Personalizing MLLMs via Reinforced Multimodal Reference Game

Deepayan Das[1], Davide Talon[2], Yiming Wang[2], Massimiliano Mancini[1], and Elisa Ricci[1,2]

[1] University of Trento, Trento, Italy
{deepayan.das, massimiliano.mancini, e.ricci}@unitn.it
[2] Fondazione Bruno Kessler, Trento, Italy
{dtalon,ywang}@fbk.eu

Abstract. Personalizing Multimodal Large Language Models (MLLMs) aims to recognize users' unique concepts from visual data and provide personalized responses. Although prior work has shown the benefit of concept descriptions and reasoning for this task, MLLM descriptions often include information, such as state and context, that does not help and may in fact hinder the unique identification of the target concept among other visually similar items. Effective descriptions of personal concepts should instead be accurate, discriminative, and free of distracting details. To achieve such descriptions, we introduce Reinforced Reference Game (RRG), a learning framework that promotes discriminative descriptions through a novel reinforced multimodal reference game. The MLLM plays both the roles of speaker and listener in a contrastive game setting, whose goal is to effectively communicate discriminative information about a target concept. Our approach formulates a verifiable contrastive reward over hard positives (dissimilar views of the same concept) and hard negatives (visually similar but different concepts). Empirically, RRG achieves state-of-the-art across multiple tasks on three personalization benchmarks. RRG generalizes to unseen domains and outperforms existing methods based on concept descriptions and personalization-specific RL frameworks. We will release code and models in the project page.

Keywords: Multimodal Large Language Models · Personalization

1 Introduction

Personalization [2, 9, 14, 24] aims to equip Multimodal Large Language Models (MLLMs) [1, 4, 19, 21, 41] with knowledge about user-defined concepts [2, 24]. In this way, an MLLM can answer questions, describe, or even refer to such concepts, providing outputs that are more engaging and aligned with user's expectations.

MLLM personalization entails several challenges. Consider the example in Fig. 1: how can we distinguish Amy's shirt from other shirts? One could argue that Amy's shirt can be recognized from its red floral patterns and creamy color. Note that these attributes (i) provide a unique, fine-grained identification of the

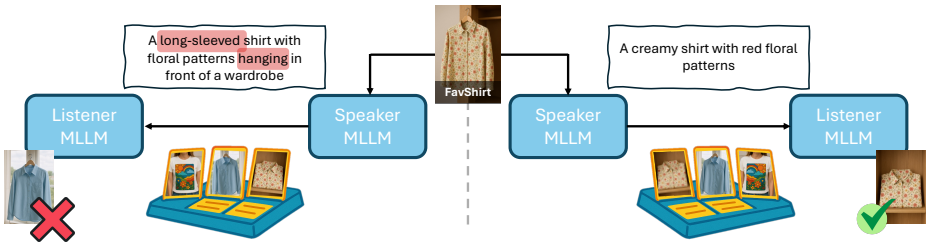


Fig. 1: The idea behind our RRG. We frame personalization as a reference game. The MLLM plays both the role of the *speaker* (describing a target concept) and the *listener* (identifying it among candidates), which encourages the speaker to generate distinctive, invariant descriptions that enable accurate identification beyond standard dense captioning. On the left a failing round, on the right a successful one.

item and (ii) the description does not have distracting information, *i.e.*, attributes are limited to the user-concept itself and do not change over time.

To equip MLLMs with this capability, early personalization approaches learn concept-specific representations at test-time [2, 24, 25]. However, these methods are computationally expensive and require retraining whenever new concepts are introduced. Recent works sidestep these limitations by either reasoning over reference images [14, 26, 34] or by introducing template-based textual descriptions [9]. Given a query image, such models retrieve the closest personal concepts in a database and reason over them directly during inference. While effective, these strategies either rely on the model's ability to identify which regions of the reference images are relevant or assume descriptions can accurately focus on the concept of interest. As a result, concept representations that contain irrelevant or mutable details can mislead the recognition of personal concepts.

Why do standard descriptions fall short? Consider Amy's personal long-sleeved t-shirt with floral patterns in Fig. 1. First, descriptions frequently mention the *concept state*: visual features may expose the current state of the concept, *e.g.*, "hanging", which is irrelevant for recognition and may confuse the model either when the state of the shirt changes or when other similar-looking shirts share the same state. Secondly, they include *context information* such as background or positional cues which can vary with viewpoint changes and thereby can hinder generalization. Finally, regular descriptions lack *comparative focus*, *i.e.*, they ignore what truly distinguishes the concept from semantically similar ones, for instance, the "floral pattern" that differentiates Amy's t-shirt from Joy's similar navy-blue one. As MLLMs are trained for generic captioning and VQA, they are ill-suited for the necessary fine-grained details needed for effective personalization.

Thus, a natural question arises: ***Can we teach an MLLM to communicate fine-grained yet essential concept features for personalization?*** To answer this question, we depart from previous efforts and treat personalization as a *reference game*, inspired by the classic "Guess Who?" game (Fig. 1). We introduce a novel MLLM learning paradigm, **R**einforced **R**eference **G**ame (RRG), to enhance personalized concept descriptions. Specifically, given an MLLM we

want to personalize, we formulate a "game" in which the model alternates between two roles: a *speaker*, which describes the target concept, and a *listener*, which must identify that concept from a set of candidate images. To succeed, the speaker must go beyond standard image captions and articulate only the distinctive and immutable features that allow the listener to select the correct concept. Crucially, we design the game to explicitly favor informative discriminative descriptions while discouraging distracting details, by exposing the listener to ambiguous candidate concepts that are difficult to distinguish. On the one hand, we present the speaker and the listener with two challenging views of the same concept: hard positives, which differ in background, viewpoint, or object state. On the other hand, we include distractor images composed of hard negatives, consisting of visually similar items from the same semantic class as the referenced concept. This hard contrastive setup provides a verifiable reward signal, enabling us to train the speaker to generate discriminative and immutable descriptions. We optimize the speaker via GRPO [13], which allows learning from verifiable game outcomes, thereby eliminating the need for explicit description supervision, which is non-trivial to obtain.

