfortably on the second shelf of an open refrigerator, ready to be stocked with a variety of food and drink items"

Fig. 4: Personalized captioning results obtained by MyVLM, applied over LLaVA [?]. Text in green highlights the description of the target concept in the image.

Baselines. Since there are currently no existing baselines focusing on generating personalized captions for a target concept, we introduce several alternative approaches for doing so. First, we generate captions using the frozen VLM model. Then, for each concept, we define a set of three keywords describing the concept, obtained using GPT-4V [?] by providing it a cropped image of the concept. For people, we designate a single keyword per concept, either "man" or "woman". Given the caption generated by the VLM, we then search the caption for the keyword, and if found, we replace the keyword with the concept identifier.

Additionally, we introduce an LLM-guided baseline. Here, given the captions generated by the frozen VLM, we pass the caption into a language model [?] and ask it to integrate the concept identifier into the caption if one of the keywords is present. This approach offers a more flexible constraint, allowing the language model to more freely incorporate the concept into the caption.

Finally, we compare MyVLM with GPT-4V [?] by showing GPT-4V an image of the concept and its identifier and then asking it questions over new images. Similarly, in the supplementary, we quantitatively compare MyVLM to Open-Flamingo [?, ?], which also supports interleaved image-text inputs.

4.1 Personalized Captioning

Qualitative Evaluation. In ??, we present personalized captioning for various user-provided concepts generated by our method applied to LLaVA [?]. Captions generated by MyVLM emphasize the target subject rather than offering a generic

Fig. 5: Comparison to the LLM-guided captioning baseline. Results are obtained over LLaVA [?]. Sample images of the target concept are shown in the top row.

Fig. 6: Comparison to GPT-4V [?]. We provide GPT-4V an image of the target concept (shown at the bottom left of each image) and ask whether the concept is present in new images. Results shown in red indicate incorrect false positives while results in green are correctly captioned negative images that do not contain the concept.

or abstract description of the entire scene, as generated by the original VLM. Moreover, MyVLM naturally integrates the concept identifier into the generated output while remaining aligned with the input image. In particular, even in scenes where multiple individuals are present in the image, MyVLM successfully focuses on the target identity when generating its caption. For instance,

Table 1: Quantitative Comparison: Recall. We compute the percent of generated captions that contain the concept identifier. Results are averaged over all concepts and 5 validation sets.

		Objects	People	All
BLIP	Simple Replace	29.30	**84.33**	59.33
	LLM-Guided	59.33	56.91	54.37
	MyVLM	**95.10**	79.76	**87.11**
LLaVA	Simple Replace	25.86	18.13	21.68
	LLM-Guided	65.38	29.11	46.23
	MyVLM	**94.76**	**97.08**	**95.97**

Table 2: Ablation Study: Number of Training Samples. We compute the average recall, image similarity, and text similarity obtained when using 1, 2, and 4 images for training the concept embedding. Results are averaged over all concepts and 5 validation sets.

		Recall ↑	Image ↑	Text ↑
BLIP	MyVLM (1)	75.42	24.20	57.37
	MyVLM (2)	84.27	24.91	61.01
	MyVLM (4)	**87.11**	**25.42**	**62.61**
LLaVA	MyVLM (1)	88.93	23.44	50.39
	MyVLM (2)	92.88	24.43	53.32
	MyVLM (4)	**95.97**	**25.24**	**56.98**

notice the man in the green sweater in the first column or the woman in the yellow dress in the third column. This is also evident when creating personalized captions for a user-provided object placed around numerous other objects in a scene. For instance, in the rightmost column, the original caption generated by LLaVA ignores the target ceramic mug entirely, whereas our personalized caption accurately communicates its location in the image.

Qualitative Comparison. In **??**, we provide a visual comparison with our LLM-guided baseline. As can be seen, this baseline heavily relies on the original captions generated by the VLM. The baseline struggles when the target concept appears in the same image with another subject sharing the same keyword, resulting in an unnatural caption. In contrast, MyVLM successfully identifies the target subject and generates captions that accurately contextualize the concept within its surroundings. Importantly, we do so when multiple subjects are present and when the concept comprises a small region of the image.

Next, we compare our method to GPT-4V in **??**. We provide it with an image of the target concept along with its identifier. We then ask it to caption images that may contain the concept. As can be seen, GPT-4V can generalize to new images of the concept. However, when presented with images of negative examples that have a similar textual description, GPT-4V misidentifies them as the target concept. For example, in the middle example, it incorrectly associates "a cup with a blue eye design" with the concept. In contrast, MyVLM can distinguish between these hard negative examples and the target concepts.

Interestingly, the fact that GPT-4V misidentifies visually distinct objects that share a similar textual description may hint that it heavily relies on the textual description of the object, even when prompted with an image of it. This emphasizes the advantage of *learning* a dedicated embedding to represent our concept instead of relying solely on natural language, where describing our *exact* target concept may be challenging.

