Personalize Your Large Vision-language Models with In-context Prompt Tuning

Yanshu Li[1], Jiaqian Li[1], Kuai Yu[2], Xi Xiao[3], Dongfang Liu[4], Tianyang Wang[3], and Ruixiang Tang[5]⋆

[1] Brown University, USA
yanshu_li1@brown.edu
[2] Columbia University, USA
[3] University of Alabama at Birmingham, USA
[4] Purdue University, USA
[5] Rutgers University, USA
ruixiang.tang@rutgers.edu

Abstract. Large vision-language models (LVLMs) have demonstrated strong general multimodal capability and are increasingly deployed in downstream systems. This trend has driven growing interest in LVLM personalization, which aims to enable models to quickly and effectively learn out-of-distribution multimodal concepts to meet user-specific needs. However, many existing methods rely on inference-time training, which reduces efficiency. They also struggle to maintain accuracy in complex multi-image, multi-concept settings. These limitations restrict the broader deployment of LVLM-based systems. Therefore, this paper proposes in-context prompt tuning (ICPT). Specifically, ICPT employs a lightweight projection module capable of operating in complex scenarios to extract fine-grained visual semantics from multiple reference images, seamlessly transforming these features alongside identity-label mappings into continuous prompts. To maximize computational efficiency, this module adaptively determines the prompt length based on the intrinsic visual complexity of each concept. Crucially, to overcome the environmental biases and cross-concept interference prevalent in real-world applications, we introduce two novel geometric regularizations. These constraints refine prompt representations by decoupling key identities from transient environmental states and separating concepts to avoid semantic confusion. Extensive experiments show that ICPT achieves state-of-the-art personalization accuracy across diverse tasks and LVLM backbones.

Keywords: Large Vision-language Models · Personalization

1 Introduction

A large vision-language model (LVLM) integrates a vision encoder with a large language model (LLM) backbone and is trained using staged learning procedures

⋆ Corresponding author.

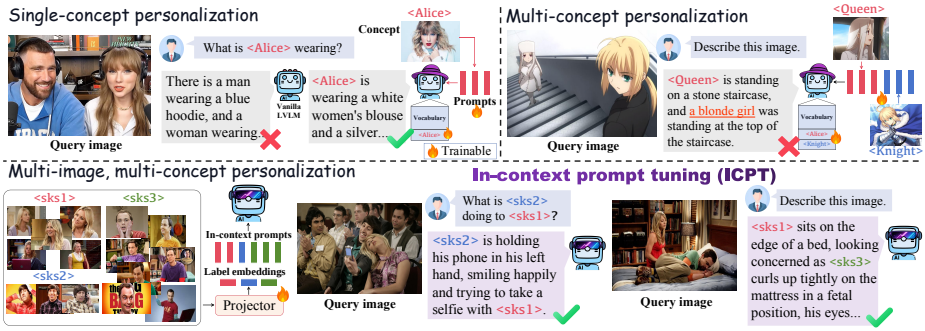

Fig. 1: Overview of LVLM personalization. Existing methods often rely on vocabulary expansion with inference-time training and struggle in multi-image and multi-concept settings. Our proposed ICPT improves efficiency and performance in complex scenarios.

[32]. By extending the capability boundary of LLMs to more real-world scenarios, LVLMs have become a core component in many modern AI systems, including interactive assistants and specialized agents [11, 12, 28].

As these systems, which are built upon general-purpose LVLMs, naturally shift toward private and user-facing deployments, ***personalization*** emerges as a critical research topic [19]. Enabling models to acquire user-specific concepts that go beyond their parametric knowledge is a key step toward fully unlocking the practical value of LVLM-based applications. As illustrated in Fig. 1, by learning the conceptual mapping between a personalized label <sks1> and the visual features of a specific individual given by a few reference images, an LVLM can subsequently leverage this concept directly to perform various tasks. However, achieving stable and efficient personalization in LVLMs remains challenging, as it requires learning both intra- and cross-modal mappings. A common solution uses in-context learning (ICL) by injecting explicit concepts into the prompt [15]. However, LVLMs are highly sensitive to few-shot prompts, which often leads to unstable concept acquisition [13]. Moreover, concept images are processed directly as visual inputs, generating many redundant tokens that reduce efficiency [14]. To address these challenges, two main lines of work have been explored. One of them focuses on post-training the entire model to improve its ICL capacity, such as PVIT [24]. However, the cost of building large-scale training datasets and performing model training further exacerbates efficiency concerns.

In contrast, soft prompt-based personalization becomes a dominant paradigm. Methods such as MyVLM [1] and Yo'LLaVA [21] learn a small set of prompts to encode user-specific concepts, enabling lightweight adaptation [2]. However, they reduce practical scalability, as each new concept requires vocabulary expansion and separate training. To avoid test-time training, PLVM [23] utilizes an additional pretrained vision encoder to extract features from a reference image and trains a cross-attention module to combine them with the concept label for prompt generation. While offering novel insights, PLVM and recent work are evaluated only on relatively simple tasks and do not address two complex

yet crucial scenarios: (1) **multi-concept** personalization, where the model must simultaneously learn and reason over multiple concepts spanning diverse categories beyond people, such as animals, objects, and scenes. (2) **multi-image** settings, where each concept is defined by multiple reference images exhibiting varying degrees of visual diversity, often requiring fine-grained and cross-image visual feature extraction. This motivates the following question: *How can we generate prompts that encode multimodal concepts with sufficient clarity for reliable reasoning in these complex settings, while avoiding inference-time training?*

In pursuit of this question, we propose in-context prompt tuning (ICPT), a personalization method that simulates ICL entirely within the representation space of frozen LVLMs. At the core of ICPT is the Adaptive Concept Projector (ACP), which extracts hierarchical multi-scale features directly from the built-in vision encoder to generate a dedicated textual label embedding alongside a continuous visual prompt for each novel concept. To balance representational capacity with inference efficiency, the ACP incorporates a Dynamic Token Router (DTR) that adaptively allocates variable prompt lengths based on the inherent visual complexity of the target concept. To guarantee pure and robust identity acquisition through these prompts, we propose two structural constraints: Contextual Variation Memory (CVM) and Margin-constrained Concept Separation (MCS). CVM maintains a memory queue of environment-induced variations, allowing learned prompts to be projected into a bounded, state-invariant subspace. Meanwhile, MCS applies a dual-modality soft-margin constraint on visual tokens and text anchors to separate identities while preserving shared semantic priors. Optimized end-to-end with carefully curated data and training objectives designed for the proposed modules and constraints, ICPT seamlessly integrates into various LVLMs and enables efficient and robust personalization at inference time. We evaluate ICPT on four common personalization tasks under both single-concept and multi-concept settings. Results across four LVLMs with different sizes and backbones demonstrate the consistently state-of-the-art personalization capability of ICPT compared to ICL and prior methods. Therefore, ICPT is an effective and practical method for future LVLM personalization.

