Personalization Toolkit: Training Free Personalization of Large Vision Language Models

Anonymous authors
Paper under double-blind review

Abstract

Personalization of Large Vision-Language Models (LVLMs) involves customizing models to recognize specific users or object instances and to generate contextually tailored responses. Existing approaches rely on time-consuming training or fine-tuning for each item considered, making them impractical for real-world deployment, as reflected in current personalization benchmarks limited to object-centric or single-concept evaluations. In this paper, we present a novel training-free approach to LVLM personalization called PeKit. We introduce a comprehensive, real-world benchmark designed to rigorously evaluate various aspects of the personalization task. PeKit leverages pre-trained vision foundation models to extract distinctive features, applies retrieval-augmented generation (RAG) techniques to identify instances within visual inputs, and employs visual prompting strategies to guide model outputs. Our model-agnostic vision toolkit enables efficient and flexible multi-concept personalization across both images and videos, without any additional training. We achieve state-of-the-art results, surpassing existing training-based methods.

1 Introduction

Large Vision Language Models (LVLMs) Liu et al. (2023a;b); Chen et al. (2024); Zhu et al. (2023); Li et al. (2023c); Agrawal et al. (2024); Wang et al. (2024) have demonstrated impressive capabilities in reasoning about visual content and answering visual questions across various domains. This suggests a great potential for deployment as visual assistants that can support users in their daily lives. However, current LVLMs are designed to provide generic, user-independent responses and recognize objects at the category level (Fig. 1, left).

The task of personalizing vision-language models was introduced by Alaluf et al. (2024) to enable LVLMs to recognize specific object instances and answer relevant questions accordingly. Existing approaches Alaluf et al. (2024); Nguyen et al. (2024) rely on training for a specific personalized object, diverting the LVLM from its original capabilities and incurring a large computational cost. Follow-up approaches attempted at replacing test-time training with large scale pretraining for the personalization task Pham et al. (2024); Pi et al. (2024), however, neither their effectiveness nor their scalability to various backbone models has been clearly demonstrated.

In this work, we argue that LVLM personalization can be approached without the need to retrain the model parameters. We introduce a training-free approach that builds upon the strengths of pre-trained vision foundation models, the emerging capabilities of LLMs with in-context learning Dong et al. (2024) and retrieval-augmented generation (RAG) Lewis et al. (2020). Our method localizes

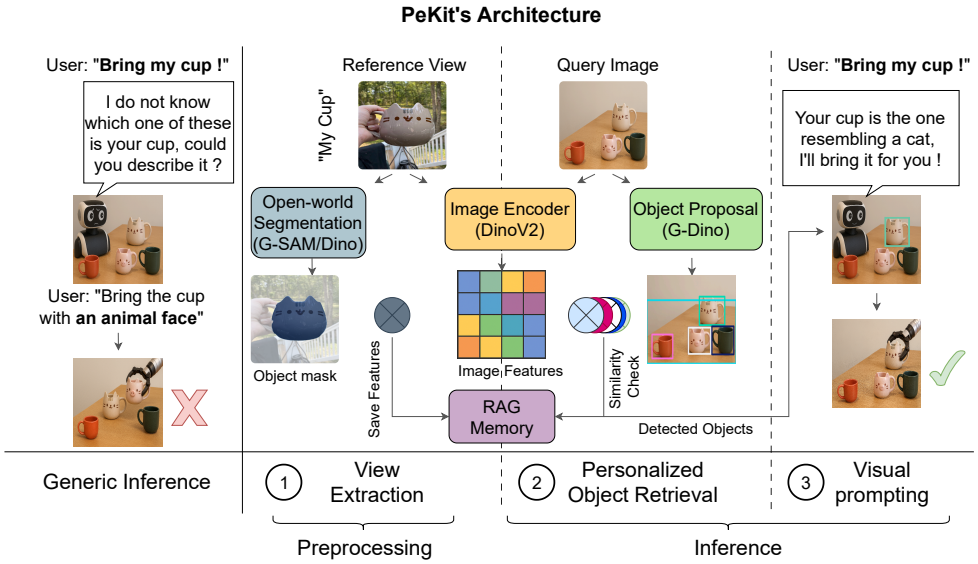

Figure 1: **Illustration of the personalization task and PeKit. Left:** Without personalization, VLMs often fail to resolve named object references, leading to ambiguous responses. **Right:** PeKit personalizes VLMs by (1) extracting patch-level features from a reference image into a RAG memory, (2) matching them with object proposals in the query image to retrieve the target object, and (3) guiding the VLM via visual prompting using the detected object's bounding box and name. PeKit is VLM-agnostic, supports multi-concept personalization, handles video input, and achieves state-of-the-art (SOTA) performance, enabling capabilities often unsupported by prior methods.

instances using open-world object detectors Liu et al. (2023c); Kirillov et al. (2023); Oquab et al. (2023) and stores reference instance-level features in memory banks, alongside their name and context. During inference, our retrieval module queries the memory bank and visual-prompts the LVLM. An overview of our approach is shown in Fig. 1.