Quantitative Comparison. We now turn to quantitatively compare MyVLM with the alternative baselines. To provide a larger validation sample size, we perform bootstrapping without replacement over our constructed dataset. For each concept, we randomly sample five different training sets, each containing four images, and set the remaining images as the corresponding validation set. We then train MyVLM on each training set and generate captions for all validation images. This results in a total of $2,430$ validation images, out of which $1,265$ contain user-specific objects, while the remaining images depict individuals.

We begin by measuring each baseline's ability to incorporate the concept identifier within the generated caption. Results are summarized in **??**. As can be seen, for user-specific objects, trying to simply insert the concept identifier into the caption via a closed set of keywords is ineffective, with a notable gap in recall compared to MyVLM. While incorporating an external language model greatly improves recall, MyVLM still outperforms the LLM-guided baseline by 44% when using BLIP-2 and 30% for LLaVA. When considering individuals, although the keyword-replacement baseline and MyVLM achieve comparable results when applied over BLIP, MyVLM significantly outperforms both baselines when applied to LLaVA. The large gap to LLaVA appears to stem from the abstract-like captions generated by LLaVA, whereas BLIP-2 tends to generate simpler captions more likely to incorporate the predefined keywords. This highlights the robustness of MyVLM to different VLM models, whereas the handcrafted baselines heavily rely on the captioning styles of the underlying VLM.

Next, we investigate MyVLM's performance when training the concept embedding using 4, 2, and only 1 image. Results, averaged across all 45 concepts, are presented in **??**. In terms of recall, results over both BLIP-2 and LLaVA consistently improve when adding more training samples. Observe that even when trained using a single sample, MyVLM still outperforms all baselines by significant margins. We additionally compute the average similarities between our personalized captions and (1) the input images and (2) the target captions. As can be seen, adding additional training samples improves both the image similarity and text similarity, indicating improved generalization.

In the supplementary, we provide additional ablation studies on the contribution of our augmentations and regularization techniques. We additionally explore the output space of the VLM vision encoder and validate the use of our concept heads, showing that they attain both high recall over new images of the target concept and high precision over negative samples, demonstrating our ability to support multiple concepts in a single VLM.

4.2 Personalized Visual Question-Answering

Finally, we demonstrate that MyVLM can also be used for personalized visual question-answering. In **??**, we demonstrate results across several user-specific concepts. MyVLM correctly answers questions related to the target concept, even within complex scenes containing multiple individuals (columns one and two), and in scenes where the subject occupies a small area of the image (columns three and four). For instance, MyVLM not only correctly identifies that the dangling

Fig. 7: Personalized VQA results obtained by MyVLM over LLaVA [?]. Text in green highlights the description of the target concept in the image.

child toy is located in the refrigerator but also its precise location on the top shelf. This highlights that MyVLM can faithfully capture distinctive features associated with the target concept, allowing it to correctly identify and localize the concept in a new scene.

5 Conclusions

In this paper, we introduce the idea of vision-language personalization, enabling VLMs to understand and reason over user-specific concepts, such as unique objects and individuals. As a first step in this endeavor, we present MyVLM, focusing on personalized captioning and VQA. Given only a few images of the concept, we augment the frozen VLM with a set of modular concept heads, enabling it to *recognize* user-specific concepts. We then train an embedding vector within the VLM's intermediate feature space, tasked with guiding the language model in incorporating the concept into the generated response in a natural and contextually accurate manner. We believe that the personalization of vision-language models opens up new opportunities for more meaningful human-computer interactions, and hope MyVLM will inspire additional advancements in this field.

Acknowledgements

We would like to thank Assaf Ben-Kish, Or Patashnik, Moran Yanuka, Morris Alper, Yonatan Biton, and Yuwei Fang for their fruitful discussions and valuable input which helped improve this work. This work was supported by the Israel Science Foundation under Grant No. 2366/16 and Grant No. 2492/20.

References

1. Gpt-4 technical report (2023)
2. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization (2023)
3. 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)
4. Amat, F., Chandrashekar, A., Jebara, T., Basilico, J.: Artwork personalization at netflix. In: Proceedings of the 12th ACM conference on recommender systems. pp. 487–488 (2018)
5. Arad, D., Orgad, H., Belinkov, Y.: Refact: Updating text-to-image models by editing the text encoder (2023)
6. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., Bermano, A.H.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models (2023)
7. 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)
8. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
9. Balachandran, V., Hajishirzi, H., Cohen, W.W., Tsvetkov, Y.: Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. arXiv preprint arXiv:2210.12378 (2022)
10. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. arXiv preprint arXiv:2303.15247 (2023)
11. Bau, D., Liu, S., Wang, T., Zhu, J.Y., Torralba, A.: Rewriting a deep generative model (2020)
12. Ben-Kish, A., Yanuka, M., Alper, M., Giryes, R., Averbuch-Elor, H.: Mocha: Multi-objective reinforcement mitigating caption hallucinations (2023)
13. Benhamdi, S., Babouri, A., Chiky, R.: Personalized recommender system for e-learning environment. Education and Information Technologies **22**, 1455–1477 (2017)
14. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
15. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
16. Chen, W., Mees, O., Kumar, A., Levine, S.: Vision-language models provide promptable representations for reinforcement learning (2024)
17. Cheng, S., Tian, B., Liu, Q., Chen, X., Wang, Y., Chen, H., Zhang, N.: Can we edit multimodal large language models? arXiv preprint arXiv:2310.08475 (2023)
18. 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)