The contributions of this paper are three-fold. **First**, we propose the Adaptive Concept Projector to extract, aggregate, and transform fine-grained information from multiple reference images into continuous prompts natively usable by LVLMs, introducing a dynamic token routing mechanism to adaptively allocate prompt length. **Second**, to optimize the semantic representation of these prompts for complex scenarios, we propose two innovative structural constraints alongside a targeted training strategy to optimize the entire framework end-to-end. **Third**, extensive experiments demonstrate that ICPT achieves superior performance and broad generalization across diverse multimodal tasks and LVLM backbones compared to existing methods.

2 Related Works

Large Vision-language Models (LVLMs). Most mainstream LVLMs consist of three core components: a vision encoder, a projector, and an LLM serving as the decoder [31]. Given multimodal inputs containing both images and text, the vision encoder, such as the Vision Transformer (ViT) used in CLIP [25], first partitions images into patches and converts them into visual tokens that encode perceptual features. These tokens are then merged and mapped by an MLP-based projector into the same latent space as textual tokens. They are then jointly processed by the LLM for unified multimodal reasoning [32]. With the progression from LLaVA1.5 [18] to more recent models like InternVL3.5 [29], and Qwen3VL [5], LVLMs have demonstrated steadily improving visual perception and understanding capabilities and are increasingly adopted as core components in a wide range of downstream applications [6, 22]. As a result, enabling effective personalization for LVLMs has become increasingly important [10].

Personalized LVLMs. Personalization aims to enable LVLMs to recognize and utilize user-specific concepts, supporting rapid semantic alignment in private and personalized scenarios [30]. In-context learning (ICL) is the most commonly used strategy, with methods such as LaMP [26] and RAP [7] leveraging retrieval augmented generation for personalization. However, in multimodal settings, ICL often exhibits instability. PVIT [24] addresses this issue by introducing additional post-training to improve the ICL capability of LVLMs. To achieve more lightweight personalization, MyVLM [1] and Yo'LLaVA [21] inject user-specific concepts through a small number of learnable soft prompts, avoiding long token-level inputs and improving stability. MC-LLaVA [2] further extends this paradigm to multi-concept scenarios. Nevertheless, these methods require retraining for each new concept due to vocabulary expansion. To reduce this overhead, PLVM [23] introduces an auxiliary pre-trained visual encoder to add new concepts without inference-time training, but its applicability is restricted to relatively simple task settings. These limitations motivate us to explore more scalable and flexible personalization methods. Meanwhile, studying training data recipes also emerges as a core topic in LVLM personalization [3, 4, 9].

3 Method

3.1 Overview

Problem setting. Each set $\mathcal{C} = \{C_1, \ldots, C_n\}$ contains n concepts, where each C_i consists of a label and a collection of reference images of varying sizes that exhibit its visual features. The concepts, namely the label–image mappings, are assumed to constitute novel knowledge for a pre-trained LVLM. Thus, given an LVLM θ, the goal of personalization is to enable the model to learn these concepts from scratch and leverage them across a variety of tasks. Although ICL aligns well with personalization, explicitly encoding all reference images in the prompt incurs high token costs and unstable performance.

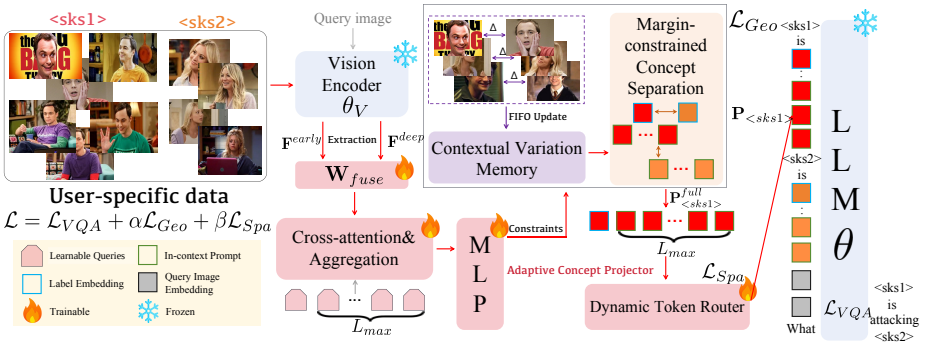

Fig. 2: The overall framework of ICPT. The components highlighted in pink constitute the core Adaptive Concept Projector.

Therefore, we propose in-context prompt tuning (ICPT), which simulates ICL in the latent space. Its pipeline is shown in Fig. 2. We first establish our real-time setting (Sec. 3.2). Next, we utilize a lightweight Adaptive Concept Projector and its dynamic token allocation mechanism (Sec. 3.3). We then introduce two strategies to enhance the multimodal semantic capture of in-context prompts: Contextual Variation Memory, which decouples transient environmental states from intrinsic identity (Sec. 3.4), and Margin-constrained Concept Separation, which mitigates cross-concept confusion while preserving shared semantic structure (Sec. 3.5). Finally, we describe the training strategy (Sec. 3.6).

3.2 Multi-concept In-context Prompts

For each concept C_i, an Adaptive Concept Projector (ACP) processes the reference images to generate a continuous in-context prompt $\mathbf{P}_i \in \mathbb{R}^{L_i \times d}$, where d is the hidden dimension of the LVLM. Unlike prior works that force a fixed prompt length for all concepts, the ACP then determines a variable token length $L_i \leq L_{max}$ based on the visual complexity of the specific concept. It simultaneously generates a token embedding $\mathbf{w}_i \in \mathbb{R}^{1 \times d}$ to represent the label of each concept. During personalization, to associate a generic label with the visual features of a concept, a two-concept LVLM input can be formatted as follows:

```
[System]: We define
<sks1> is w_1 : P_1
<sks2> is w_2 : P_2
[User]: (Query Image) What is <sks2> doing to <sks1>?
```

Here, the textual labels (`<sks1>`, `<sks2>`) are retained, enabling the frozen LVLM to output them in open-ended responses using its pre-trained vocabulary. $\mathbf{w}_i$ serves as the dedicated semantic anchor for each label, while the subsequent $\mathbf{P}_i$ provides visual evidence. Under this real-time setting, the challenge is ensuring that these in-context prompts and their anchors precisely capture the identities while maintaining disentanglement during multi-concept routing.