Results on established benchmarks Yo'LLaVA [24], MyVLM [2] and PerVA [9] demonstrate that our trained description model effectively improves personalization performance on captioning and VQA tasks, outperforming existing training-based methods. Notably, RRG outperforms alternative description-based approaches [9], and reinforcement learning (RL)-based ones [26], demonstrating both the benefits of the high-quality concept descriptions and the multimodal reference game as an RL training strategy.
Our main **contributions** are summarized below:

- We introduce a novel perspective on MLLM personalization, by formulating the task as a communication game, where success depends on communicating the essential, immutable attributes that uniquely identify a user-defined concept among visually similar candidates;
- Building on this formulation, we propose RRG, a new approach in which the MLLM alternates between speaker and listener roles in a multimodal reference game, learning to generate uniquely discriminative concept descriptions. The speaker is updated through reinforcement learning, where the reward is derived from game outcomes.
- RRG achieves state-of-the-art performance for the captioning task on three datasets and superior VQA accuracy, outperforming existing training-based methods and generalizing effectively to unseen concepts.

2 Related Work

Personalization with MLLMs. The problem of personalization has been explored across several vision tasks. In text-to-image generation, the idea is to bind target concepts to custom identifiers derived from a handful of reference images, either through methods such as textual inversion [12] or fine-tuning

[18, 32, 36]. For segmentation, recent methods establish correspondences between reference and query images to estimate the target concept's location [22, 33, 40, 43], which is then used to guide prompting of SAM [17]. Other solutions have drawn on diffusion models [33, 40], customized generative models [38], or cycle-consistency constraints [6]. In this study, we narrow the focus to personalization in MLLMs, aiming to enable recognition, visual question answering, and captioning with references to user-specific concepts.

Early contributions such as MyVLM [2] and Yo'LLaVA [24] adapt inversion-based strategies from generative modeling, assigning each object a distinctive latent code, represented either as a concept vector [2] or a pseudo-token [24, 25]. To reduce adaptation time, recent efforts incorporate large-scale tuning [5, 9, 14, 28, 29], improving performance at the cost of extensive pretraining or expensive pairwise-matching. Most similar to us, [9] leverages the model's implicit knowledge to describe via natural language the reference concept for later matching and [26] employs an RL-based post-training approach for improving MLLM recognition capabilities. However, (i) descriptions generated independently for each concept lack comparative information about visually similar alternatives and (ii) naive application of RL strategies may lead to shortcuts for concept recognition rather than producing a better concept characterization (see Sec. 4). Unlike prior methods, RRG reframes the problem as a communication game leveraging RL to learn descriptions that uniquely identify the target concept for personalization.

GRPO-based post-training methods for MLLMs. Reinforcement Learning has recently emerged as a powerful paradigm for post-training large language models (LLMs), allowing them to align better with human preferences and downstream objectives. Approaches such as Reinforcement Learning from Human Feedback (RLHF) and its gradient-based variants have demonstrated remarkable success in improving reasoning and factual consistency in large-scale LLMs [27, 37]. Among these, GRPO (Group Relative Policy Optimization) has become particularly popular due to its stability and efficiency in optimizing non-differentiable objectives during instruction tuning [35]. It enables large models to learn from relative ranking signals among multiple sampled generations, thus avoiding explicit human annotation of reward scores. Building upon this, methods such as Visual-RFT [23] and Vision-R1 [20] have shown that GRPO-guided optimization can significantly improve few-shot performance in vision-centric tasks such as object detection, classification, and visual reasoning. Most relevant to our work, RePIC [26] introduces RL-based post-training for personalizing MLLMs, demonstrating that GRPO can effectively enhance recognition of user-specific visual concepts. While RePIC focuses on optimizing the reasoning module to predict the correct personalized concept name for a given concept, our approach diverges in its objective. We employ GRPO to fine-tune the speaker model to generate rich, fine-grained, and state-invariant descriptions.

Learning via communication games. The principles behind multimodal reference games have been studied in various contexts. For instance, [10] showed how to build an agent that can guess which objects in images an oracle is referring to, akin to referring object segmentation [16]. Early works on visual dialog [8]