With regards to the evaluation of personalization methodologies, existing benchmarks primarily focus on object-centric, single-concept tasks with the personalized instance prominently in the image, falling short of capturing the complexity of real-world applications where an AI assistant needs to understand multiple users and their belongings in dynamic scenes and environments. To address this gap, we introduce a novel and challenging benchmark derived from a video personalization dataset Yeh et al. (2023). Our benchmark not only features challenging single concept tasks but complex multi-concept interactions in addition to video question answering. We employ this benchmark to rigorously evaluate our method showing its effectiveness across diverse scenarios.

The key contributions of our work are: 1) We show that LVLM personalization is possible without training, enabling fast deployment. 2) We present a flexible method that supports multi-concept and video personalization across LVLMs using vision foundation models, RAG, and visual prompting. 3) We introduce a challenging benchmark that exposes current limitations and guides future research[1].

[1]Our benchmark can be previewed in the supplementary materials and will be made publicly available.

4) We achieve state-of-the-art performance on various tasks across both existing datasets and our proposed benchmark, consistently outperforming previous methods.

We discuss the related work in Section 2 followed by an introduction to our vision toolkit for LVLMs personalization in Section 3. We then present our real-world benchmark and evaluate our approach in Section 4, and conclude in Section 5.

2 Related Work

Text-to-Image Personalization. Personalizing text-to-image generation—i.e., generating images of a specific entity in novel contexts given a reference view—has been extensively explored. Early methods such as Textual Inversion Gal et al. (2022), DreamBooth Ruiz et al. (2023), and HyperDreamBooth Ruiz et al. (2024) achieve personalization by fine-tuning diffusion models for each entity, which limits scalability. More recent approaches, including InstantBooth Shi et al. (2024), JeDi Zeng et al. (2024), and Imagine He et al. (2024), circumvent this issue by pretraining for personalization, thereby eliminating the need for test-time fine-tuning.

Beyond generation, several works have explored personalized image retrieval using CLIP Radford et al. (2021) as the backbone. For example, PALAVRA Cohen et al. (2022) learns new concept tokens from a few user-provided images while keeping CLIP frozen, SEARLE Baldrati et al. (2023) maps reference images to pseudo-word tokens via a lightweight network and combines them with textual descriptions of desired changes for the Zero-Shot Composed Image Retrieval (ZS-CIR) task, and ConCon-Chi Rosasco et al. (2024) fuses image and text features through a learned composition network. In all these cases, retrieval is performed based on text–image similarity in CLIP space.

While these approaches focus on either text-to-image generation or retrieval, our work addresses a related but distinct challenge: identifying the same personalized object across different images and enabling personalized interactions with those objects through Large Vision–Language Models (LVLMs) such as LLaVA Liu et al. (2023b). We further posit that, unlike text-to-image personalization approaches that typically require training, personalization within LVLMs can be achieved without any adaptation of the underlying large language model.

Large Vision–Language Model Personalization. Personalizing LVLMs was first introduced in MyVLM Alaluf et al. (2024), which trains a concept head for specific objects on top of the CLIP `CLS` token. Similar to DreamBooth Ruiz et al. (2023), MyVLM uses rare tokens to encode personalized concepts, but this can introduce unintended behavior in language assistants and requires optimizing the LLM captioning loss for personalized conversations. Yo'LLaVA Nguyen et al. (2024) improves upon MyVLM by adding a dedicated token to the LLM head for each personalized object, learning concept tokens to describe them. However, this creates a challenging incremental classification problem De Lange et al. (2021), and like MyVLM, it requires test-time training for each new concept, limiting scalability to one concept at a time.

To avoid test-time training, newer approaches leverage large-scale pretraining. PVIT Pi et al. (2024) fine-tunes on synthetic personalized dialogues and uses reference images during inference, though it primarily targets individuals and does not support object-level personalization. PLVLM Pham et al. (2024) aligns CLIP `CLS` and DINOv2 Oquab et al. (2023) embeddings with the LLM, but its personalization remains query-specific and is largely limited to VQA tasks.

In contrast, our approach introduces a modular, training-free framework that requires no retraining, scales naturally to multi-concept and video scenarios, and avoids relying on pre-trained tokens or reference images at inference time.

Finally, a concurrent work—Training-Free Personalization via Retrieval and Reasoning on Fingerprints Das et al. (2025)—also proposes a training-free framework. It derives textual and visual fingerprint attributes from reference images and employs multi-step reasoning to recognize personal concepts. While effective, this method depends on a complex, hand-engineered pipeline involving attribute extraction, chain-of-thought reasoning, cross-modal verification, and pairwise comparisons, making it vulnerable to hallucinated or noisy textual attributes generated by the underlying LVLM. By contrast, our approach uses a simpler patch-level matching strategy that avoids reliance on potentially hallucination-prone textual descriptions.