3.3 Adaptive Concept Projector (ACP)

In realistic settings, reference images of a concept are often not clean or well-isolated. Instead, they may contain visual distractions such as co-occurring objects and complex backgrounds. To address this issue, existing approaches either employ an auxiliary mask generator to crop concept-specific regions or introduce an additional vision encoder to extract embeddings dedicated to personalization.

To eliminate such external dependencies while improving efficiency, we directly leverage the built-in vision encoder θ_V of the LVLM. For each reference image, we feed it into θ_V and, inspired by [20], extract visual representations from its early and deep layers. By utilizing these hierarchical embeddings, we capture multi-level visual characteristics of objects within the image, enabling fine-grained concept identification without requiring external supervision. Specifically, let $I_{i,j}$ denote the j-th reference image of concept C_i, where $j \in \{1, \ldots, N_i\}$. We extract its patch-wise embeddings from the designated early and deep layers of θ_V, denoted respectively as $\mathbf{F}_{i,j}^{early}, \mathbf{F}_{i,j}^{deep} \in \mathbb{R}^{S \times d_v}$, where S is the sequence length of the patches and d_v is the feature dimension of the vision encoder. We fuse these multi-scale features via channel-wise concatenation followed by a linear projection weight $\mathbf{W}_{fuse}$:

$$\mathbf{F}_{i,j} = [\mathbf{F}_{i,j}^{early} \parallel \mathbf{F}_{i,j}^{deep}]\mathbf{W}_{fuse} \in \mathbb{R}^{S \times d}, \tag{1}$$

where $\parallel$ denotes the concatenation operator and $\mathbf{W}_{fuse} \in \mathbb{R}^{2d_v \times d}$ projects the fused features into the LVLM's hidden dimension d.

After obtaining the fused visual embeddings for each image, we employ the ACP, which consists of a single-layer cross-attention module followed by an MLP, to transform them into in-context visual prompts and their corresponding textual label embeddings. We initialize a fixed set of $L_{max} + 1$ learnable latent queries $\mathbf{Q} \in \mathbb{R}^{(L_{max}+1) \times d}$. The ACP distills the dense visual features into a fixed-capacity representation via cross-attention:

$$\mathbf{v}_{i,j} = \text{Softmax}\left(\frac{(\mathbf{Q}\mathbf{W}_q)(\mathbf{F}_{i,j}\mathbf{W}_k)^\top}{\sqrt{d}}\right)(\mathbf{F}_{i,j}\mathbf{W}_v) \in \mathbb{R}^{(L_{max}+1) \times d}, \tag{2}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The visual centroid across the N_i reference images is aggregated as $\bar{\mathbf{v}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{v}_{i,j}$. This centroid is then processed by the MLP to generate a joint representation $[\mathbf{w}_i; \mathbf{P}_i^{full}] = \text{MLP}(\bar{\mathbf{v}}_i)$, where the first token $\mathbf{w}_i \in \mathbb{R}^{1 \times d}$ serves as the dedicated word embedding for the text label, and the remaining L_{max} tokens form the intermediate full-capacity visual prompt $\mathbf{P}_i^{full} \in \mathbb{R}^{L_{max} \times d}$.

As different concepts vary in visual complexity, we aim to allocate prompts dynamically to balance accuracy and efficiency. To this end, the ACP incorporates a Dynamic Token Router (DTR) applied exclusively to the visual prompt. The DTR utilizes a lightweight linear routing head and a Sigmoid activation σ to assign a continuous score $s_{i,l}$ to each of the L_{max} visual tokens:

$$s_{i,l} = \sigma(\mathbf{p}_{i,l}^{full}\mathbf{w}_{router} + b_{router}) \in (0, 1), \quad \text{for } l \in \{1, \ldots, L_{max}\}, \tag{3}$$

where $\mathbf{p}_{i,l}^{full} \in \mathbb{R}^{1\times d}$ is the l-th token of $\mathbf{P}_i^{full}$, $\mathbf{w}_{router} \in \mathbb{R}^{d\times 1}$, and b_{router} is a bias term. During training, to preserve end-to-end differentiability, we apply soft-masking via element-wise multiplication: $\mathbf{P}_i = \mathbf{s}_i \odot \mathbf{P}_i^{full}$, where $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,L_{max}}]^\top$ and $\odot$ denotes row-wise broadcasting multiplication. During inference, we dynamically prune redundant visual tokens by retaining only the L_i tokens satisfying $s_{i,l} > \tau$, guaranteeing $L_i \geq 1$. The discrete label embedding $\mathbf{w}_i$, acting as the global text anchor, is preserved without pruning. This ensures each concept occupies an optimal, minimal footprint in the LVLM context.

In summary, the ACP defines the prompt-generation process. It takes C_i's hierarchical visual features as input and produces a full-capacity visual prompt $\mathbf{P}_i^{full}$ and an embedding of the concept label $\mathbf{w}_i$. The visual prompt is then dynamically routed into a soft-masked prompt $\mathbf{P}_i$, which is incorporated into the LVLM prompt alongside $\mathbf{w}_i$, introduced in Sec. 3.2. However, to ensure the representations capture pure, disentangled concepts, we regularize them during training via two geometric constraints, detailed next.

3.4 Contextual Variation Memory

Although the generated full-capacity prompt $\mathbf{P}_i^{full}$ encodes the target identity, it also retains biases from environmental factors such as background and lighting. If unaddressed, the LVLM may overfit to these cues rather than the true identity. In multi-concept settings, estimating such noise from intra-concept differences is unreliable because the aggregated centroid is itself biased by visual context.

