Ego: Embedding-Guided Personalization of Vision-Language Models

Soroush Seifi* Simon Gardier Vaggelis Dorovatas* Daniel Olmeda Reino Rahaf Aljundi
Toyota Motor Europe
{soroush.seifi, vaggelis.dorovatas}@external.toyota-europe.com
{simon.gardier, daniel.olmeda.reino, rahaf.al.jundi}@toyota-europe.com

Abstract

AI assistants that support humans in daily life are becoming increasingly feasible, driven by the rapid advancements in multimodal language models. A key challenge lies in overcoming the generic nature of these models to deliver personalized experiences. Existing approaches to personalizing large vision language models often rely on additional training stages, which limit generality and scalability, or on engineered pipelines with external pre-trained modules, which hinder deployment efficiency. In this work, we propose an efficient personalization method that leverages the model's inherent ability to capture personalized concepts. Specifically, we extract visual tokens that predominantly represent the target concept by utilizing the model's internal attention mechanisms. These tokens serve as a memory of that specific concept, enabling the model to recall and describe it when it appears in test images. We conduct a comprehensive and unified evaluation of our approach and SOTA methods across various personalization settings including single-concept, multi-concept, and video personalization, demonstrating strong performance gains with minimal personalization overhead.

1. Introduction

Large Language Models (LLMs) have recently demonstrated impressive capabilities in understanding, reasoning, and generating text across diverse domains [23, 30, 33]. Extending these capabilities to the multimodal domain, Large Vision-Language Models (LVLMs) have achieved notable success in tasks such as image captioning [27], visual question answering [18, 20], and embodied navigation [32]. These advances position LVLMs as promising general-purpose perceptual agents. As human-AI interaction becomes increasingly multimodal, the ability of LVLMs to detect, understand, and reason about individual users and their belongings is essential for personalized experience.

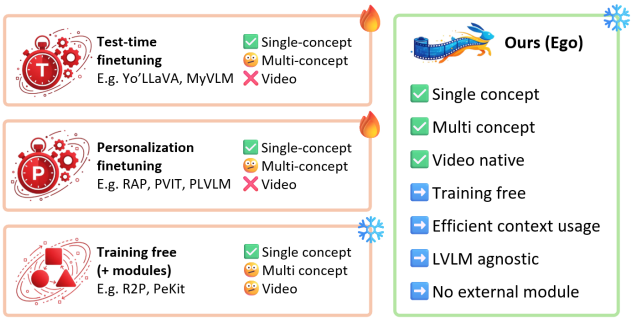

Figure 1. **Personalization approaches vs *Ego***. Existing methods typically require test-time or LVLM fine-tuning, or depend on external vision modules, and often fail to support multi-concept or video-level personalization. In contrast, *Ego* is training-free, LVLM-agnostic, requires no external modules, and efficiently enables single-concept, multi-concept, and video-native personalization within a unified framework.

Personalization of Large Vision-Language Models seeks to adapt pre-trained models to recognize, describe, and reason about user-specific entities [1, 24]. Unlike general-purpose LVLMs that operate at the category level, personalization focuses on capturing the unique visual and semantic traits of specific subjects—such as a person, object, or pet—using limited reference data. This capability unlocks a wide range of applications, including personalized text generation, user-specific assistants, and long-term embodied agents that maintain consistent knowledge of users and their environments. More broadly, effective personalization bridges the gap between general visual-language understanding and individualized, context-aware intelligence.

Despite increasing interest in personalization, current approaches face several practical limitations. Many methods rely on test-time finetuning for each individual subject [1, 24], which significantly hampers scalability, especially on resource-constrained edge devices. Some recent works attempt to bypass test-time finetuning by leveraging large-scale training and instruction tuning of LVLMs to generate personalized outputs [11, 26]. However, even after training for personalization, these models typically require

*Providing contracted services at Toyota Motor Europe.