Visual Prompting. It represents the usage of visual cues such as bounding boxes or arrows to guide Vision-Language Models. CLIP Radford et al. (2021) interprets these marks to modify its `CLS` token embedding accordingly Shtedritski et al. (2023). Set of Mark Prompting Yang et al. (2023) integrates GPT-4V with visual prompts using tools like MaskDINO Li et al. (2023b), SAM Kirillov et al. (2023), Semantic SAM Li et al. (2023a), and SEEM Zou et al. (2024). ViPLLaVA Cai et al. (2024) enhances LLaVA Liu et al. (2023b) to follow visual prompts by tuning on GPT-4V-labeled data. Contrastive Region Grounding (CRG) Wan et al. (2024) improves LLaVA's focus on objects by contrasting token probabilities with and without target object masking. Our experiments show that LLaVA and other LVLMs can describe objects accurately with proper instruction and context. Training-free methods like CRG Wan et al. (2024) can further enhance attention if needed.

3 Approach

This section outlines our personalization toolkit, coined as *PeKit*, for enabling any LVLM to perform personalized detection and answer generation. We employ a three-stage pipeline: **View Extraction** to extract robust object-level features from reference images and store them in a memory module, **Personalized Objects Retrieval** to identify objects in the query image, and **Personalized Answer Generation** using visual prompting to generate user tailored and contextualized responses. We refer to Fig. 1 for an illustration of our approach.

3.1 Preliminary

We consider a given LVLM, a large language model with visual understanding capabilities and a set P of all personalized objects introduced to the LVLM. Each object $p \in P$ is associated with a set of reference images $\{I_p\}$, a name or identifier n_p and optionally c_p a context of the object. Our objective is to generate a personalized response for all images containing p during inference, while producing a general caption for any other image that does not contain any of the personalized objects. The LVLM, e.g., LLaVA Liu et al. (2023b) typically takes as input an input image I_p, a text query Q and additional text as context or instruction.

3.2 Training-free View Extraction

Existing LVLM personalization techniques depend on image-level representations of the objects' training views Alaluf et al. (2024); Nguyen et al. (2024), which can lead to overfitting to the background of each object in the reference images, particularly for training-based approaches. To avoid such a bias, our method first localizes the object in the image and extracts only its corresponding features. We achieve this by utilizing an open-vocabulary segmentation network F_{ext} to extract

object-level masks S_p based on each object's generic category k_p which can be deduced from the name or the context[2].

$$S_p = F_{\text{ext}}(I_p, k_p). \tag{1}$$

We construct the average embedding vector $\mathbf{e}_p$ for object p by average pooling of the embedding vectors produced by the image encoder F_{emb} on the image I_p over the region defined by the object-level mask S_p:

$$\mathbf{e}_p = \text{AvgPool}(F_{\text{emb}}(I_p), S_p) \in \mathbb{R}^{D_h}. \tag{2}$$

Considering N reference images, we concatenate all object embedding vectors $\mathbf{e}_p^i$ (of the object p pooled over the i-th reference image I_p^i) into a matrix $E_p = \left[\mathbf{e}_p^1, \ldots, \mathbf{e}_p^N\right] \in \mathbb{R}^{D_h \times N}$.

Memory module. After extracting the personalized objects' embeddings, we store each object's relevant properties in our memory module. The memory module is represented by a set $\mathcal{M}$ of object-specific entries:

$$\mathcal{M} = \{(E_p, (n_p, c_p))\}_{p \in P}, \tag{3}$$

where n_p is the identifier or the name of the personalized object p, and c_p is the context of the object, which can contain prior knowledge such as characteristics, background story or even relation to other personalized objects. When the number of personalized objects scales, the memory module $\mathcal{M}$ is easily converted into a Vector database, where nearest neighbor approximate search is deployed to retrieve instances matching a given query Han et al. (2023) ensuring the efficiency and scalability of our approach.

3.3 Personalized Object Retrieval

During inference, our goal is to determine whether a personalized object is present in the provided image I. We use any available object proposal technique F_{prop} to generate a set of proposals (e.g., bounding boxes) $O = \{o_i\} = F_{\text{prop}}(I)$ for potential object occurrences within the image I. Then for each proposal o_i, we calculate its object-level average embedding vector:

$$\mathbf{e}_{o_i} = \text{AvgPool}(F_{\text{emb}}(I), o_i). \tag{4}$$

We define the retrieval module $\mathcal{R}$ that takes an object embedding vector $\mathbf{e}_{o_i}$ and retrieves the memory entry $(E_j, (n_j, c_j))$ for a matching object j as:

$$\mathcal{R}(\mathbf{e}_{o_i}) = \begin{cases} (E_j, (n_j, c_j)), \\ \quad \text{if } j = \arg\max_l \{\text{sim}(E_l, \mathbf{e}_{o_i})\} \\ \quad \text{and } \text{sim}(E_j, \mathbf{e}_{o_i}) > \tau \\ \phi, \text{ otherwise} \end{cases} \tag{5}$$

where $\text{sim}(E_l, \mathbf{e}_{o_i}) = \arg\max_k \left\{\text{sim}(\mathbf{e}_l^k, \mathbf{e}_{o_i})\right\}$. We calculate the proposal's similarity to the embeddings of the training views for all personalized objects. Any similarity measure (e.g., cosine similarity) can be employed for this purpose. We set a constant threshold τ to identify the personalized objects. We discard the object proposals in which no matching object is found by the retrieval module $\mathcal{R}$. Our method inherently supports the detection of multiple personalized objects.

[2]Although using the semantic category k_p is recommended for optimal performance, our method attains state-of-the-art results even without it, relying solely on the generic category 'main'. See Appendix A.5 for further discussion.

3.4 Personalized Answer Generation

Once a personalized object is identified, our method generates captions specifically about that object, distinct from general captions a standard LVLM would produce. This involves emphasizing the detected object and incorporating prior knowledge about it. We achieve this through visual prompting by overlaying bounding boxes on the image and querying the LVLM to generate captions or answer questions focused on these objects. We use distinctive colors to differentiate recognized objects. We provide the LVLM with the object identifier n_j (e.g., the instance name) and possibly a context c_j for each personalized object. The LVLM incorporates this context and responds to queries using the given name n_j. For multiple personalized objects, we instruct the LVLM for each bounding box, name, and context. We refer to Appendix C.1 for the exact prompt format.

3.5 Video-personalization

To conduct personalized video question answering (VQA), we apply our method to the video frames. We first identify any presence of a personalized object, then each personalized object is consistently visually prompted using the same bounding box color across all frames to maintain identity. The video frames—including those with visual prompts and those without any detected personalized objects—are fed into a video large language model.

The model is then prompted to generate captions or answer questions using each detected object's identifier n_j, an optional context c_j, and the unique bounding box color used for the visual prompting of that object instance (Appendix C.1).

3.6 Choice of Vision Tools

We deploy GroundedSAM Ren et al. (2024) as the open-vocabulary segmentation network F_{ext} and ablate using GroundingDINO Liu et al. (2023c), where the mask is represented by the object's bounding box. We use DINOv2 Oquab et al. (2023) as the image encoder F_{emb} to extract patch-level features of the personalized objects. Image encoders trained with a self-supervised objective, such as DINOv2, produce distinctive features that improve the re-identification of personalized objects as we empirically show in Appendix A.1 in a comparison with CLIP Radford et al. (2021). During inference, GroundingDINO Liu et al. (2023c) is queried with the generic term 'object' to generate object proposals.

4 Experiments

In this section we present the considered benchmarks and introduce our LVLM personalization benchmark. We then compare our approach with SOTA personalization methods and baselines under various settings.

4.1 Existing Benchmarks

We consider the datasets from **Yo'LLaVA** Nguyen et al. (2024) and **MyVLM** Alaluf et al. (2024). Yo'LLaVA includes 40 categories of objects, buildings, and people, with 4–10 images per category for training and validation. It also features a VQA benchmark with multiple-choice questions (A/B). MyVLM comprises 29 object categories, each with 7–17 images, using 4 images for training and the rest for validation, with final accuracies averaged over 5 runs. For both datasets, we use only reference images and the semantic category of each object for training view extraction (see 3.2), discarding all other data, such as the ground-truth captions and negative images.

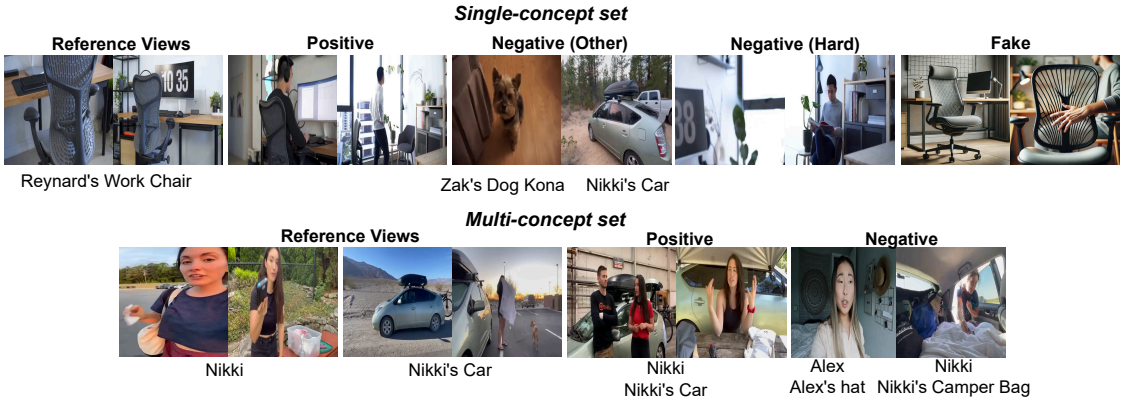

Figure 2: **Our proposed evaluation set This-Is-My-Img, built on the *This-Is-My* dataset Yeh et al. (2023)**: Example reference views and validation samples from the single-concept category *Reynard's Work Chair* and the multi-concept category *Nikki* and *Nikki's Car*.