To isolate and filter this noise without identity leakage, we construct a **Contextual Variation Memory** (CVM), denoted $\mathcal{M}_{ext}$, dynamically maintained as a First-In-First-Out (FIFO) queue with a fixed capacity limit K. During training, we concurrently perform two operations: populating this memory with environmental variations and utilizing it to regularize the prompts of the target concepts. To populate $\mathcal{M}_{ext}$ without introducing significant additional overhead, we reuse the representations already computed during the ACP's forward pass. For each concept C_i in the current batch with $N_i \geq 2$ reference images, we randomly sample one pair of distinct indices (x, y). To extract a sharp environmental transition without the blurring effect of multi-image averaging, we bypass centroid aggregation and independently pass $\mathbf{v}_{i,x}$ and $\mathbf{v}_{i,y}$ through the shared MLP to obtain their single-image full-capacity visual prompts, $\mathbf{P}_{i,x}^{full}$ and $\mathbf{P}_{i,y}^{full}$. In practical personalization scenarios, the provided reference images are typically selected to preserve stable or correlated identity components across different views. As a result, these identity components are expected to largely offset each other in the visual latent space. Thus, their Euclidean difference yields a directional matrix that captures environment-induced variations:

$$\boldsymbol{\Delta}_{i,x,y} = \mathbf{P}_{i,x}^{full} - \mathbf{P}_{i,y}^{full} \in \mathbb{R}^{L_{max}\times d}. \tag{4}$$

We then normalize this matrix to obtain $\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}_{i,x,y}/\|\boldsymbol{\Delta}_{i,x,y}\|_F$ and push it into $\mathcal{M}_{ext}$. Once the memory queue reaches its capacity limit K, the oldest

transitions are dequeued as new ones are added. Meanwhile, we utilize this updated memory to regularize the multi-image target concepts being optimized. For the aggregated visual prompt $\mathbf{P}_i^{full}$ of a target concept, we normalize it onto the unit hypersphere ($\hat{\mathbf{P}}_i^{full} = \mathbf{P}_i^{full}/\|\mathbf{P}_i^{full}\|_F$) and explicitly penalize its projection onto the variation matrices currently stored in $\mathcal{M}_{ext}$. Because both operands are Frobenius-normalized, their Frobenius inner product is mathematically equivalent to their cosine similarity. This constraint encourages the prompt to suppress components aligned with dominant contextual variation directions:

$$\langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{d}} \rangle_F = 0, \quad \forall \hat{\mathbf{d}} \in \mathcal{M}_{ext}. \tag{5}$$

By satisfying this constraint across all discrete matrices in $\mathcal{M}_{ext}$, we geometrically drive the learned prompt into the orthogonal null space of the continuous, identity-agnostic subspace spanned by the memory queue. By equipping the matrix space $\mathbb{R}^{L_{max} \times d}$ with the Frobenius inner product, we formally denote this contextual variation subspace as $\hat{\mathcal{V}}_{ext} = \text{span}(\mathcal{M}_{ext})$.

This regularization provides an explicit decomposition and bound on state-induced interference. Consider the LVLM evaluating a novel query image at test time. Let $\hat{\mathbf{q}}$ denote the Frobenius-normalized representation of a concept observed under a novel condition, and let $\hat{\mathbf{v}}_{id}$ denote the ideal, identity-dominant reference representation of that same concept. The deviation caused by the novel environment is defined as $\boldsymbol{\delta}_{state} \triangleq \hat{\mathbf{q}} - \hat{\mathbf{v}}_{id}$. We can orthogonally decompose this state-induced shift with respect to our learned subspace $\hat{\mathcal{V}}_{ext}$:

$$\boldsymbol{\delta}_{state} = \boldsymbol{\delta}_{state}^{\parallel} + \boldsymbol{\delta}_{state}^{\perp}, \quad \boldsymbol{\delta}_{state}^{\parallel} \in \hat{\mathcal{V}}_{ext}, \quad \boldsymbol{\delta}_{state}^{\perp} \perp \hat{\mathcal{V}}_{ext}. \tag{6}$$

Because $\hat{\mathcal{V}}_{ext}$ captures a broad spectrum of environmental shifts, $\boldsymbol{\delta}_{state}^{\parallel}$ represents the predictable environmental interference, while $\boldsymbol{\delta}_{state}^{\perp}$ is the unpredictable, out-of-span residual. The following theorem formally proves that enforcing Eq. 5 during training neutralizes this interference during inference.

Theorem 1 (State-Induced Interference Bound). *If the learned prompt satisfies the orthogonality constraint $\langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{d}} \rangle_F = 0$ for all $\hat{\mathbf{d}} \in \mathcal{M}_{ext}$, then the Frobenius inner product (which is equivalent to cosine similarity for Frobenius-normalized matrices) between the prompt and the novel query decomposes as*

$$\langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{q}} \rangle_F = \langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{v}}_{id} \rangle_F + \langle \hat{\mathbf{P}}_i^{full}, \boldsymbol{\delta}_{state}^{\perp} \rangle_F, \tag{7}$$

and the contribution of state-induced deviation is bounded by

$$\left| \langle \hat{\mathbf{P}}_i^{full}, \boldsymbol{\delta}_{state}^{\perp} \rangle_F \right| \leq \left\| \boldsymbol{\delta}_{state}^{\perp} \right\|_F. \tag{8}$$

Remark 1. Theorem 1's proof is provided in the Appendix. It establishes that enforcing orthogonality to the subspace $\hat{\mathcal{V}}_{ext}$ attenuates interference aligned with the stored variation directions. As $\mathcal{M}_{ext}$ is continuously updated with diverse, Frobenius-normalized pairwise transitions during training, it provides a progressively richer empirical reference frame for common environment-induced changes.

3.5 Margin-constrained Concept Separation

When multiple concepts populate the LVLM's visual dictionary simultaneously, cross-concept interference can manifest in two distinct pathways: *visual-space collision* between the candidate visual prompts ($\mathbf{P}_i^{full}$ and $\mathbf{P}_j^{full}$) during cross-attention, and *textual-space collision* between the generated label embeddings ($\mathbf{w}_i$ and $\mathbf{w}_j$) during self-attention. Enforcing global orthogonality between two concepts by driving their similarity toward zero can induce catastrophic feature loss. For example, if two concepts are both specific dog breeds, forcing their embeddings to be orthogonal harms their shared semantic priors, collapsing the representations out-of-distribution.