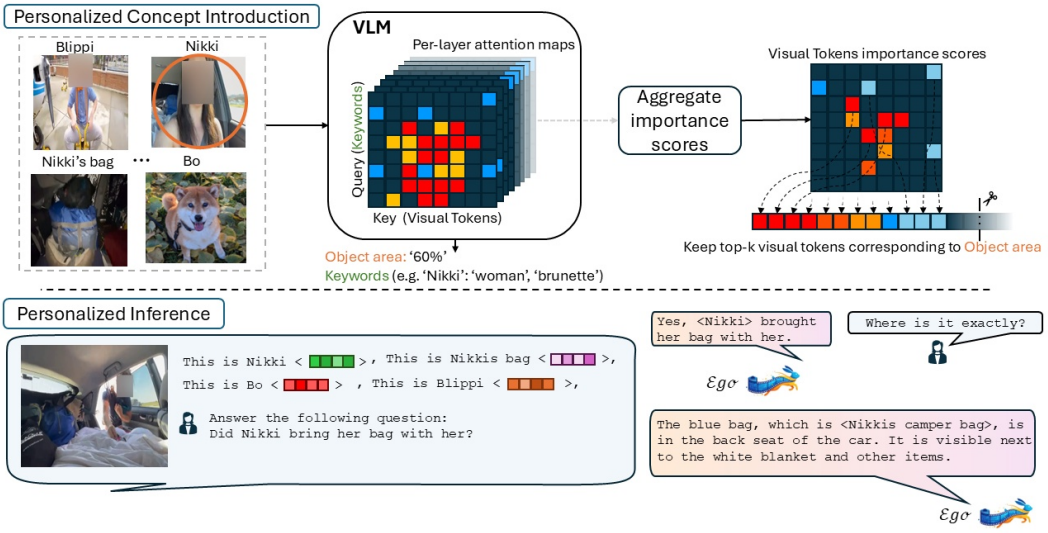

Figure 2. **Our proposed method *Ego*. Personalized Concept Introduction:** The LVLM is tasked to estimate the subject area in the reference image and generate keywords describing its main characteristics. *Ego* identifies the most representative visual tokens via keywords cross cross-attention and creates a concept memory. **Inference:** Given a test image, the LVLM in *Ego* accesses internal concept memories **in context** to recall and reason about known subjects in the image. *Ego* requires neither additional training nor external modules.

reference views of the personalized concepts during inference [11, 25, 26], adding computational overhead and forcing the model to reprocess the personalized input each time. On the other hand, training-free approaches often depend on heavily engineered architectures or external vision modules [7, 31], leading to increased system complexity and inference-time overhead, refer to Fig.1 for an illustration.

Leveraging the strong image understanding and in-context learning capabilities of modern LVLMs [5, 15], our method eliminates the need for additional training, fine-tuning, or external tools. We address personalization by enabling the model to build an internal memory of personalized concepts. Upon introduction of a concept via one or a few reference views, we task the model to describe its distinctive features. During this process, we extract the most attended visual tokens—those the model deems most representative—and aggregate them into a compact concept visual memory. At inference, these concepts memories are provided in context, guiding the model to recognize and personalize its response when the subject reappears, Fig. 2 illustrates our approach.

We term our method *Ego*, short for Embedding-Guided Personalization of Vision-Language Models. *Ego* operates entirely without additional training, external modules, while minimizing inference-time overhead.

In comparing our approach with existing state-of-the-art methods, we identified substantial inconsistencies in the datasets and evaluation protocols employed across prior work. To ensure fairness and reproducibility, we perform a unified evaluation of representative personalization techniques on diverse datasets and tasks, including recognition, visual question answering, and captioning. Our evaluation spans multiple personalization scenarios, covering single-concept, multi-concept, and video personalization where applicable. *Ego* achieves the strongest training-free performance in single-concept personalization and delivers significant improvements over both training-based and non-training approaches in multi-concept and video personalization.

Our key contributions are as follows: 1) We propose *Ego*, a training-free personalization method that requires no fine-tuning, external tools, or architectural changes. 2) We conduct a unified evaluation of state-of-the-art approaches establishing a comprehensive testbed for future works. 3) *Ego* achieves state-of-the-art performance with minimal compute overhead. 4) *Ego* supports single-concept, multi-concept, and video personalization within a unified model-generic framework.

2. Related Work

Personalization of pretrained models refers to adapting the behavior of generic models to fit specific user profiles, concepts, or styles, and it has attracted growing interest across multiple domains. Large Language Model (LLM) personalization focuses on tailoring models to user or group preferences derived from past interactions or explicit feedback (e.g., [17, 34, 36]), while image and video generation personalization enables integrating specific concepts into generated media (e.g., [28, 29]). Encoder-based vision-language personalization typically addresses concept retrieval (e.g., [6, 10]). In this work, we focus on Large Vision-Language Model (LVLM) personalization, which

aims to enhance LVLMs' ability to understand, reason, and respond to visual input by incorporating knowledge about personalized concepts such as individuals and their belongings. Current LVLM personalization approaches can be broadly categorized into the following lines of work.

Test-Time Fine-Tuning. Early attempts to LVLM personalization approach the task via a dedicated training phase for each introduced concept. **MyVLM** [1] learns binary classification heads for each personalized concept and injects concept identifiers through a Q-former mechanism, while **Yo'LLaVA** [24] employs prompt-tuning with trainable prefix tokens per concept and fine-tunes LLM classifier weights. Both rely on caption-level supervision and QA pairs, focusing on object-centric evaluation without addressing multi-concept or video personalization [31].

Finetuning for Personalization. Training approaches remove test-time tuning requirement by enabling models to recognize personalized concepts directly from reference images [11, 25, 26]. **PLVLM** [25] trains an aligner module using DINOv2 embeddings but focuses on human-centric evaluation without exploring generalization to other categories. **PVIT** [26] builds a synthetic dialogue dataset for personalized conversations and fully fine-tunes the LVLM on it. **RAP** [11] retrieves candidate concepts via visual similarity and uses LoRA-based fine-tuning on large scale paired data. Although these methods avoid per-concept tuning, they remain resource-intensive and rely on reference views at inference, which can introduce context-length bottlenecks. Furthermore, training biases the model towards the constructed personalization paired data, limiting scalability to multi-concept setting, as shown in our experiments.