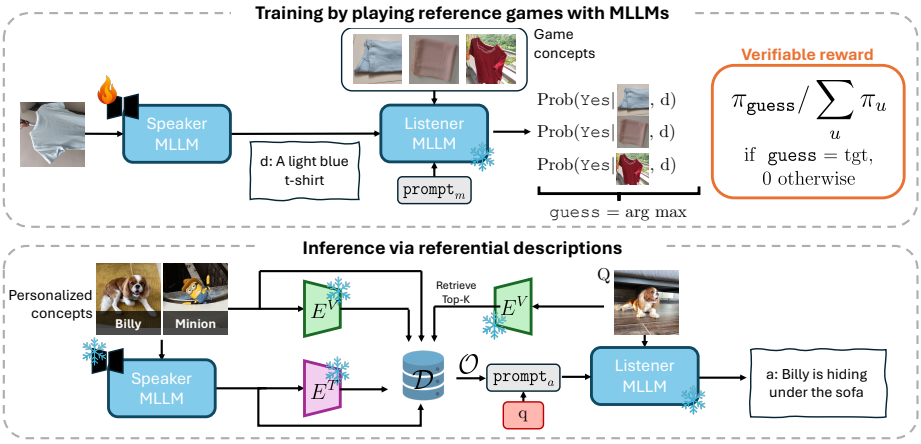


Fig. 2: Reinforced Reference Game (RRG). We revisit personalization through the lenses of a multimodal reference game, where an MLLM should learn how to describe concepts. (Top) During training when acting as a speaker, the MLLM receives as input an image of a concept, producing in output a description. When the MLLM acts as the listener, it receives as input the description from the speaker and a set of images: the goal of the game is to guess to which image the speaker is referring to, assigning the highest matching probability. Based on the outcome of the game, we compute a verifiable reward for RL, updating the speaker. (Bottom) During inference, the speaker outputs descriptions for each of the personalized concepts. These are used as reference for retrieving the personalized concept from a database and aiding MLLM reasoning.

studied interactions between a model and a human, answering questions given a specific image. Our work takes inspiration from [7], in which two agents play a guessing game: a speaker communicates an image to a listener using only natural language attributes. Following the theory of mind [30,39], the speaker should adapt to what the listener can understand, modifying its language accordingly. Notably, this game has been shown to be applicable for building interpretable classification models [3]. In this work, we neither focus on interpretable interactions, nor on full dialogues. Instead, we exploit the abstraction capabilities needed to solve the reference game as a tool for agents to learn how to generate descriptions of concepts tailored for personalization.

3 Reinforced Reference Game

We introduce **R**einforced **R**eference **G**ame, a new method that aims to obtain descriptions tailored for MLLM personalization. We first formalize the personalization problem (Sec. 3.1) and present how a reference game can be instantiated on MLLMs for improved personalization using online RL (Sec. 3.2). Finally, we describe how we exploit this newly learned capability during inference (Sec. 3.3).

3.1 Problem formulation

MLLM personalization aims to inject knowledge of specific personal concepts into a MLLM, in such a way that the model can recognize those personal items and answer queries about them. Formally, let us consider the MLLM Φ, as a generative framework, taking as input an image $v \in \mathcal{V}$ and a textual query $t \in \mathcal{T}$, and producing as output a textual response, *i.e.*, $\Phi : \mathcal{V} \times \mathcal{T} \to \mathcal{T}$. In the personalization task, the user provides a set $P = \{p_1, \cdots, p_n\}$ of n *personalized concepts*. Each concept $p \in P$ is a tuple $p = (V_p, t_p)$ composed of a set of reference images $V_p = \{v_p^1, \cdots, v_p^k\}$ and a corresponding name t_p, associated to that specific concept. It is important to note that, because these are personal concepts, the MLLM *has not* been previously exposed to any images of $p \in P$ or to the specific names used to reference them. Thus, the set P represents user-specific knowledge that the MLLM must acquire through personalization.

3.2 Playing reference games with MLLMs

In the personalization scenario of Sec. 3.1, our goal is to build a MLLM $\Phi_{\texttt{speak}}$ that produces descriptions useful for personalization. But, how should the speaker behave? First, since users may introduce multiple personal concepts over time, the speaker should generate *discriminative* captions *without* requiring access to all personalized concepts simultaneously. This ensures that the description of one concept is independent of the others, avoiding the need to update the entire set of descriptions whenever a new concept is added. Second, the speaker should *generalize* to unseen concepts so that it does not require re-training whenever new concepts arrive. However, supervising the speaker with unique captions poses annotation challenges, as descriptions depend on comparing concepts, a task that is also cognitively challenging for humans.

Multimodal reference games offer a principled framework to meet these requirements. We have two collaborative agents facing an imperfect communication channel: a *speaker* and a *listener*. Given an image containing a personal concept, the speaker produces a textual description that the listener must use, and based solely on this description, it should identify the concept the speaker is referring to. Formally, let $C = \{c_1, \cdots, c_n\}$ be the set of *game concepts*. Similarly to a personalized set P, to each concept $c \in C$, we associate a name t_c and a set of images V_c. Note that, C is a disjoint set of P (*i.e.*, $C \cap P = \emptyset$), as we want the MLLM to generalize its communication capabilities to *arbitrary*, unseen concepts. Starting from an image featuring the game concept c_{tgt}, the speaker describes the concept it observes, whose description is then used by the listener to identify c_{tgt} from a set of candidate images V. The game is *successful* if the listener guesses c_{tgt}, *i.e.*, the image associated to the target concept, and unsuccessful otherwise. Intuitively, successful descriptions re-identify the target concept by capturing features that are both *immutable*, *i.e.*, invariant across views and states, and *distinctive*, meaning that they tell the concept apart from similar ones.

Game setup. To personalize a given MLLM, we instantiate the game with both $\Phi_{\texttt{speak}}$ and $\Phi_{\texttt{list}}$ sharing the same backbone model. The speaker $\Phi_{\texttt{speak}}$ is further

adapted through LoRA [15] parameters θ, which steer it toward producing more discriminative and effective descriptions. A game round proceeds as follows: we begin from the game concepts set C and sample a target concept $c_{tgt} \in C$. From its associated image set V_{tgt}, one image v^i_{tgt} is provided to the speaker, while a different image v^j_{tgt}, with $i \neq j$, is given to the listener to be recognized.