4.2 Real-world Personalization Benchmark

Yo'LLaVA Nguyen et al. (2024) and **MyVLM** Alaluf et al. (2024) datasets primarily focus on object-centric scenarios in controlled settings that lack real-world complexity, such as background clutter, partial occlusions, and diverse contexts, limiting their effectiveness for evaluating robust LVLM personalization.

To address this gap, we introduce a more realistic evaluation set called **This-Is-My-Img** benchmark, based on the *This-Is-My* dataset Yeh et al. (2023), originally developed for video-level detection of personalized instances in realistic settings where objects may be partially occluded, in the background, or poorly lit. We design several splits to evaluate specific capabilities of personalization methods.

Reference Views: As in prior work, we extract five reference images per object from training segments. Unlike earlier datasets, objects in these frames may appear under suboptimal conditions—e.g., partially visible, in the background, or poorly lit (Fig. 3 Appendix).

Single-Concept Eval Set: Set of images designed to evaluate a single personalized object. Fig. 2 shows examples. For each personalized object, images are labeled as:

- *Positive*: This subset of the dataset includes only images where the personalized object appears and is used to test how well the method can correctly detect all occurrences of that object, that is, recall or *positive accuracy* in **Yo'LLaVA** Nguyen et al. (2024) and **MyVLM** Alaluf et al. (2024). This split consists of 2257 images.
- *Negative (Hard)*: Images taken from video segments of the personalized object **where the object is not visible**. This tests method's overfitting to the background associated with the personalized object in the reference views. This splits consists of 631 images.
- *Negative (Other)*: Images sourced from video segments of other personalized objects. This set measures method's robustness to intra-instance and the general dataset bias. This split consists of 29341.

- *Negative (Fake)*: AI-generated images (via GPT-4o) that mimic the personalized object in different perspectives and environments while making sure the specific characteristics of the fake object are different from the real personalized object. These samples are used to assess the robustness of the methods against high visual similarity combined with a distribution shift. This split consists of 140 images.
- *Single-Concept VQA:* Multiple choice questions based on positive images, following the Yo'LLaVA benchmark VQA format. This set consists of 70 image and QA pairs.

Multi-Concept Set: Images containing pairs of personalized concepts, such as a person-object pair (e.g. Nikki and Nikki's car) or a person-person pair (e.g. Casey and Casey's son). To construct these, we augment the personalized instances from the original This-Is-My dataset with additional instances that frequently co-occur with them. Refer to Appendix B.1 for the original personalized instances and Appendix B.2 for the proposed personalized instance pairs.

we collect images from all segments that contain pairs of personalized concepts and split them into two sets for each pair:

Positives: For a pair of personalized objects (e.g. Alex and Alex's hat), this set includes images with both instances of the pair appearing together while negatives include all images from other pairs in the dataset. See Fig. 2 for examples. The positive and negative sets include 55 and 550 images respectively.

- *Multi-Concept VQA:* Provides open-ended questions for each image in the multi-concept set that address both objects depicted. This set includes 55 images and QA pairs.

Video-QA Set: To rigorously examine our method's ability to generalize across the video modality and dynamic contexts, we construct a Video-QA benchmark grounded in the This-Is-My dataset. In this benchmark, original video clips are coupled with open-ended questions concerning the personalized object in each video, providing a means to evaluate both temporal reasoning and the method's adaptability to video-based inputs. This set includes 267 QA pairs. See appendix B.3 for more details on the construction of this set.
Our evaluation benchmark presents a comprehensive challenge for LVLM personalization, requiring recognition and reasoning over personalized objects in realistic, cluttered, and dynamic conditions.