Training-Free Methods: As LVLMs become more powerful, training-free methods aim to enable personalization without altering the LVLM, prioritizing efficiency and scalability. **R2P** [7] generates descriptive attributes per concept and detects them via top-k retrieval using external vision models [14, 27]. **PeKit** [31] decouples object detection from LVLM reasoning using an external segmentation network and a DINOv2-based memory bank. These methods still depend on external modules and trade off training for additional test-time computation. In contrast, our approach, *Ego*, tackles personalization by leveraging the LVLM's strong visual understanding capabilities. We create visual concept memories that compress subject attributes and are accessed in context by the LLM, requiring minimal overhead equivalent to textual prompting at inference.

3. Ego: Embedding-Guided Personalization

Our approach is motivated by the observation that recent powerful LVLMs are capable of cross-referencing objects in multiple input images and reasoning about objects in short videos [2, 38]. This shows an inherent ability to recognize objects across different images or video frames, implying that the model internally assigns discriminative embeddings to individual objects for recognition and tracking.

Leveraging this insight, we want to extract these discriminative embeddings from the LVLM's intermediate representations and provide them as in-context information, enabling the model to identify and reason about personalized concepts. The proposed method is modular, scalable to multiple concepts and video inputs, and compatible with any number of reference views. The following sections detail each component of our approach.

3.1. Preliminaries

Given a Large Vision Language model $\mathcal{M}$ composed of an LLM, and a visual encoder with a projector (collectively referred to as VP) that maps input image(s) into the LLM embedding space referred to hereafter as **visual tokens**. Our objective is to enable $\mathcal{M}$ to generate textual outputs tailored to personalized concepts. For a concept c to be added to the personalized concept set C, we receive a set of reference images $\{R_c\}$ along with its concept name n_c. Our method works with one or multiple reference images. For each reference view R_c of concept c, we query $\mathcal{M}$ with an instruction I to produce key descriptive words. Formally:

$$T = \mathcal{M}(R_c, I);\ T = \text{LLM}(\mathbf{X}_R, I),\ \text{where}\ \mathbf{X}_R = \text{VP}(R_c), \tag{1}$$

$\mathbf{X}_R \in \mathbb{R}^{N_r \times D}$, N_r is the number of visual tokens extracted from a reference image and D is the LLM token embedding dimension, and T is the textual output of the LLM. From T we filter out punctuations and keep textual tokens corresponding to the LLM output keywords $\mathbf{W}$. Our objective is to select a compact subset of visual tokens that best represent the personalized subject in the given reference image. This is motivated by recent works highlighting redundancy in visual tokens and showing that selecting an informative subset increases efficiency and can often outperform using the full image or video [4, 8, 12].

3.2. Attention-Guided Embedding Extraction

A reference image R_c, mapped into visual tokens $\mathbf{X}_R$ by VP, contains the subject and unrelated background. We aim to identify and extract a minimal subset of visual tokens $\mathbf{X}_R^c \in \mathbb{R}^{K \times D}$ (where $K << N_r$) that captures the unique characteristics of a given concept c. This approach serves two main purposes. First, efficiency: representative tokens can be aggregated from multiple reference views while keeping the overall token count per concept manageable for personalized inference. Second, relevance: it is essential to remove tokens that do not represent the subject, such as those corresponding to background elements, so the concept representation remains focused and accurate.

To identify concept-specific embeddings, we analyze the attention maps of the LLM layers, specifically, we focus on

cross-modal attention of the keywords tokens $\mathbf{W}$ to the visual tokens $\mathbf{X}_R$. We hypothesize that the representative visual tokens corresponding to the descriptive keywords will receive the highest attention scores by the keywords embeddings and, the higher the attention score, the higher the importance of the visual token. First, we describe how to compute importance scores from attention maps per visual token given all LLM attention heads and layers, then we explain how we identify the most relevant layers for visual objects understanding in a given $\mathcal{M}$.

For a layer l and attention head h, the full attention matrix is computed on the embedding $\mathbf{X}^l$ after mapping it to Query $\mathbf{Q}^{l,h}$, Key $\mathbf{K}^{l,h}$ and Value $\mathbf{V}^{l,h}$

$$\mathbf{A}^{l,h} = \text{Softmax}\left(\frac{\mathbf{Q}^{l,h}(\mathbf{K}^{l,h})^\top}{\sqrt{d_k}}\right), \tag{2}$$

where d_k is the dimension of attention heads. We extract the cross attention matrix $A^{l,h}_{wr} \in \mathbb{R}^{N_w \times N_r}$ computed on the embedding at layer l, $\mathbf{X}^l = [\mathbf{X}^l_R, \mathbf{X}^l_W]$ where N_w is the number of tokens in the key descriptive words. The extracted $A^{l,h}_{wr}$ represents the cross attention scores between the keywords embedding $\mathbf{X}^l_W$ and the visual tokens $\mathbf{X}^l_R$.