Hard positives *vs.* Hard negatives. A key component in our game is how we populate the set of candidates V. Since a more challenging game encourages the speaker to be more precise, we construct V using hard examples. On the one hand, hard positives reflect challenging views of the target concept c_{tgt}: variations from v^i_{tgt} to v^j_{tgt} may involve changes in illumination, background, or state of the concept (*e.g.*, a blouse being folded or hung). These examples encourage the model to focus on concept-invariant cues while disregarding contextual and appearance variations. On the other hand, hard negatives acting as distractors need to be discriminated based on their fine-grained, instance-specific attributes, ensuring that the distinction cannot rely on coarse category-level features. To this end, we rely on U concepts from the same semantic class as c_{tgt} and select one image per concept as distractors. For the nature of hard negatives and positives, we identify PerVA dataset [9] as the most suitable to support our training among other personalization datasets, as it features (i) images of different concepts sharing the same semantic class, ideal for hard negatives, and (ii) diverse views of the same concept with significant appearance changes, ideal for constructing hard positives. The game round considers recognizing c_{tgt} among candidates $V = \{v^j_{tgt}\} \cup \bar{V}_{tgt}$, with $\bar{V}_{tgt}$ denoting the image set associated to hard negatives.

Game start. Given this setup, the speaker produces a description $d = \Phi_{\texttt{speak}}(v^i_{tgt})$ for the target image, and the listener must determine which image in V the description d refers to. As MLLMs struggle with multi-image inputs [11], we treat the problem as one with binary output.

More precisely, for each of the candidates $v_u \in V$ the listener evaluates:

$$\rho_u = \texttt{Prob}_{\Phi_{\texttt{list}}}(\texttt{Yes} \mid v_u, d, \texttt{prompt}_m), \tag{1}$$

where $\texttt{prompt}_m$ denotes the prompt used to probe the description-candidate match: **‘‘Does this description match the provided image? Answer Yes or No’’**. In other words, Eq. (1) computes the probability that description d refers to the concept v_u as evaluated by the probability of the `Yes` token.

The listener's final guess is obtained by gathering the information among the candidates, *i.e.*, $\texttt{guess} = \arg\max_u \rho_u$.

Speaker training with contrastive verifiable reward. We rely on GRPO [13], an online reinforcement learning algorithm to update parameters based on the relative preferences within a set of rollouts, while minimizing the KL divergence between the updated and the reference policy π_{ref}. The `guess` outcome of the game provides a verifiable reward signal, defined as:

$$r_{\text{match}} = \begin{cases} \rho_{\texttt{guess}} / \sum_u \rho_u, & \text{if } \texttt{guess} = tgt, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where the reward is given by the normalized probabilities of candidates, thus accounting for model uncertainty on its guess. See *Supp. Mat.* for more details.

Hence, the GRPO objective $\mathcal{R}(\theta)$ is:

$$\mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S}\frac{1}{|d_s|}\sum_{t=1}^{|d_s|}R_t^s(d_s,\theta)-\beta D_{KL}\left(\pi_\theta \,||\, \pi_{\text{ref}}\right)\right], \tag{3}$$

where the average is on S rollouts $\{d_s\}_{s=1}^{S} \sim \Phi_{\texttt{speak}}(v_{tgt}^i)$ when describing an image $v_{tgt}^i \in V_{tgt}$ of a concept $c_{tgt} \in C$. $R_t^s(d_s,\theta)$ is defined as:

$$\min\left(r_t^s(\theta)\hat{A}_t^s, \text{clip}\left(r_t^s(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t^s\right), \tag{4}$$

where $r_t^s(\theta)$ is the importance sampling ratio and $\hat{A}_t^s$ the normalized advantage accounting for the different rollouts descriptions.

Success in the reference game translates into the ability to describe the invariant and discriminative features of a concept, aligning the speaker's behavior with the objectives of concept characterization necessary for MLLM's personalization. It is important to emphasize that as the training game concepts set C comprises concepts different from the user-defined personal concepts P, the model must generalize its learned behavior to unseen user-provided concepts during personalization at inference.

Note that the generalization is over unseen concepts, not unseen *types* of discriminative attributes (*e.g.*, color *vs.* texture). So the speaker can reuse attribute types already seen in training.

3.3 Inference via referential descriptions

Inspired by prior work, we leverage R2P [9] retrieval-based strategy for inference. Personal concepts P are stored in a database where their visual information is accompanied with referential descriptions from the speaker, and the corresponding concept name. Formally, let $v_p \in V_p$ be the visual information of the p-th personal concept, with t_p its associated name. We can define d_p as the corresponding speaker-based description $d_p = \Phi_{\texttt{speak}}(v_p)$. A personal database is constructed as:

$$\mathcal{D} = \{t_p, v_p, d_p, f_p^V, f_p^T\}_{p\in P}, \tag{5}$$

where $f_p^V = E^V(v_p)$ denotes the visual embedding of the reference image v_p of a CLIP-like [31] encoder E^V. For the descriptions we compute textual embeddings $f_p^T = E^T(d_p)$ using the text encoder E^T. Starting from the query image Q, most relevant concepts according to visual and textual similarity are retrieved:

$$s_{Qp} = \frac{1}{2}\left(\langle f_p^V, E^V(Q)\rangle + \langle f_p^T, E^V(Q)\rangle\right), \quad p \in P, \tag{6}$$

and the Top-K elements P^K are selected as candidates. Finally, the personalized model reasons on the available concept options $\mathcal{O} = \{(t_p, d_p)\}_{p\in P^K}$, *i.e.*, possible candidates, to provide a personalized answer. To this end, the referential

descriptions of selected candidates are compared to the query image to form a closed-set problem:

> $\texttt{prompt}_a(\mathcal{O}, q)$ = "Consider concept-description pairs $\{\mathcal{O}\}$ and choose among one of the concepts. Please answer $\{q\}$:",

where $\texttt{prompt}_a$ is the answer templating function and q refers to the task specific instruction, *i.e.*, recognition, recall or captioning. The final answer evaluates as:

$$a = \Phi_{\texttt{list}}(Q, \texttt{prompt}_a(\mathcal{O}, q)). \tag{7}$$

Hence, the personalized MLLM $\Phi_{\texttt{list}}$ can produce tailored captions or respond to user-specific queries, conditioned on the user's prompt. Importantly, if the user explicitly refers to the concept by name t_p in the prompt, the system can directly fetch the associated textual description d_p from $\mathcal{D}$, bypassing the retrieval-and-reasoning stage of the strategy. Please, refer to the *Supp. Mat.* for the implementation details.

4 Experiments

We evaluate RRG across several established personalization benchmarks, including MyVLM [2], Yo'LLaVA [24], and PerVA dataset [9]. We begin by outlining the experimental setting (Sec. 4.1), followed by a comparison against state-of-the-art approaches (Sec. 4.2) and motivating experiments on the role of descriptions (Sec. 4.3). Finally, we present an ablation study analyzing the reward function used in our method and show some qualitative results (Sec. 4.4).

4.1 Experimental setting

Dataset. We experiment with three personalization benchmarks: MyVLM [2], Yo'LLaVA [24], and PerVA [9]. The MyVLM benchmark defines 29 personalized concepts, each representing user-specific visual concepts such as pets, accessories, or personal belongings, designed to test the model's ability to recognize and reason on fine-grained visual identity cues. Yo'LLaVA extends this evaluation to 40 categories covering a broader range of concepts, including people and buildings, thereby offering a more comprehensive benchmark for real-world personalization requests. Finally, we leverage 30 concepts from the PerVA dataset for training the game, while we use the remaining 269 for testing. PerVA stress-tests the personalization on visually ambiguous instances with challenging state changes.
Downstream tasks and metrics. We follow prior work [9, 14, 24] and assess the effectiveness of RRG across three personalization tasks: *object recognition*, *caption generation*, and *personalized VQA*. For object recognition, the model must determine whether a query image contains the personalized concept of interest. Positive samples depict the target concept, whereas negative samples contain other concepts from the benchmark. We treat this task as binary classification and report recall (`Pos.` ↑), specificity (`Neg.` ↑), and their balanced average (`Wtd` ↑).

For captioning, the objective is to determine whether the model can correctly reference the personalized concept without being explicitly prompted with its name. We report Precision (`Prec.` ↑), Recall (`Rec.` ↑), and their harmonic mean (`F1` ↑), computed per concept and then macro-averaged across all concepts within each dataset. Finally, for the personalized VQA task, the model answers closed-set questions concerning the personalized concepts and report the answer `Accuracy` (↑). All experiments are averaged on three random seeds.

Baselines. We compare RRG with several state-of-the-art methods for MLLM personalization. Among training-based approaches, MyVLM [2] and Yo'LLaVA [24] leverage test-time training to incorporate new personalized concepts. Instead, RAP [14] (and its Qwen2VL variant, `RAP-Qwen`) and RePIC leverage large-scale pre-training and a retrieval-augmented paradigm with an instruction-tuned MLLM to produce personalized responses, with supervised fine-tuning and personalized RL, respectively. We also compare with R2P [9], a training-free approach leveraging textual descriptions for personalization. For fair comparison, we adapted the approach for Qwen2VL removing its costly pairwise matching step, maintaining an equal inference budget across all baselines. Following the VQA task protocol of [9], as baselines we include `LLaVA+prompt` and `GPT-4V+Vprompt`, where the reference description and image are leveraged, respectively.

4.2 Comparison with state-of-the-art models

Table 1 reports captioning and recognition performance when comparing RRG to the state of the art.

Captioning. RRG consistently achieves the best results on captioning, on all datasets. On MyVLM, RRG yields substantial gains across all metrics, reaching a 95.0 F1 and markedly surpassing training-based baselines such as RePIC (76.7 *vs*. 95.0 F1) and RAP-Qwen (+12% Rec.). A similar pattern emerges on Yo'LLaVA, where RRG improves F1 by +16.1% over the strongest training-based model. Finally, on PerVA, our reference-game yields a large margin improvement on F1 over the second-best method (86.9 *vs*. 67.3 of R2P-Qwen), highlighting the strength of our discriminative and accurate descriptions in character-

Table 1: Comparison with state of the art. Recognition and captioning performance (↑) on established personalization benchmarks. Captioning metrics are macro-averaged (Precision/Recall/F1).