4.3 Metrics and Evaluation Setting

Single-Concept Metrics. We follow the metrics and notation introduced in Yo'LLaVA Nguyen et al. (2024) and MyVLM Alaluf et al. (2024), reporting *Positive Accuracy (Recall)*—the proportion of true positive instances correctly identified, *Negative Accuracy (Specificity)*—the proportion of true negative instances correctly identified, and *Weighted Accuracy*, which averages Positive and Negative accuracies across n personalized objects. In addition, we report *Precision*, defined as the ratio of true positive predictions to the total number of predicted positives, averaged over n personalized objects. We argue that this provides a more comprehensive evaluation, capturing the trade-off between the above metrics and precision. VQA accuracy on single-concept datasets is measured as the percentage of correctly answered multiple-choice questions. In line with MyVLM Alaluf et al. (2024), on the MyVLM benchmark, we also report *CLIPScore*, which evaluates the similarity between the generated caption and the image, and *Personalization Recall*, which measures how

MyVLM Dataset				
Method/Metric	Precision	Accuracy		
		Positive	Negative	Weighted
MyVLM Alaluf et al. (2024)	–	96.6	90.9	93.8
Yo' LLaVA Nguyen et al. (2024)	–	97.0	95.7	96.4
PeKit (G-DINO)	79.1	94.3	98.8	96.55
PeKit (G-SAM)	**82.3**	**97.6**	**99**	**98.3**
Yo' LLaVA Dataset				
Method/Metric	Precision	Accuracy		
		Positive	Negative	Weighted
Yo' LLaVA Nguyen et al. (2024)	–	**94.9**	89.8	92.4
PeKit (G-DINO)	**77**	89.9	**98.9**	94.4
PeKit (G-SAM)	74.8	91	98.7	**94.9**

Table 1: **Visual Recognition performance (%) on MyVLM and Yo'LLaVA datasets.** PeKit achieves state of the art performance and improves significantly over existing methods.

effectively the model incorporates the personalized object's name into the caption on the MyVLM dataset, defined as the fraction of generated captions that correctly mention the target concept in the image.

Multi-Concept Metrics. Positive accuracy is defined as the proportion of images in which both personalized objects are correctly detected. Negative accuracy is defined as the proportion of images in which at least one of the two personalized objects is correctly **not** detected.

Video and Multi-concept VQA Metric. For open-ended VQA validation, we adopt an auto-grading protocol inspired by Maaz et al. (2023), using ChatGPT-3.5 to assess the accuracy of the model's responses. The GPT model is instructed to evaluate whether the predicted and ground-truth answers are contextually similar rather than exact matches (see Appendix C.2 for details).

Implementation Details. The modularity of our method makes it generic regarding the choice of the LVLM model. We use LLaVA1.5-13B Liu et al. (2023a) as our primary LVLM model for consistency with previous works. We pair PeKit with InternVL2-26B Chen et al. (2024) , a state-of-the-art LVLM to ablate. We use LLaVA-OneVision-Qwen2-7B as our default video model and ablate with InternVideoChat2.5 as a superior video model. We utilize Cosine Similarity with a constant threshold of τ=0.75 for detecting personalized objects across all datasets (3.3). Refer to Appendix A.7 for further details on compute and memory overhead of our personalization toolkit.

4.4 Visual-Recognition

Previous Datasets Tab. 1 compares our method to MyVLM Alaluf et al. (2024) and Yo'LLaVA Nguyen et al. (2024) on their respective benchmarks. On MyVLM benchmark, PeKit achieves a new SOTA, improving both positive and negative accuracy with an average gain of 1.9%. On Yo'LLaVA benchmark, it increases negative accuracy by 8.9%, with slightly lower positive accuracy, resulting in a 2.5% improvement in weighted accuracy. While PeKit may miss some difficult instances, its low false positive rate helps avoid incorrect personalized outputs, favoring generic captions when uncertain. We ablate the choice of an open-world object detector, G-DINO Liu et al. (2023c), compared to the open-world semantic segmentation model, G-SAM Ren et al. (2024), for F_{ext} in Eq. 1. The results show that our method outperforms previous methods in both cases. However, the more precise segmentation model, which extracts only patches of the object of interest, achieves the best accuracy on average.

This-Is-My-Img Single-Concept Dataset						
Method/Metric	Precision	Accuracy				Avg.
		Positive	Negative			
			Other	Hard	Fake	
MyVLM	8.0	**88.1**	14.6	4.7	54.2	33.9
Yo'LLaVA	42.1	87.1	84.4	61.9	**61.4**	67.3
PeKit	**90.1**	69.0	**99.9**	**96.0**	59.3	**82.8**

This-Is-My-Img Multi-Concept Dataset				
Method/Metric	Precision	Accuracy		Avg.
		Positive	Negative	
Yo'LLaVA	10.0	**83.6**	25.0	54.3
PeKit	**96.1**	45.4	**99.8**	**72.6**

Table 2: **Visual Recognition Performance (%) on This-Is-My-Img dataset.** MyVLM and Yo'LLaVA tend to capture the dataset's bias and produces many false positives when the images are taken from similar scenes regardless of the object's presence.