Given a set of relevant LLM layers L, we compute an importance score for each visual token $\mathbf{X}^l_R[j]$ by maximizing attention scores over heads and layers and averaging the scores from the different keywords tokens .

$$I_j = \frac{1}{|L|}\sum_{l\in L}\frac{1}{H}\sum_{h=1}^{H}\left(\frac{1}{N_w}\sum_{n=1}^{N_w}\mathbf{A}^{l,h}_{wr}[n,j]\right). \tag{3}$$

With an importance score per visual token, we select the K_c most important visual tokens of $\mathbf{X}_R$ as representative of the concept introduced for personalization. Specifically:

$$\mathcal{P}^{\text{ordered}} = \overbrace{\text{sort}_\uparrow}^{\text{restore order}}\left(\overbrace{\text{argsort}_\downarrow(\mathbf{I})[1:K_c]}^{\text{select top-}K_C}\right),$$

$$\mathbf{X}^c_R = \mathbf{X}_R[\mathcal{P}^{\text{ordered}}, :].$$

As noted before, we select a subset of the visual tokens $\mathbf{X}^c_R \in \mathbb{R}^{K_c \times D}$ that received the highest attention. When provided with multiple reference views N_v, we process each image independently to extract the top visual tokens which will be concatenated into one matrix: $\mathbf{X}^c_R \in \mathbb{R}^{N_v * K_c \times D}$. We construct $\mathbf{X}^c_R$ to represent the LVLM memory of the personalized concept and its most unique visual characteristics, acting as the visual highlights of a given subject. During inference, the model receives each concept's visual memory $\mathbf{X}^c_R$ along with its name n_c, and is tasked with determining whether the personalized concept(s) appear in the inference image and respond to the input query accordingly.

3.3. Concept Memory Size

We select a small set of visual tokens, denoted as K_c, to capture the model's memory of the personalized concept. The optimal number of representative tokens depends on the size of the object in the reference image. For instance, when the object occupies a small area (e.g., a phone or a pair of shoes), only a few tokens are sufficient. Conversely, for larger subjects such as a person in a high-resolution profile image, a greater number of tokens is preferred.

To build a compact memory of a personalized subject while considering the area it occupies in the image, we leverage the LVLM's inherent ability to estimate the size of the region taken up by the main subject in the reference image. Given a reference image R_c of a newly introduced concept c, we first ask the LVLM to estimate the percentage of the area that the subject occupied in the image (α_c) and then estimate the appropriate number of visual tokens as $K_c = min(K, \frac{\alpha_c \times N_r}{100})$, where K is the maximum number of visual tokens allowed per reference image. Setting a maximum number $K << N_r$ is essential to maintain efficiency during inference and to extract only patches representing the key attributes of the subject.

3.4. Layer Selection

LVLMs encode information across layers with varying levels of abstraction. Our objective is to identify the layers that exhibit the strongest interaction between visual tokens and the generated keywords for a given subject. Prior work suggests that mid-to-late layers enhance visual representations [13], dominate visual-to-text information flow [16], and enable efficient fusion through attention to text-relevant regions [9]. However, there is no established method for determining vision-relevant layers in a specific LVLM. To address this, we propose an automatic procedure that identifies where the generated text interacts most with the visual tokens in an image, using an external calibration set.

We sample a subset of images from the COCO 2017 training split [19], each containing a single category and one instance of that category. Using the ground-truth segmentation mask for each object instance, we determine which visual tokens correspond to the object. The LVLM is then tasked with describing the main foreground object in each image. For every layer, we compute the overlap between the top K patches (ranked by our importance score) and the segmentation mask, and rank layers based on their average overlap. The top L layers are selected.

This process yields a set of layers where image descriptions typically interact most strongly with the visual information of the main subject, serving as a proxy for our task of understanding a subject in a reference image. Note that this calibration is performed only once per LVLM.

3.5. Ego Inference

Ego accesses the model's internal states and attention maps to construct a model's own memory $\{\mathbf{X}_R^c, n_c\}$ of each personalized concept. At inference, we retrieve the memories of the personalized concepts and inject them into the context of the LLM as soft prompts. We then instruct the LVLM to check if the provided concepts are present in the image and to respond to the query accordingly. When the number of the personalized concepts increases beyond the context limit of a given LVLM, a filtering step can be done based on the similarity of the query image (in the LLM embedding space) to the stored concepts' memories. Note that by storing the visual tokens in the LLM embedding space, we avoid the need for reprocessing the reference view at test-time by the visual encoder and allow for a small memory and compute footprint during inference, as we show in the experiments 4. Our soft prompting leverages recent LVLMs capability of In Context Learning [3, 5], alleviating the need for additional stages of training and alignment. This mechanism supports single-concept, multi-concept, and video-level personalization with a unified procedure.

4. Experiments

We present a comprehensive evaluation of *Ego* across diverse personalization settings. We compare its performance against both training-free and training-based SOTA methods, analyze run-time, and conduct ablation studies to assess the impact of key design choices.

4.1. Experimental Settings

Existing personalization methods vary widely in datasets and evaluation metrics [1, 11, 24, 31]. Even when using the same dataset, evaluation splits often differ [7, 11]. To ensure fair comparison, we standardize all experimental settings—model backbones, preprocessing, and evaluation protocols—and reproduce prior results under identical conditions. Our implementation will be publicly released.

Compared Methods We benchmark representative SOTA approaches. For finetuning-based methods, we select RAP [11], which learns from large-scale synthetic personalization examples across tasks. For training-free post-hoc methods, we include R2P [7] and PeKit [31], both neither fine-tune nor modify the base model. We exclude test-time finetuning methods (MyVLM [1], Yo'LLaVA [24]) due to their impractical per-concept fine-tuning and consistent underperformance [7, 11, 31].