Method	Recognition			Captioning		
	Pos.	Neg.	Wtd	Prec.	Rec.	F1
MyVLM Dataset [2]						
MyVLM [2]	96.6	90.9	93.8	-	-	-
Yo'LLaVA [24]	97.0	95.7	96.4	-	-	-
R2P-Qwen [9]	86.9	95.4	91.5	94.4	94.1	93.4
RAP-LLAVA [14]	94.4	**98.8**	96.6	90.1	78.7	82.0
RAP-Qwen [26]	97.2	94.8	95.5	88.6	85.2	84.9
RePIC [26]	**98.1**	95.6	**96.8**	84.2	73.2	76.7
RRG (Ours)	91.5	92.0	91.8	**95.7**	**95.4**	**95.0**
Yo'LLaVA Dataset [24]						
Yo'LLaVA [24]	94.9	89.8	92.4	-	-	-
R2P-Qwen [9]	86.2	95.3	90.8	89.8	87.5	86.3
RAP-LLAVA [14]	86.0	**99.2**	92.2	83.1	70.0	76.3
RAP-Qwen [26]	**97.2**	87.5	92.3	76.3	70.7	69.4
RePIC [26]	91.9	91.1	91.5	79.5	57.7	62.7
RRG (Ours)	97.0	92.3	**94.7**	**91.9**	**89.3**	**88.6**
PerVA Dataset [9]						
MyVLM [2]	66.0	58.5	62.2	-	-	-
Yo'LLaVA [24]	75.1	69.0	72.0	-	-	-
R2P-Qwen [9]	88.2	91.8	90.0	73.0	67.7	67.3
RAP-LLAVA [14]	92.9	85.2	89.0	63.8	49.5	51.8
RAP-Qwen [26]	93.2	92.9	**93.1**	66.0	50.4	52.8
RePIC [26]	88.8	**96.5**	92.6	62.4	44.3	48.9
RRG (Ours)	**94.0**	88.3	91.1	**89.5**	**87.9**	**86.9**

Table 2: Comparison with state of the art by category. F1 Captioning performance (↑) on Yo'LLaVA dataset.

Method	L	C	O	P	A	Avg.
R2P-Qwen [9]	**100**	83.6	88.5	75.0	**92.1**	86.3
RePIC [26]	36.4	86.2	75.0	53.7	83.5	62.7
RAP-Qwen [14]	56.3	**96.7**	78.3	**84.1**	70.0	76.3
RRG (Ours)	98.7	93.8	**92.2**	75.7	88.2	**88.6**

Table 3: VQA results in terms of answering accuracy (↑) in Yo'LLaVA [24].

Method	Accuracy
Yo'LLaVA [24]	92.9
GPT-4V + Vprompt [24]	86.6
LLaVA [24]	89.9
LLaVA + prompt [24]	92.5
R2P-Qwen [9]	94.1
RAP-LLaVA [14]	93.2
RRG (Ours)	**95.3**

izing the concepts (see Fig. 4). Crucially, RRG consistently outperforms R2P-Qwen that leverages textual fingerprints for concept characterization, improving performance across all benchmarks (*e.g.*, +1.6 F1 on MyVLM and +2.3 F1 on Yo'LLaVA). Tab. 2 offers a breakdown of captioning F1 across different categories of concepts on Yo'LLaVA (**L**andmark, **C**haracter, **O**bject, **P**erson and **A**nimal): while trained on PerVA objects, RRG generalizes to new concepts yielding to best or second-best performance across all categories, including those much different from training ones (*e.g.*, landmark). Experiments in the *Supp. Mat.* further demonstrate that RRG is not specific to PerVA. When the same candidate-construction protocol is applied to MC-LLaVA [5] concepts, it yields consistent improvements over the zero-shot speaker.

Recognition. As for recognition (Tab. 1), RRG achieves state-of-the-art Wtd on Yo'LLaVA (+2.3% Wtd over the second best) and competitive performance on PerVA, while we observe a performance gap on MyVLM dataset. Our analysis suggests that such limited performance may stem from hallucinations due to the presence of both the reference image v_p and the query image Q as input. This leads the model to confirm the presence of the described object even when it is absent [42, 44] (see *Supp. Mat.* for a more detailed discussion and a controlled study). In contrast, captioning requires no reference images at inference, avoiding the visual overload induced by multiple retrieved images and limiting cross-modal interference, with RRG achieving substantial gains on this more challenging task. Nevertheless, when compared with the description-based R2P-Qwen, our approach consistently achieves better results (*e.g.*, +4.7% on Yo'LLaVA, +1.1% on PerVA). These results confirm the advantage of reference-game-based descriptions over personalization alternatives that rely on textual fingerprints.

Personalized VQA. Table 3 reports the personalized VQA performance on Yo'LLaVA. Our method achieves the highest VQA accuracy of 95.3%, outperforming the next-best baseline, R2P-Qwen (94.1%). RRG improves over the large-scale pretrained RAP-LLAVA by 2.1%. This improvement highlights the benefit of using accurate and discriminative descriptions generated by our speaker, which provide more unique, informative attributes for the listener MLLM during personalized VQA.

Table 4: The effect of descriptions on personalization. Recognition and captioning performance (↑) on MyVLM [2] and Yo'LLaVA [24] datasets. Captioning metrics are macro-averaged (Precision/Recall/F1).