To address this issue across modalities, we introduce **Margin-constrained Concept Separation** (MCS). Rather than driving the cross-concept Frobenius inner product (which equates to cosine similarity for Frobenius-normalized matrices) to zero, we implement this constraint using a soft margin boundary $m \in (0, 1)$ and enforce it dually on both the continuous visual prompts and the discrete label embeddings:

$$\langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{P}}_j^{full} \rangle_F \leq m, \quad \text{and} \quad \langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j \rangle \leq m, \quad \forall i \neq j, \tag{9}$$

where $\hat{\mathbf{w}}_i = \mathbf{w}_i / \|\mathbf{w}_i\|_2$ denotes the L_2-normalized textual label embedding, and the vector inner product corresponds to their cosine similarity.

This bounded projection allows the representations to dynamically preserve shared features up to the similarity threshold m, while making unique identity embeddings remain separated enough to prevent confusion in both textual and visual spaces. By imposing these pre-DTR constraints, we ensure that the in-context prompts are less affected by visual noise and cross-concept interference.

3.6 Training

Training Data. To equip ICPT with robust multi-image, multi-concept personalization capabilities, we construct a dedicated, high-quality training dataset due to the lack of suitable existing resources. We first collect 350 concepts from four sources: three existing training sets (MyVLM, Yo'LLaVA, and MC-LLaVA) and additional concepts curated from public media repositories. Each concept is assigned a generic label that is independent of the LVLM's prior knowledge. The selected concepts cover diverse categories, including people, animals, objects, and scenes. For each concept, we prepare a pool of 10 reference images that provide informative visual cues. These images are either directly collected real-world captures or generated using image editing by Nano Banana2[1], allowing us to explicitly introduce controlled variations in appearance, context, and background within each concept. During training, the number of reference images sampled per concept ranges from 1 to 6 and is non-uniformly distributed.

Based on these concepts, we construct diverse query images and user queries. In total, we collect 2000 query images, each containing between 0 and 5 concepts.

[1] https://gemini.google/overview/image-generation/

To ensure robust reasoning and explicitly penalize text-prior hallucinations, the data distribution is carefully balanced: single-concept settings account for 35% of the data, multi-concept cases comprise 60%. For each query image, we utilize Gemini3.1-Pro[2] to generate 10 candidate user queries spanning a wide range of multimodal tasks, including existence recognition, multiple-choice visual question answering (VQA), and MSCOCO-style [16] captioning. Finally, we conduct two rounds of manual filtering to retain the five highest-quality user queries per image for training. Detailed statistics are provided in the Appendix.

Training Objective. To optimize the ACP while keeping the LVLM backbone frozen, we adopt a three-term multi-task objective:

$$\mathcal{L} = \mathcal{L}_{VQA} + \alpha \mathcal{L}_{Geo} + \beta \mathcal{L}_{Spa}. \tag{10}$$

Here, $\mathcal{L}_{VQA}$ represents the standard cross-entropy loss over the generated answer tokens y_t of length T. Given the user query X_{query}, query image I_{query}, and the set of active concepts $\mathcal{C}$, the VQA loss is formulated as:

$$\mathcal{L}_{VQA} = -\frac{1}{T} \sum_{t=1}^{T} \log P(y_t | y_{<t}, X_{query}, I_{query}, \{\mathbf{w}_i, \mathbf{P}_i\}_{i \in \mathcal{C}}). \tag{11}$$

Meanwhile, $\mathcal{L}_{Geo}$ geometrically unifies the constraints introduced in Sec. 3.4 and Sec. 3.5 into a learning objective. For a batch containing active concepts $\mathcal{C}$ and a sampled subset of normalized variation matrices $\hat{\mathcal{M}}_b$, we define:

$$\mathcal{L}_{Geo} = \sum_{i \in \mathcal{C}} \left(\frac{1}{|\hat{\mathcal{M}}_b|} \sum_{\hat{\mathbf{\Delta}} \in \hat{\mathcal{M}}_b} \langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{\Delta}} \rangle_F^2 + \frac{1}{|\mathcal{C}|-1} \sum_{\substack{j \in \mathcal{C} \\ j \neq i}} \left[\max\left(0, \langle \hat{\mathbf{P}}_i^{full}, \hat{\mathbf{P}}_j^{full} \rangle_F - m\right)^2 + \max\left(0, \langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j \rangle - m\right)^2 \right] \right). \tag{12}$$

Finally, $\mathcal{L}_{Spa}$ applies an L_1 penalty to the DTR's predicted scores, actively encouraging the ACP to drop redundant visual tokens and minimize the expected length of in-context prompts:

$$\mathcal{L}_{Spa} = \frac{1}{|\mathcal{C}| \cdot L_{max}} \sum_{i \in \mathcal{C}} \sum_{l=1}^{L_{max}} s_{i,l}. \tag{13}$$

4 Experiment

4.1 Experimental Setup

Evaluation. Existing LVLM personalization benchmarks generally fail to satisfy the two evaluation requirements of our setting: support for multi-image, multi-concept personalization, and an out-of-distribution test set with respect to the training data. Therefore, to maintain consistency with existing protocols, we adapt data from prior benchmarks and follow their construction principles to

[2] https://deepmind.google/models/gemini/pro/

enable effective evaluation. We collect concepts from MyVLM, Yo'LLaVA, MC-LLaVA, MMPB [9], and additional concepts curated and annotated from public media, resulting in a total of 200 concepts. All selected concepts are verified to be absent from the training set. For each concept, we obtain 1 to 6 reference images through collection or synthesis, with the number of images uniformly distributed across concepts. At test time, all reference images for each concept are used for personalization. These concepts correspond to 300 collected query images, each containing between 0 and 6 concepts, where 6 represents an out-of-distribution setting. We then construct 10 high-quality user queries for each query image by combining existing annotations with those generated by Gemini3.1-Pro. These queries cover the three task types used in training and a novel open-ended VQA task. We report performance separately for each task type under single-concept and multi-concept settings, comprising 1050 and 2100 cases, respectively. Among them, 150 cases correspond to open-ended VQA without any query image, where the LVLM answers solely based on the user query.

Models and baselines. We conduct experiments on LLaVA-NeXT-7B [17], LLaVA-NeXT-34B, InternVL3-8B [33], and Qwen3VL-8B [5]. Results on LLaVA-NeXT-7B serve as the primary comparison. We benchmark against multiple personalization baselines, including ICL, MyVLM [1], Yo'LLaVA [21], PVIT [24], RAP [7], PeKit [27], PLVM [23], and MC-LLaVA [2]. We also include GPT-4o with ICL as a closed-source baseline. On the other three LVLMs, we compare ICPT with ICL and PLVM to demonstrate its generalization across different LVLM backbones. For closed-source baselines, we reproduce their pipelines to the extent possible and report the best scores. For baselines originally designed only for single-concept personalization, we first report their results under the original setting. We then adapt them to the multi-concept settings strictly following their official implementations and report their best scores. If a method cannot be properly adapted, we mark its entries as "-". For methods that expand LVLM's vocabulary, we retrain them using *in-distribution* data to ensure a fair comparison.