Personalization Tasks and Metrics We cover three tasks:

- **Recognition.** Determine whether a concept is present in the query image, responding with *Yes* or *No*. ecognition metrics are computed using the ground-truth validation images of each concept as positive examples, while images from all other categories serve as negative examples for that concept. Each image is evaluated for every concept, assessing robustness to intra-category variation. We avoid pooled negatives [1, 24] for reproducibility.
  Metrics: Previous works have adopted different metrics for evaluating recognition accuracy, including positive accuracy (Recall), negative accuracy (Specificity), and their weighted combination [1, 24]. PeKit [31] additionally incorporates Precision into its reporting. In line with [11], we standardize all comparisons using three conventional binary classification metrics: `Precision`, `Recall`, and `F1-score`. For a single concept, `Precision` represents the fraction of true positives among all predicted positives $\frac{\text{TP}}{\text{TP+FP}}$. Recall captures the fraction of true positives among all ground-truth positives $\frac{\text{TP}}{\text{TP+FN}}$. The F1-score is the harmonic mean of Precision and Recall $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.
  Overall performance is reported by averaging Precision and Recall across all concepts in the dataset, and computing the F1-score from these averaged values. For multi-concept, we compute the same metrics for each concept pair, where a model's response is considered correct only if it identifies all concepts present in the query image.
- **Visual Question Answering (VQA).** The model is presented with a question concerning a specific concept(s) within the query image, expressed through their personalized names. The question may be formulated as either multiple-choice or open-ended. This task assesses the model's ability to effectively collaborate with the user by accurately interpreting and reasoning on personalized references. **Metrics:** We measure `accuracy` as the fraction of questions answered correctly. For multiple-choice questions, we determine correctness through string matching with the ground-truth answer. For open-ended questions, we use ChatGPT to automatically assess whether the model's response is semantically and contextually aligned with the ground-truth answer, following an evaluation protocol similar to [22].
- **Captioning.** the model is required to generate a textual description of the query image that correctly identifies and incorporates the relevant concept name from all available concepts in the dataset. Unlike the recognition and VQA tasks, where the query is constructed based on a known ground-truth concept for a given query and no retrieval is necessary, this task additionally evaluates the model's ability to select the appropriate concept for a given image before generating the caption. **Metrics:** We employ `captioning recall`, calculated as the fraction of query images for which the generated caption correctly references the concept or concept-pair shown in the image, with results averaged across all concepts.

Datasets. We cover a diverse suite of datasets designed to assess the methods' adaptability across *single-concept*, *multi-concept*, and *video-based* personalization scenarios.

Table 1. **Recognition.** Performance comparison across methods and settings using InternVL3 (14B) [38] and Qwen2.5-VL (7B) [2]. Best results are shown in **bold**, and second-best results are underlined. For training-free methods, the Training Time column reflects the average time required for concept introduction. *Ego* attains state-of-the-art recognition accuracy across datasets while introducing concepts with minimal overhead. Note that RAP dataset is limited to a single-reference training set, and R2P [7] does not support multi-concept tasks.

Method	Model	Training Time ↓	Single Concept									Multi Concept					
			MyVLM [1]			Yo'LLaVA [24]			This-is-my (Single) [31]			This-is-my (Multi)			RAP (Multi) [11]		
			Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑
1 Reference View																	
RAP [11]	Intern	24hrs	63.4	98.2	77.0	47.8	93.7	63.3	83.4	**91.3**	**87.1**	**100.0**	62.0	76.5	90.7	**100.0**	95.1
R2P [7]	Intern	5.98s	54.1	93.4	68.5	53.1	85.6	65.5	61.0	76.3	67.7	–	–	–	–	–	–
Ego (Ours)	Intern	1.40s	86.0	94.8	90.2	77.2	83.4	80.2	81.3	77.0	79.1	93.9	78.2	88.6	**100.0**	96.9	**98.4**
	Qwen	**1.25**s	76.9	90.3	83.1	65.9	90.2	76.2	73.3	67.6	70.3	83.4	**87.3**	85.3	94.1	78.1	85.4
5 Reference Views																	
PeKit [31]	–	21.3s	82.3	97.6	89.2	74.8	91.0	82.1	**90.1**	69.0	78.1	96.1	45.4	61.6	–	–	–
Ego (Ours)	Intern	7.00s	**87.7**	**99.0**	**92.8**	**85.0**	86.4	**85.7**	87.2	65.9	75.1	**100.0**	81.8	**90.9**	–	–	–
	Qwen	6.25s	72.1	95.9	82.3	67.9	**98.7**	80.5	77.6	71.9	74.6	92.1	74.5	82.4	–	–	–