Method	MyVLM						Yo'LLaVA					
	Recognition			Captioning			Recognition			Captioning		
	Pos.	Neg.	Wtd	Prec.	Rec.	F1	Pos.	Neg.	Wtd	Prec.	Rec.	F1
zs-speaker	88.4	90.2	89.3	89.9	89.4	89.3	84.7	**94.0**	89.3	87.4	85.5	84.6
listener-training	89.1	90.3	89.7	92.6	92.2	92.0	88.3	92.4	90.4	89.0	87.3	86.3
speaker-training	91.3	**93.0**	**92.2**	93.0	92.9	92.3	95.0	92.8	93.9	89.1	87.6	87.0
RRG (Ours)	**91.5**	92.0	91.8	**95.7**	**95.4**	**95.0**	**97.0**	92.3	**94.7**	**91.9**	**89.3**	**88.6**

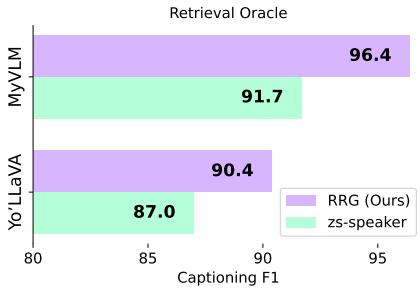
4.3 The role of referential descriptions

We conduct experiments to motivate the importance of descriptions and the reference game. Tab. 4 reports personalization results on established MyVLM and Yo'LLaVA datasets in terms of both recognition and captioning.

To assess the benefit of enhancing descriptions, we first compare RRG against the pre-trained zero-shot speaker (`zs-speaker`), which is asked to simply describe the personal concept given the reference image. The consistent improvements by RRG indicate that refining concept descriptions with RRG enhances personalization in both recognition and captioning tasks (*e.g.*, +5.4 Wtd on Yo'LLaVA and +5.7 F1 in captioning on MyVLM).

Furthermore, we investigate whether personalization benefits more from enhancing concept descriptions (training the speaker) or from strengthening concept recognition (training the listener). To this end, we compare RRG against the `listener-training` baseline in which the personalized model is optimized to recognize objects from zs-speaker descriptions. The results consistently favor prioritizing description quality: RRG yields stronger personalization gains than directly training concept recognition (*e.g.*, +2.1 Wtd in MyVLM, +2.3 in captioning F1 in Yo'LLaVA). Intuitively, low-quality original descriptions, either lacking discriminative cues or containing irrelevant, distracting attributes, impair listener training. Notably, the reference game enhances description quality, surpassing a speaker trained on the same data without the game: RRG outperforms `speaker-training`, where the speaker is optimized with an MLLM reward model in which the MLLM serves as a verifier, predicting the presence of the target concept conditioned on both the reference image and the generated description (*e.g.*, 95.0 *vs*. 92.3 F1 on MyVLM and 94.7 *vs*. 93.9 on Yo'LLaVA Wtd). We refer the reader to the *Supp. Mat.* for further details on the baselines and for experiments with different listeners: a stronger reward listener improves F1 further, while RRG descriptions continue to outperform zero-shot ones when evaluated with a different inference listener.

Finally, as RRG is a retrieval-augmented strategy, we isolate the impact of improved descriptions on both retrieval and listener reasoning. To this end, in Fig. 3 we consider a *retrieval oracle* setting in which the target concept is guaranteed to be included among the retrieved candidates. Under this controlled condition, we



Method	MyVLM		Yo'LLaVA	
	hit@1	hit@2	hit@1	hit@2
zs-speaker	85.0	96.6	84.0	97.6
R2P-Qwen [9]	**92.2**	96.3	**88.1**	94.3
RePIC [26]	91.5	94.9	73.5	82.5
RAP-Qwen [14]	91.5	94.9	73.5	82.5
RRG (Ours)	88.7	**98.5**	85.4	**98.2**

Table 5 & Fig. 3: The effect of descriptions on reasoning and retrieval. (Left) Captioning performance (↑) on MyVLM [2] and Yo'LLaVA [24] under the retrieval oracle setting. (Right) Retrieval performance (↑) of different retrieval-based approaches.

explicitly evaluate the effect of descriptions on final personalization performance, and compare descriptions produced by a zero-shot speaker (`zs-speaker`) for the personal concept against those generated by RRG. A consistent gain highlights that enhancing concept descriptions benefits listener reasoning at inference time.

Similarly, the retrieval results in Tab. 5 show that more accurate descriptions directly improve retrieval performance, with RRG outperforming the second-best zero-shot speaker in Yo'LLaVA (98.2 *vs*. 97.6 hit@2). We also report performance of previous retrieval-augmented personalization strategies [9, 14, 26] as reference, with RRG achieving the best hit@2 results on both benchmarks.

4.4 Additional results

Ablation on rewards. We perform an ablation study to assess the contribution of the proposed reward strategy. Results are reported in Tab. 6. The `binary` baseline, which binarizes the reward based on whether the listener's guess is correct, provides a competitive starting point but underperforms across most metrics. Instead, using the proposed reward r_{match} (Eq. (2)) yields strong captioning results on both datasets (F1 equal to 95.0 on MyVLM and to 88.6 on Yo'LLaVA), confirming that the normalized probability signal provides a richer training objective than a binary reward. Results are on par in the recognition task where RRG yields +0.8 gain in Yo'LLaVA but a marginal drop in MyVLM.

Qualitative results on concept descriptions. Figure 4 presents qualitative examples which compare the concept descriptions generated by RRG with those produced by the `zs-speaker` and the training-free `R2P-Qwen` approach. As shown, both zs-speaker and R2P-Qwen tend to include unrelated or state-specific features, such as "hanging on a hanger" or "sitting on a plaid blanket" that vary across images and can distract the listener MLLM during inference. In contrast, our finetuned speaker consistently avoids such location- or state-dependent artifacts, generating concise descriptions that focus strictly on stable, identity-defining

Table 6: Ablations on RRG reward. Recognition and captioning performance (↑) on MyVLM [2] and Yo'LLaVA [24] datasets. Captioning metrics are macro-averaged (Precision/Recall/F1).