Implementation details. The MLP used in the ACP has three linear layers, where the hidden dimension is set to twice the embedding dimension of the LVLM, and GeLU is used as the activation function. During training, only the parameters of the ACP and its DTR module are updated, while the LVLM backbone remains frozen. We use AdamW as the optimizer with a learning rate of 1e-4 and a batch size of 8. Across different LVLMs, we fix the full capacity of each in-context prompt to $L_{\max} = 20$ and the capacity limit of the CVM to $K = 128$, set the DTR decision threshold to $\tau = 0.5$, and use a soft margin of $m = 0.15$. For visual feature extraction, we consistently take representations from the third layer and the third-to-last layer of the LVLM vision encoder. The coefficients α and β are selected in a model-specific manner. For example, on LLaVA-NeXT-7B, we set $\alpha = 0.3$ and $\beta = 0.5$. All experiments are conducted on two NVIDIA H200 GPUs. All additional details of Sec. 4.1 are provided in the Appendix.

Table 1: Performance across four personalization task types under two settings, and their weighted average, on LLaVA-NeXT-7B. For existence recognition, we report recall; for multiple-choice VQA (MVQA), accuracy; for open-ended VQA (OVQA), BLEU; and for captioning, we report the average of concept recall in the generated captions and their embedding similarity to the ground-truth captions using DeBERTa-v3-large [8].

Method	Recognition			MVQA			OVQA			Captioning		
	Single	Multi	Weight	Single	Multi	Weight	Single	Multi	Weight	Single	Multi	Weight
GPT-4o + ICL	0.726	0.653	0.702	0.695	0.651	0.661	0.663	0.605	0.619	0.674	0.608	0.630
Vanilla	0.500	0.500	0.500	0.565	0.370	0.413	0.367	0.325	0.335	0.208	0.159	0.175
ICL	0.632	0.560	0.608	0.705	0.609	0.630	0.572	0.483	0.504	0.289	0.137	0.188
MyVLM	0.718	-	-	0.645	-	-	0.548	-	-	0.493	-	-
Yo' LLaVA	0.709	0.640	0.686	0.680	0.597	0.615	0.593	0.451	0.485	0.506	0.385	0.425
PVIT	0.752	0.633	0.712	0.695	0.647	0.658	0.642	0.537	0.562	0.623	0.467	0.519
RAP	0.725	0.621	0.690	0.710	0.627	0.645	0.651	0.529	0.558	0.580	0.493	0.522
PeKit	0.684	0.647	0.672	0.700	0.657	0.667	0.680	0.524	0.561	0.658	0.486	0.543
PLVM	0.794	0.719	0.769	0.760	0.673	0.692	0.695	0.607	0.628	0.664	0.507	0.559
MC-LLaVA	0.824	0.785	0.811	0.815	0.709	0.733	0.689	0.630	0.644	0.686	0.538	0.587
ICPT (Ours)	**0.873**	**0.859**	**0.868**	**0.850**	**0.774**	**0.791**	**0.716**	**0.697**	**0.702**	**0.713**	**0.624**	**0.654**

4.2 Main Results

ICPT demonstrates strong performance across different personalization tasks and settings. As summarized in Table 1, ICPT achieves state-of-the-art results across all settings for the four task categories, outperforming existing personalization baselines and GPT-4o with ICL. Compared with directly inserting concepts through ICL, ICPT improves both efficiency and effectiveness while using fewer tokens. On the existence recognition task, ICPT surpasses the prior SOTA baseline MC-LLaVA by 0.057 in weighted score, while using only **13.6** tokens per concept on average, compared with the fixed 16 tokens used by MC-LLaVA. ICPT also shows larger gains in more complex scenarios. For example, on open-ended VQA, it improves over MC-LLaVA by 0.027 in the single-image setting and by 0.067 in the multi-image setting. These results confirm that our design effectively addresses the challenges of multi-image and multi-concept personalization. In addition, the strong performance on open-ended VQA and captioning indicates that introducing in-context prompts does not impair the generative ability of the LVLM, allowing ICPT to integrate naturally into existing LVLM-based systems. Qualitative examples are shown in Fig. 3.

ICPT exhibits promising generalization across different LVLM sizes and backbones. Table 2 summarizes the performance of ICPT on three additional LVLMs, where it consistently achieves SOTA results across all settings. Results on LLaVA-NeXT-7B and LLaVA-NeXT-34B show that ICPT adapts well across model scales when $L_{\max}$ is properly configured. ICPT also performs strongly on InternVL3 and Qwen3VL, two recent LVLM architectures, demonstrating its robustness across different backbones. We further observe that ICPT continues to improve personalization performance even when the underlying

Table 2: Performance across four personalization task types under two settings when the personalization methods are applied to three additional LVLM backbones.

LVLM	Method	Recognition			MVQA			OVQA			Captioning		
		Single	Multi	Weight	Single	Multi	Weight	Single	Multi	Weight	Single	Multi	Weight
LLaVA-NeXT-34B	Vanilla	0.500	0.500	0.500	0.585	0.407	0.447	0.392	0.316	0.334	0.231	0.147	0.175
	ICL	0.695	0.584	0.658	0.715	0.646	0.661	0.636	0.573	0.588	0.328	0.215	0.253
	PLVM	0.812	0.721	0.782	0.780	0.682	0.704	0.723	0.668	0.681	0.740	0.545	0.610
	ICPT	**0.913**	**0.885**	**0.904**	**0.875**	**0.807**	**0.822**	**0.746**	**0.715**	**0.722**	**0.762**	**0.614**	**0.663**
InternVL3-8B	Vanilla	0.500	0.500	0.500	0.635	0.507	0.535	0.386	0.379	0.381	0.235	0.263	0.254
	ICL	0.738	0.705	0.727	0.825	0.702	0.729	0.726	0.618	0.644	0.575	0.538	0.550
	PLVM	0.795	0.730	0.773	0.820	0.695	0.723	0.734	0.597	0.630	0.806	0.762	0.777
	ICPT	**0.945**	**0.906**	**0.932**	**0.905**	**0.893**	**0.896**	**0.740**	**0.706**	**0.714**	**0.828**	**0.804**	**0.812**
Qwen3VL-8B	Vanilla	0.500	0.500	0.500	0.600	0.604	0.603	0.403	0.364	0.373	0.306	0.257	0.273
	ICL	0.775	0.743	0.764	0.865	0.826	0.835	0.758	0.642	0.670	0.635	0.572	0.593
	PLVM	0.820	0.735	0.792	0.825	0.799	0.805	0.760	0.668	0.690	0.784	0.708	0.733
	ICPT	**0.963**	**0.924**	**0.950**	**0.920**	**0.867**	**0.879**	**0.776**	**0.724**	**0.736**	**0.835**	**0.821**	**0.826**