- **Single-concept.** We utilize three datasets: *MyVLM* [1], *Yo'LLaVA* [24], and *This-is-my-img* [31]. The *MyVLM* dataset comprises 29 object categories, and we evaluate the single-concept captioning task on this dataset. *Yo'LLaVA* offers broader coverage, featuring 40 categories including objects, cartoon characters, Vietnamese public figures, and architectural landmarks; it also provides multiple-choice (A/B) questions for the VQA task. The *This-is-my-img* dataset presents 14 object and person categories collected from YouTube videos, emphasizing personalized visual understanding under in-the-wild conditions, in addition to single-concept multiple-choice (A/B) VQA questions. To ensure consistent evaluation across all methods, we establish fixed training (reference views) splits with 1 and 5 images per dataset.
- **Multi-concept.** We use both the This-is-my-img [31] and RAP [11] datasets. The multi-concept extension of This-is-my-img expands the original dataset by adding co-occurring people and objects, covering 22 concepts within 11 concept pairs, this is accompanied by a question–answer pair for open-ended VQA task. We evaluate the multi-concept VQA and captioning tasks on this split. The RAP dataset provides a multi-concept validation split spanning 16 concepts and 8 concept pairs, including people, cartoon characters, and animals, offering a challenging benchmark for multi-entity personalization.
- **Video.** We adopt the video question-answering dataset introduced by [31]. This evaluation is conducted on 267 question–answer pairs derived from the original validation segments of the *This-is-my* dataset [35], enabling analysis of the model's temporal reasoning and consistency in personalized Video QA.

Models Early methods for LVLM personalization were based on earlier, less performant LVLMs such as MiniGPT-4 [37] and LLaVA-1.5 [21]. Such models could not process multiple images and lacked visual in-context learning capability [15]. In this evaluation, we focus on more powerful LVLMs that demonstrate strong capabilities across a wide range of vision tasks [1]. Such LVLMs inherently possess better generalization and reasoning capabilities, making effective personalization especially valuable for real-world and complex visual understanding tasks. We evaluate SOTA personalization methods when paired with stronger, stable models while maintaining a feasible scale of model parameters for practical use cases. Namely, we adopt **InternVL3-14B** [38] as the primary LVLM for our method and for all reproduced baselines.We further evaluate our method when paired with a smaller model, **Qwen2.5-VL-7B-Instruct** [2], analyzing the relationship between model capacity and training-free personalization performance.

Implementation Details. We refer to the appendix for the extended implementation details.

4.2. Recognition

We evaluate recognition performance under two reference-view settings: (a) **1-view**, supported by RAP, R2P, and *Ego*, and (b) **5-views**, supported only by PeKit and *Ego*. As shown in Tab. 1, *Ego* achieves the highest F1-scores across most datasets and settings, including challenging multi-concept scenarios that reflect real-world personalization. Moreover, *Ego* requires minimal concept processing time, relying only on brief descriptions rather than extensive reasoning or attribute selection. RAP, despite large-scale finetuning (210k samples) and full reference views, is constrained by top-k retrieval (set to 3), limiting detectable concepts and causing in-context saturation. While RAP excels in Recall, its low Precision reveals a tendency to over-predict, a byproduct of finetuning. Performance drops sharply in multi-concept cases, since the training dataset doesn't represent this case, underscoring the dependency of

[1] VLM ranking: MMBench.

Table 2. **VQA Acc. and Captioning Recall.** Performance comparison (1 ref. view and InternVL3). *Ego* attains competitive accuracy on single-concept VQA while achieving SOTA performance in both multi-concept and video settings. *Ego* delivers superior captioning-recall. R2P only supports single-concept personalization and RAP does not support video personalization.

Method	Sample Runtime (s ↓)	Single Concept			Multi Concept		Video
		VQA (Acc. ↑)		Captioning (Recall ↑)	VQA (Acc. ↑)	Captioning (Recall ↑)	VQA (Acc. ↑)
		Yo'LLaVA	This-is-my	MyVLM	This-is-my	This-is-my	This-is-my
RAP [11]	7.8	**97.6**	90.0	65.6	53.7	43.6	–
R2P [7]	7.0	94.0	82.0	77.5	–	–	–
PeKit	**5.8**	94.6	**92.0**	81.1	51.8	35.2	59.9
Ego (Ours)	6.0	92.3	88.0	**91.3**	**72.2**	**70.9**	**70.0**

training-based methods on curated datasets. R2P, though training-free, inherits the same top-k restriction and fails in multi-concept settings with notably low precision. PeKit performs strongly in single-concept detection but struggles in multi-concept cases because it relies on a *single fixed similarity threshold*, which introduces imbalance: a value suitable for one concept may be overly strict or lenient for another, leading to false negatives and significantly reducing recall, despite maintaining high precision in multi-concept scenarios. *Ego* overcomes these limitations by extracting compact, discriminative visual memory from the VLM's embedding space, guided by its internal attention scores. This directs the model toward the most informative regions from the **model's own view**, filtering out background clutter and avoiding the distractions introduced by full reference images, a common failure mode in RAP and R2P. On RAP, *Ego* improves F1-score by 3.3%, and on the more challenging This-is-my dataset the margin increases to **12%**, showing robustness to occlusion, blur, and other real-world factors. Its ability to leverage multiple reference views further strengthens performance in complex multi-concept settings.