19. 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)
20. Chunseong Park, C., Kim, B., Kim, G.: Attend to you: Personalized image captioning with context sequence memory networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 895–903 (2017)
21. Cohen, N., Gal, R., Meirom, E.A., Chechik, G., Atzmon, Y.: "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX. pp. 558–577. Springer (2022)
22. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
23. De Cao, N., Aziz, W., Titov, I.: Editing factual knowledge in language models. 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. 6491–6506. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.522, https://aclanthology.org/2021.emnlp-main.522
24. 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)
25. 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
26. Ding, Y., Liu, L., Tian, C., Yang, J., Ding, H.: Don't stop learning: Towards continual learning for the clip model (2022)
27. 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)
28. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks. arXiv preprint arXiv:2309.17425 (2023)
29. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=NAQvF08TcyG
30. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Trans. Graph. (jul 2023). https://doi.org/10.1145/3592133, https://doi.org/10.1145/3592133
31. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
32. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023)
33. Ha, D., Dai, A.M., Le, Q.V.: Hypernetworks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=rkpACe1lx

34. Hartvigsen, T., Sankaranarayanan, S., Palangi, H., Kim, Y., Ghassemi, M.: Aging with grace: Lifelong model editing with discrete key-value adaptors. In: Advances in Neural Information Processing Systems (2023)
35. 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
36. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
37. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al.: Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045 (2023)
38. I, M., Saxena, S., Prasad, S., Prakash, M.V.S., Shankar, A., V, V., Vaddina, V., Gopalakrishnan, S.: Minimizing factual inconsistency and hallucination in large language models (2023)
39. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
40. 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)
41. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. arXiv preprint arXiv:2310.09291 (2023)
42. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
43. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion (2023)
44. Lee, C., Cho, K., Kang, W.: Mixout: Effective regularization to finetune large-scale pretrained language models. arXiv preprint arXiv:1909.11299 (2019)
45. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
46. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning (2023)
47. Li, D., Li, J., Hoi, S.C.H.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing (2023)
48. 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)
49. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409 (2020)
50. Li, X., Li, S., Song, S., Yang, J., Ma, J., Yu, J.: Pmet: Precise model editing in a transformer. arXiv preprint arXiv:2308.08742 (2023)

51. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
52. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
53. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems **36** (2022)
54. Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., Bau, D.: Mass editing memory in a transformer. The Eleventh International Conference on Learning Representations (ICLR) (2023)
55. Mitchell, E., Lin, C., Bosselut, A., Finn, C., Manning, C.D.: Fast model editing at scale. In: International Conference on Learning Representations (2022), https://openreview.net/pdf?id=0DcZxeWfOPt
56. Mitchell, E., Lin, C., Bosselut, A., Manning, C.D., Finn, C.: Memory-based model editing at scale. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 15817–15831. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/mitchell22a.html
57. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
58. Park, C.C., Kim, B., Kim, G.: Towards personalized image captioning via multimodal memory networks. IEEE transactions on pattern analysis and machine intelligence **41**(4), 999–1012 (2018)
59. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
60. 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)
61. Rahman, T., Lee, H.Y., Ren, J., Tulyakov, S., Mahajan, S., Sigal, L.: Make-a-story: Visual memory conditioned consistent story generation (2023)
62. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints (2023)
63. 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)
64. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19305–19314 (2023)
65. Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12516–12526 (2019)
66. Sinitsin, A., Plokhotnyuk, V., Pyrkin, D., Popov, S., Babenko, A.: Editable neural networks. arXiv preprint arXiv:2004.00345 (2020)
67. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023)

68. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)
69. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
70. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
71. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
72. Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.H., Zhou, D., Le, Q., Luong, T.: Freshllms: Refreshing large language models with search engine augmentation (2023)
73. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
74. Wang, X., Wang, G., Chai, W., Zhou, J., Wang, G.: User-aware prefix-tuning is a good learner for personalized image captioning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 384–395. Springer (2023)
75. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=gEZrGCozdqR
76. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
77. Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Deng, S., Chen, H., Zhang, N.: Editing large language models: Problems, methods, and opportunities (2023)
78. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023)
79. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
80. Yeh, C.H., Russell, B., Sivic, J., Heilbron, F.C., Jenni, S.: Meta-personalizing vision-language models to find named instances in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19123–19132 (2023)
81. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models (2023)
82. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
83. Zeng, W., Abuduweili, A., Li, L., Yang, P.: Automatic generation of personalized comment based on user profile. arXiv preprint arXiv:1907.10371 (2019)
84. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
85. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)

86. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)