MyVLM Dataset		
Method	CLIPScore	Personal. Recall
MyVLM Alaluf et al. (2024)	27.6	94.76
PeKit (LLaVA)	**30.2**	**97.1**

Yo'LLaVA Dataset	
VQA Accuracy	
LLaVA Nguyen et al. (2024)	89.9
Yo'LLaVA Nguyen et al. (2024)	92.9
PeKit (LLaVA)	93.4
PeKit (InternVL)	**95.9**

This-Is-My-Img Single-Concept Dataset	
VQA Accuracy	
LLaVA	72.8
Yo'LLaVA	67.1
PeKit (LLaVA)	77.1
PeKit (InternVL)	**84.2**

This-Is-My-Img Multi-Concept Dataset	
VQA Accuracy	
LLaVA	49.0
Yo'LLaVA	12.7
PeKit (LLaVA)	56.3
PeKit (InternVL)	**63.6**

This-Is-My-Video Dataset	
VQA Accuracy	
LLaVA-OneVision-Qwen2	23.0
PeKit (LLaVA-OneVision-Qwen2)	35.0
InternVideoChat2.5	56.6
PeKit (InternVideoChat2.5)	**61.3**

Table 3: **Visual Question Answering Performance.** PeKit achieves SOTA performance on all benchmarks. Cited entries are taken from their respective papers.

This-Is-My-Img Benchmark. Tab. 2 reports the performance on our benchmark. For single-concept images, **MyVLM** shows high positive accuracy but poor negative accuracy (14.6% on *other* and 4.7% on *hard* negatives), suggesting it focuses on scene context over specific objects and tends to provide mostly positive responses to test images. **Yo'LLaVA** performs similarly but better, with 87.1% positive accuracy and 84.8% on *other* and 61.9% on *hard* negative samples. In contrast,

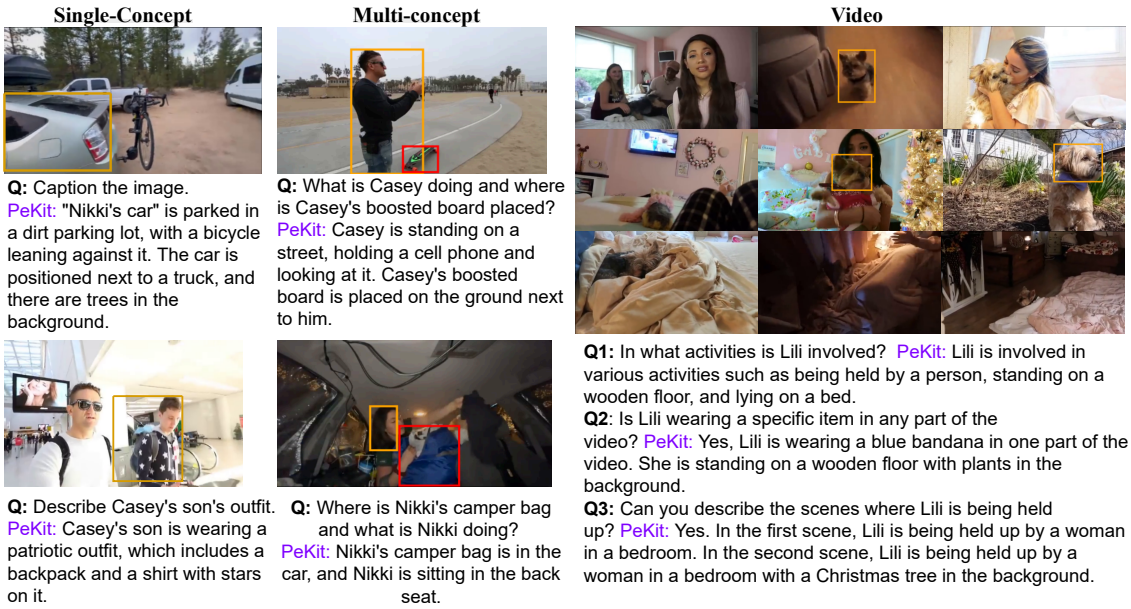

Figure 3: **Qualitative Results:** PeKit handles a range of personalization tasks, encompassing both single- and multi-concept personalization in images and videos. For video personalization, the VLM model can reliably track the target object across frames using only a few confidently annotated instances. One representative frame is shown per scene.

PeKit offers stronger negative accuracy (99.9% and 96.0%) with a precision of 90.1%—48% higher than Yo'LLaVA—indicating fewer false positives.

On fake images, differences narrow. All models struggle with objects stylistically close to the personalized one with image-level methods like MyVLM and Yo'LLaVA potentially having a higher chance to detect the domain shift. However, PeKit also demonstrates robustness to domain variations, despite some difficulty with highly similar objects. Overall, PeKit achieves the best balance, with an average accuracy of 82.86%, outperforming Yo'LLaVA (67.38%) and MyVLM (33.92%).

In multi-concept settings, PeKit shows slightly lower accuracy in dual-object detection but achieves much higher precision and negative accuracy, leading to an 18.3% overall accuracy gain over Yo'LLaVA. These results highlight our training-free method's superior robustness and lower false positive rate, especially in complex real-world scenarios.