LVLM already has strong ICL ability, which PLVM fails to maintain. On the two latest LVLMs, InternVL3 and Qwen3VL, PLVM performs worse than direct ICL on several tasks, whereas ICPT consistently provides stable gains. These results indicate that ICPT can effectively leverage advances in foundation LVLMs and highlight its practical value for building next-generation applications.

Beyond the above multi-image, multi-concept evaluation, we further evaluate ICPT on four standard public benchmarks to validate its effectiveness. Table 3 reports the average results across models on the YoLLaVA, MyVLM, MC-LLaVA, and MMPB. For the first three benchmarks, all concepts appearing in the ICPT training set are excluded from evaluation to ensure a fair comparison, as these benchmarks assume vocabulary expansion-based personalization. ICPT consistently achieves the best performance across all four benchmarks, demonstrating that its superiority extends beyond our customized evaluation setting.

Table 3: ICPT's performance on four standard multimodal personalization benchmarks.

	Yo'LLaVA	MyVLM	MC-LLaVA	MMPB
ICL	0.805	0.640	0.493	0.589
MC-L	0.921	0.964	0.904	0.824
PLVM	0.875	0.905	0.857	0.814
ICPT	**0.926**	**0.972**	**0.915**	**0.853**

4.3 Ablation Studies

We first examine the effectiveness of the two proposed prompt constraints and their corresponding learning objective. The results in Fig. 4 show that both CVM and MCS, together with $\mathcal{L}_{Geo}$, improve performance under both single-concept and multi-concept settings, with larger gains in the multi-concept case. This indicates that the proposed mechanisms are well optimized during training and effectively enhance the semantic quality of the in-context prompts.

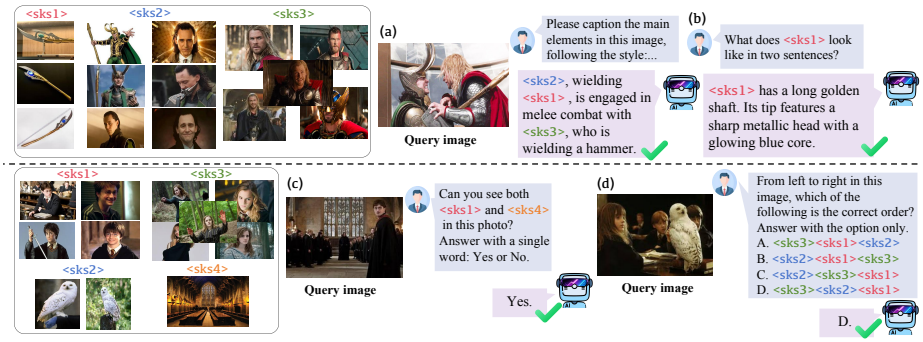

Fig. 3: Qualitative examples of ICPT on four tasks: (a) captioning, (b) open-ended VQA without a query image, (c) existence recognition, and (d) multiple VQA. More examples and failure case analyses are provided in the Appendix.

We then analyze the impact of the prompt full capacity $L_{\max}$ and the DTR mechanism. One ablation replaces DTR with random pruning at the same ratio. As shown in Fig. 5, personalization performance generally improves as the prompt length increases, since longer prompts can encode richer visual features. The improvement continues until a threshold. Within this range, the performance gain brought by DTR also becomes more pronounced. Beyond this threshold, however, longer prompts tend to encode redundant features or noise, which leads to performance degradation. In this regime, dynamic pruning becomes even more important. These results indicate that both proper prompt length selection and adaptive pruning are critical for soft prompt-based methods.

For MCS, we design three ablation settings. We apply MCS only to label embeddings or only to full-capacity prompts while keeping the similarity threshold at $m = 0.15$, and we further test the case where $m = 0$, which enforces strict orthogonality. Results in Fig. 6 show that applying MCS only to label embeddings causes a larger performance drop than applying it only to prompts. This suggests that multi-concept confusion mainly arises from overlap in the visual representation space. Nevertheless, applying constraints to both modalities remains important because LVLMs have limited ability to clearly separate them. When $m = 0$, the model overemphasizes separation and weakens the beneficial role of prior knowledge during personalization. This result highlights the importance of selecting an appropriate margin m. We provide more comprehensive ablation studies in the Appendix, covering the choice of vision encoder, the selection of feature extraction layers, the coefficients of each loss term, the DTR decision threshold τ, and the CVM capacity K.

4.4 Analyses

We conduct efficiency analyses of ICPT. With the prompt full capacity set to 20, ICPT achieves 12% and 18% lower inference latency on the full test set compared with PLVM and MC-LLaVA, which use fixed 16-token prompts for each concept.

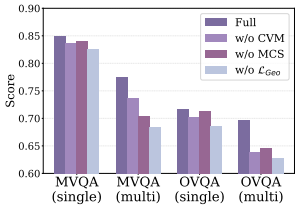

Fig. 4: Ablation study on the proposed constraints.

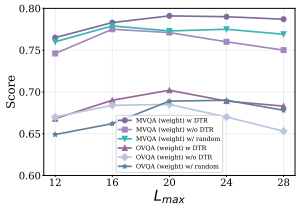

Fig. 5: Ablation study on the DTR mechanism.

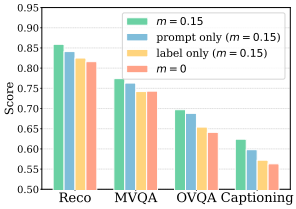

Fig. 6: Ablation study on the MCS setting.

This improvement results from using the model's native vision encoder, together with the lightweight design of ACP and the dynamic pruning enabled by DTR.