In our protocol, precision and recall are computed over all dataset examples, where positives represent only a small fraction—roughly 1/#concepts. As a result, predicting a non-existent concept severely penalizes precision, as reflected in baseline performance. This design mirrors real-world conditions, where personalized concepts are rare across diverse inputs. A robust model should therefore predict concepts accurately and default to generic outputs under uncertainty. *Ego* achieves a strong precision–recall balance and demonstrates scalability by delivering competitive results even with the smaller Qwen2.5VL backbone. Full comparisons to base models are provided in the Appendix.

4.3. Visual Question Answering

Tab. 2 reports VQA performance in both single- and multi-concept settings. In the single-concept regime, *Ego* performs close to RAP, even though RAP is trained on data containing VQA-specific supervision. Compared to R2P, *Ego* underperforms on Yo'LLaVA but achieves a clear improvement on the more challenging This-is-my dataset.

The advantages of *Ego* become more evident in the multi-concept scenario. On the multi-concept split of This-is-my, *Ego* exceeds RAP by nearly 20%, reflecting its ability to preserve multiple personalized concepts. Moreover, *Ego* applies directly to video personalization without any modifications and outperforms the PeKit pipeline, underscoring its generality across personalization modalities.

4.4. Captioning Recall

In Tab. 2 we present our results on personalized captioning. We observe a clear advantage of *Ego* over prior methods. In MyVLM dataset, *Ego* improves performance by 14% over R2P, while in the challenging multi-concept setting of This-is-my, it achieves nearly 30% improvement over RAP. Notably, *Ego* also achieves faster inference by using a precomputed, informative subset of visual tokens from the reference views for in-context conditioning, avoiding the need to process full reference images with the vision encoder. Unlike previous methods that rely on full reference views and external retrieval modules, *Ego* leverages the VLM's inherent capabilities: (1) strong discriminative power to extract unique visual features for each concept while filtering out background noise, and (2) in-context learning to build an In-Context Memory of concepts and retrieve the correct concept internally through attention during caption generation. This parallel, integrated approach reduces latency, eliminates additional stages, and improves retrieval quality—highlighting the benefit of our method over using the raw reference view, an advantage that was not apparent in simpler tasks such as recognition and VQA.

4.5. Ablation

We analyze the main design choices of *Ego*. Our method relies on keywords produced by the model to extract key visual tokens representing the model's memory of a concept. We evaluate the performance of the model provided with the keywords as the concept memory instead (**Keywords Only**). To evaluate the quality of our visual token selection, we consider two baselines, uniform selection of K visual tokens (Sec.3.3) from the reference view (**Uniform Visual**), and using all the visual tokens in the reference view (**Full Visual**). K here is set to 20% for both Uniform and *Ego*. We also consider the combination of keywords and the full

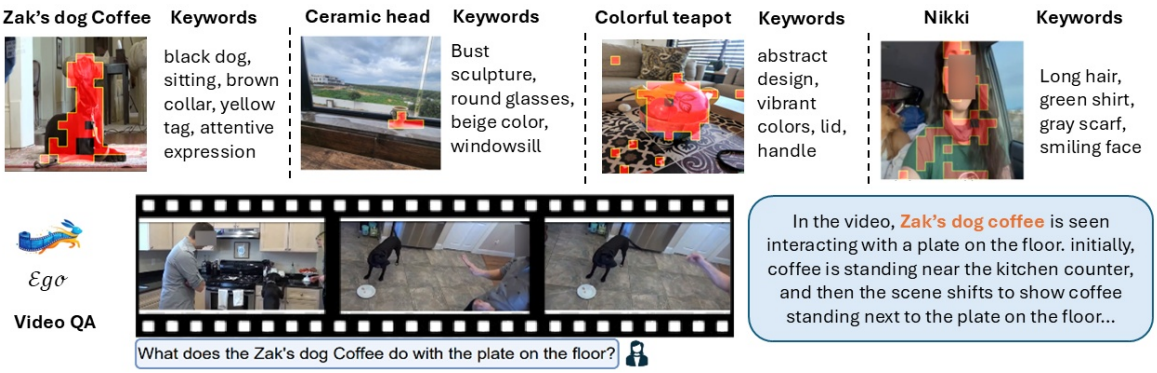

Figure 3. **Qualitative results.** Top row: keywords and highlighted patches of selected visual tokens (*Ego* concept memory) for various concepts, illustrating their representativeness and adaptability to object size. Bottom row: *Ego* demonstrates Video QA capability.

visual tokens (**Full Visual + Keywords**).

Results (F1 score) based on InternVL3 (14B) and the Yo'LLaVA dataset in the recognition task are reported in Tab. 3. It can be observed that descriptive words alone lack the discriminative power to strongly identify the concept, compared to the visual tokens counterpart, and it does not provide any additional useful information but rather a distraction when combined with full visual tokens of the concept. This confirms the importance of constructing a **visual concept memory**. *Ego* significantly outperforms Uniform. Notably, *Ego* with 5 reference views improves significantly over the Full Visual (+1.7%) while using the same number of visual tokens. This indicates that our attention-based selection strategy succeeds in identifying the key visual patches of a concept. We refer to the following analysis and ablations in the Appendix: 1) The effect of dynamic concept memory size K_c versus the default size K (Sec.3.3). 2) Using cross attention to full concept description rather than the keywords, showing the utility of keywords selection. 3) Repeated sampling from the LLM output during keywords generation with results showing no significant impact. 4) Our automatic layer selection strategy (Sec.3.4) compared to tuned layer selection based on downstream performance, showing the stability of our layer selection on the two models. 5) Comparison to the LVLM with full reference views provided in-context on various tasks, showing the effectiveness and efficiency of *Ego*. Note that this baseline does not scale to cases with many personalization concepts. 6) An evaluation of *Ego* across smaller-scale LVLMs to assess how model size influences personalization performance. 7) A comparison between our attention-guided embedding extraction and an alternative approach that uses full segmentation masks, highlighting the benefits of our proposed method.