Overall, this evaluation set illustrates the difficulty of the personalization task beyond object-centric images, moving closer to realistic scenarios for intelligent visual assistants. This opens the door for future research on this challenging benchmark.

4.5 Visual-Question Answering

In Tab. 3, we assess PeKit performance in answering questions about personalized objects across various scenarios. We evaluate our approach using the Yo'LLaVA Nguyen et al. (2024) and This-Is-My-Img benchmarks. For a comprehensive comparison, we also evaluate PeKit against MyVLM on the MyVLM dataset using CLIPScore Hessel et al. (2021) and Personalization Recall, as it does not include a VQA validation set.

Our method consistently outperforms baseline models on both single- and multi-concept VQA tasks. On the MyVLM dataset, PeKit surpasses the MyVLM method in generating image-aligned captions (as measured by CLIPScore) and in correctly incorporating personalized concept names (as measured by Personalization Recall). On the Yo'LLaVA dataset, PeKit outperforms Yo'LLaVA despite requiring no additional tuning, special tokens, or modifications to the original LVLM.

On the This-Is-My-Img benchmark, Yo'LLaVA fails to improve over the base LLaVA model, whereas PeKit achieves a notable performance gain of approximately 5%, demonstrating strong capabilities in both visual recognition and question answering. On the multi-concept set, Yo'LLaVA substantially reduces the performance of the original VLM, whereas our approach consistently achieves an improvement of approximately 7% over the original model. Additionally, PeKit generalizes well to the video domain (3.5), unlike prior approaches, which are not readily extendable to video. In this setting, our method improves the original model's performance by 12% and continue to show gains when paired with a more capable instruction-following backbone.

It is worth mentioning that while our method demonstrates superior performance compared to prior training-based approaches, it is subject to certain limitations, which are discussed in Appendix D.

4.6 Qualitative Results

Fig. 3 presents examples of PeKit performing various tasks on the proposed benchmark, This-Is-My-Img. In the left column, PeKit successfully identifies personalized single-concept objects, even in ambiguous scenarios involving multiple instances from the same object category. In the middle column, it demonstrates the ability to answer questions involving multi-concept pairs. Lastly, without any modifications, the model can be applied to video inputs by visually prompting the model with detected personalized objects in the video. Further qualitative comparisons with the base LLaVA model, MyVLM, and Yo'LLaVA are presented in Appendices E.2, E.3, and E.4 while Appendix E.1 presents a real-world demonstration of PeKit, where it is integrated into a robot to locate and grasp a personalized object.

4.7 Ablation and Analysis

A comprehensive analysis of our method is provided in Appendix.A, where we analyze robustness under various settings. Specifically, we examine the effects of altering the backbone encoder F_{emb}, the number of reference images per object (N), the number of personalized objects (P) in the dataset, the detection threshold (τ), semantic category (k_p) and the name (n_p) of the personalized objects.

5 Discussion and Conclusion

In this work, we introduced PeKit, a training-free, plug-and-play toolkit for LVLM personalization that combines vision foundation models with retrieval-augmented generation and visual prompting. Our approach outperforms existing training-based methods without requiring any fine-tuning or additional data beyond the personalized concept inputs. To evaluate our method, we proposed a challenging new benchmark that reflects more realistic scenarios involving multi-concept and video personalization, significantly increasing the complexity of the visual recognition task. PeKit demonstrates strong robustness and scalability across both multi-concept and video settings. Our benchmark reveals strong performance and shows some improvement points, particularly in handling complex temporal and multi-concept reasoning. We believe that PeKit establishes a new efficient training-free baseline for future research in LVLM personalization.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL `https://arxiv.org/abs/2410.07073`.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. *arXiv preprint arXiv:2403.14599*, 2024.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12914–12923, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pp. 558–577. Springer, 2022.

Deepayan Das, Davide Talon, Yiming Wang, Massimiliano Mancini, and Elisa Ricci. Training-free personalization via retrieval and reasoning on fingerprints. *arXiv preprint arXiv:2503.18623*, 2025.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Yikun Han, Chunjiang Liu, and Pengfei Wang. A comprehensive survey on vector database: Storage and retrieval technique, challenge. *arXiv preprint arXiv:2310.11703*, 2023.

Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023a.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *arXiv preprint arXiv:2406.09400*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Personalized large vision-language models, 2024. URL https://arxiv.org/abs/2412.17610.

Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*, 2024.

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

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Andrea Rosasco, Stefano Berti, Giulia Pasquale, Damiano Malafronte, Shogo Sato, Hiroyuki Segawa, Tetsugo Inada, and Lorenzo Natale. Concon-chi: Concept-context chimera benchmark for personalized vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22239–22248, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6527–6536, 2024.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8543–8552, 2024.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *ECCV*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Chun-Hsiao Yeh, Bryan Russell, Josef Sivic, Fabian Caba Heilbron, and Simon Jenni. 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.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.