We also investigate an open question in LVLM personalization: whether training data volume or diversity is more important. We characterize training data diversity along three dimensions: the visual distribution of reference images per concept, the distribution of the number of concepts per input, and the variety of task types. The results show that a low-volume, high-diversity training setup substantially outperforms its high-volume counterpart. This indicates that exposing the model to complex visual contexts, varied concept combinations, and diverse task formats is far more important than simply increasing the number of collected concepts. As LVLM capabilities advance, effective personalization should move beyond brute-force data memorization and instead serve as a generalizable semantic bridge that aligns concepts with the LVLM's reasoning in the latent space. These analyses are presented in detail in the Appendix.

5 Conclusion

We propose in-context prompt tuning (ICPT), a novel and effective framework that advances the capability of LVLMs to handle complex, multi-concept personalization tasks. By fully shifting the personalization process into the latent space through simulated ICL, our method circumvents the heavy computational burden of inference-time training, allowing users to introduce multiple new concepts on the fly. Central to this capability is the Adaptive Concept Projector (ACP), which efficiently distills reference images into dynamic-length visual prompts and dedicated semantic anchors. Furthermore, to ensure these representations remain robust against background noise and cross-concept entanglement, we design two geometric constraints: Contextual Variation Memory and Margin-constrained Concept Separation. These designs are validated through extensive experiments. We envision that ICPT will provide a strong and practical baseline for future research on LVLM personalization.

References

1. Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., Cohen-Or, D.: Myvlm: Personalizing vlms for user-specific queries. In: European Conference on Computer

Vision. pp. 73–91. Springer (2024) 2, 4, 11
2. An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., et al.: Mc-llava: Multi-concept personalized vision-language model. arXiv preprint arXiv:2411.11706 (2024) 2, 4, 11
3. An, R., Yang, S., Zhang, R., Shen, Z., Lu, M., Dai, G., Liang, H., Guo, Z., Yan, S., Luo, Y., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. arXiv preprint arXiv:2505.14671 (2025) 4
4. An, R., Zeng, K., Lu, M., Yang, S., Zhang, R., Ji, H., Zhang, Q., Luo, Y., Liang, H., Zhang, W.: Concept-as-tree: Synthetic data is all you need for vlm personalization. In: Synthetic Data for Computer Vision Workshop@ CVPR 2025 (2025) 4
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 4, 11
6. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., Shou, M.Z.: Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv preprint arXiv:2306.08640 (2023) 4
7. Hao, H., Han, J., Li, C., Li, Y.F., Yue, X.: Rap: Retrieval-augmented personalization for multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14538–14548 (2025) 4, 11
8. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021) 12
9. Kim, J., Kim, W., Park, W., Do, J.: Mmpb: It's time for multi-modal personalization. arXiv preprint arXiv:2509.22820 (2025) 4, 11
10. Kim, J., Choi, J., Chay, W., Kyung, D., Kwon, Y., Jo, Y., Choi, E.: Propersim: Developing proactive and personalized ai assistants through user-assistant simulation. arXiv preprint arXiv:2509.21730 (2025) 4
11. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. Foundations and Trends in Computer Graphics and Vision **16**(1-2), 1–214 (2024) 2
12. Li, F., Han, S., Lee, C.H., Feng, S., Jiang, Z., Sun, Z.: A new era in human factors engineering: A survey of the applications and prospects of large multimodal models. International Journal of Human–Computer Interaction pp. 1–14 (2025) 2
13. Li, Y., Cao, Y., He, H., Cheng, Q., Fu, X., Xiao, X., Wang, T., Tang, R.: M²IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=9ffYcEiNw9 2
14. Li, Y., Yang, J., Shen, Z., Han, L., Xu, H., Tang, R.: Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6619–6627 (2026) 2
15. Li, Y., Yang, J., Yang, Z., Li, B., Han, L., He, H., Yao, Z., Chen, Y.V., Fei, S., Liu, D., et al.: Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6610–6618 (2026) 2

16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 10
17. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024) 11
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 4
19. Lyu, H., Jiang, S., Zeng, H., Xia, Y., Wang, Q., Zhang, S., Chen, R., Leung, C., Tang, J., Luo, J.: Llm-rec: Personalized recommendation via prompting large language models. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 583–612 (2024) 2
20. Meng, L., Yang, J., Tian, R., Dai, X., Wu, Z., Gao, J., Jiang, Y.G.: Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. Advances in Neural Information Processing Systems **37**, 23464–23487 (2024) 6
21. Nguyen, T., Liu, H., Li, Y., Cai, M., Ojha, U., Lee, Y.J.: Yo'llava: Your personalized language and vision assistant. Advances in Neural Information Processing Systems **37**, 40913–40951 (2024) 2, 4, 11
22. Park, S., Song, Y., Lee, S., Kim, J., Seo, J.: Leveraging multimodal llm for inspirational user interface search. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–22 (2025) 4
23. Pham, C., Phan, H., Doermann, D., Tian, Y.: Personalized large vision-language models. arXiv preprint arXiv:2412.17610 (2024) 2, 4, 11
24. Pi, R., Zhang, J., Han, T., Zhang, J., Pan, R., Zhang, T.: Personalized visual instruction tuning. arXiv preprint arXiv:2410.07113 (2024) 2, 4, 11
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 4
26. Salemi, A., Mysore, S., Bendersky, M., Zamani, H.: Lamp: When large language models meet personalization. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7370–7392 (2024) 4
27. Seifi, S., Dorovatas, V., Reino, D.O., Aljundi, R.: Personalization toolkit: Training free personalization of large vision language models. arXiv preprint arXiv:2502.02452 (2025) 11
28. Sun, Q., Liu, Z., Ma, C., Ding, Z., Xu, F., Yin, Z., Zhao, H., Wu, Z., Cheng, K., Liu, Z., et al.: Scienceboard: Evaluating multimodal autonomous agents in realistic scientific workflows. arXiv preprint arXiv:2505.19897 (2025) 2
29. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 4
30. Xu, B., Feng, J., Lu, S., Luo, Y., Yan, S., Liang, H., Lu, M., Zhang, W.: Jarvis: Towards personalized ai assistant via personal kv-cache retrieval. arXiv preprint arXiv:2510.22765 (2025) 4
31. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024) 4
32. Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., Yu, D.: Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601 (2024) 2, 4

33. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 11