4.6. Qualitative Results

Fig. 3 illustrates how our attention mechanism localizes representative visual tokens in reference views from the This-is-my [35] and MyVLM [1] datasets. It can be seen how our dynamic size selection strategy effectively reduces redundant tokens and background noise. It also illustrates *Ego* ability to track concepts in a video and answer questions accordingly. Further qualitative results are in the Appendix.

Table 3. **Ablation**. *Ego* outperforms uniform visual token selection and Full Visual under the same In Context tokens budget.

Method	% of Visual Tokens	Word Injection	F1 Score ↑
Keywords	✗	✓	71.3
Full Visual	100%	✗	<u>84.1</u>
Full Visual + Keywords	100%	✓	82.5
Uniform	20%	✗	77.7
Ego (1-view)	20%	✗	80.4
Ego (5-view)	20%	✗	**85.7**

5. Discussion

In this work, we aim to establish a strong and efficient post-hoc LVLM personalization method. Our approach assumes that the LVLM has robust visual understanding capabilities and may not perform well with older models. However, given the availability of powerful open-source alternatives, we do not anticipate the need to rely on less performant models. It is worth noting that full reliance on the LVLM requires an instruction prompt tailored to the specific model.

For evaluation, we prioritized fairness and reproducibility: we re-evaluated major personalization baselines on the same datasets under a unified set of expressive metrics. Our experiments cover most existing personalization datasets across diverse tasks—Recognition, VQA, and Captioning—as well as different personalization scenarios, including single-concept, multi-concept, and video.

Ego demonstrates a strong balance between efficiency and performance, outperforming state-of-the-art methods and baselines. Nevertheless, our results indicate significant room for improvement. We envision our evaluation protocol serving as a testbed for future personalization research.

References

[1] Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pages 73–91. Springer, 2024. 1, 3, 5, 6, 8

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6

[3] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550, 2024. 5

[4] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3

[5] Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. True multimodal in-context learning needs attention to the visual context. In *Second Conference on Language Modeling*, 2025. 2, 5

[6] 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*, pages 558–577. Springer, 2022. 2

[7] 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. 2, 3, 5, 6, 7

[8] Vaggelis Dorovatas, Soroush Seifi, Gunshi Gupta, and Rahaf Aljundi. Recurrent attention-based token selection for efficient streaming video-llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[9] Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. VisiPruner: Decoding discontinuous cross-modal dynamics for efficient multimodal LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18896–18913, Suzhou, China, 2025. Association for Computational Linguistics. 4

[10] 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. 2

[11] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Rap: Retrieval-augmented personalization for multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14538–14548, 2025. 1, 2, 3, 5, 6, 7

[12] Xiaohu Huang, Hao Zhou, and Kai Han. LLM-VTP: LLM-reasoned visual token pruning for efficient multi-modal video understanding, 2025. 3

[13] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 4

[14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 3

[15] Zhiqi Kang, Rahaf Aljundi, Vaggelis Dorovatas, and Karteek Alahari. Online in-context distillation for low-resource vision language models. *arXiv preprint arXiv:2510.18117*, 2025. 2, 6

[16] Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision-language models. *arXiv preprint arXiv:2509.17588*, 2025. 4

[17] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. Large language models in law: A survey. *AI Open*, 5:181–196, 2024. 2

[18] 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*, pages 19730–19742. PMLR, 2023. 1

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[21] 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*, pages 26296–26306, 2024. 6

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 5

[23] Ben Mann, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1(3):3, 2020. 1

[24] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo'llava: Your personalized language and vision assistant. *Advances in Neural Information Processing Systems*, 37:40913–40951, 2024. 1, 3, 5, 6

[25] Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*, 2024. 2, 3

[26] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*, 2024. 1, 2, 3

[27] 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*, pages 8748–8763. PmLR, 2021. 1, 3

[28] 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*, pages 22500–22510, 2023. 2

[29] 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*, pages 6527–6536, 2024. 2

[30] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023. 1

[31] Soroush Seifi, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Personalization toolkit: Training free personalization of large vision language models. *arXiv preprint arXiv:2502.02452*, 2025. 2, 3, 5, 6

[32] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023. 1

[33] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023. 1

[34] Likang Wu, Zhaopeng Qiu, Zhi Zheng, Hengshu Zhu, and Enhong Chen. Exploring large language model for graph data understanding in online job recommendations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 9178–9186, 2024. 2

[35] 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*, pages 19123–19132, 2023. 6, 8

[36] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024. 2

[37] 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. 6

[38] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3, 6