Method	MyVLM						Yo'LLaVA					
	Recognition			Captioning			Recognition			Captioning		
	Pos.	Neg.	Wtd	Prec.	Rec.	F1	Pos.	Neg.	Wtd	Prec.	Rec.	F1
binary	**92.9**	91.6	**92.3**	93.5	94.2	93.4	95.4	**92.4**	93.9	88.9	86.8	85.5
r_{match} **(Ours)**	91.5	**92.0**	91.8	**95.7**	**95.4**	**95.0**	**97.0**	92.3	**94.7**	**91.9**	**89.3**	**88.6**

	zs-speaker	R2P-Qwen	RRG (Ours)
	The toy is a gray plush elephant with floppy ears... is sitting on a red and blue plaid blanket.	Large floppy ears, gray color, black eyes, white tag with 'lambs & Ivy' branding, sitting position.	The plush elephant has large, floppy ears, textured feet, and a soft, fuzzy body
	The toy is a small, round, purple creature with a mischievous expression... positioned on top of a metallic object.	Resembles the character Gengar, placed on top of a metal container.	The toy is a small, purple, spiky creature with a mischievous expression and a pink nose
	The cloth is a black dress with a V-neckline, long sleeve,..,hanging on a hanger with a white tag attached.	Asymmetrical neckline, full-sleeve design, Smooth texture, hanging on a hanger.	The black dress features a distinctive cut with flowing, sheer fabric creating an elegant silhouette.

Fig. 4: Qualitative examples of RRG speaker descriptions. We compare the descriptions generated by RRG with zs-speaker and R2P-Qwen for three different concepts. Distracting attributes (*e.g.*, background or state-specific) are highlighted.

attributes (*e.g.*, "floppy ears, textured feet", "with flowing fabric") of each personalized concept. Moreover, RRG descriptions add discriminative details that further aid concept recognition, *e.g.*, "spiky creature" and "sheer fabric". These invariant descriptions allow the listener to accurately disambiguate among competing personalized concepts, even when the query image differs significantly from the reference ones. More qualitative results are in *Supp. Mat.*

5 Conclusions

We addressed MLLM personalization by proposing a novel learning framework based on a reinforced reference game. In the game, a multimodal large language model learns to communicate a concept to a listener, resulting in personalized

descriptions that are increasingly more discriminative and accurate representations of personal concepts. We validated the advantage of our method, RRG, in comparison to state-of-the-art methods through experiments across three tasks on three benchmark datasets. Our findings show that fine-grained, immutable descriptions significantly enhance downstream recognition, captioning, and VQA performance, without requiring large-scale training or expensive inference-time pairwise reasoning. As future work, we will explore novel joint training strategies for the speaker and listener to enable more efficient and adaptive communication.

6 Acknowledgement

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). This work was supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007), European Union (Next Generation EU), and the EU Horizon ELLIOT (No. 101214398) project. This research has also received funding from the European Union's Horizon Europe research and innovation actions under grant agreement No 101215032.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 1
2. Alaluf, Y., Richardson, E., Tulyakov, S., Aberman, K., Cohen-Or, D.: Myvlm: Personalizing vlms for user-specific queries. In: ECCV (2024) 1, 2, 3, 4, 9, 10, 12, 13, 14, 6
3. Alaniz, S., Marcos, D., Schiele, B., Akata, Z.: Learning decision trees recurrently through communication. In: CVPR (2021) 5
4. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) 1
5. 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) 4, 11, 8
6. Cohen, N., Gal, R., Meirom, E.A., Chechik, G., Atzmon, Y.: "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In: ECCV (2022) 4
7. Corona Rodriguez, R., Alaniz, S., Akata, Z.: Modeling conceptual understanding in image reference games. In: NeurIPS (2019) 5
8. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017) 4
9. Das, D., Talon, D., Wang, Y., Mancini, M., Ricci, E.: Training-free personalization via retrieval and reasoning on fingerprints. In: ICCV (2025) 1, 2, 3, 4, 7, 8, 9, 10, 11, 13
10. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: CVPR (2017) 4

11. Dingjie, S., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. In: Conference on Language Modeling (2024) 7
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023) 3
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 3, 7, 1
14. Hao, H., Han, J., Li, C., Li, Y.F., Yue, X.: Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant. In: CVPR (2025) 1, 2, 4, 9, 10, 11, 13, 5
15. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022) 7, 1
16. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014) 4
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023) 4
18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) 4
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) 1
20. Li, Z., Yu, W., Huang, C., Liu, R., Liang, Z., Liu, F., Che, J., Yu, D., Boyd-Graber, J., Mi, H., et al.: Self-rewarding vision-language model via reasoning decomposition. arXiv preprint arXiv:2508.19652 (2025) 4
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 1
22. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: ICLR (2024) 4
23. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025) 4
24. Nguyen, T., Liu, H., Li, Y., Cai, M., Ojha, U., Lee, Y.J.: Yo'LLaVA: Your personalized language and vision assistant. In: NeurIPS (2024) 1, 2, 3, 4, 9, 10, 11, 12, 13, 14, 6
25. Nguyen, T., Singh, K.K., Shi, J., Bui, T., Lee, Y.J., Li, Y.: Yo'chameleon: Personalized vision and language generation. In: CVPR (2025) 2, 4
26. Oh, Y., Mok, J., Chung, D., Shin, J., Park, S., Barthelemy, J., Yoon, S.: Repic: Reinforced post-training for personalizing multi-modal language models. In: NeurIPS (2025) 2, 3, 4, 10, 11, 13, 5
27. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022) 4
28. Pham, C., Phan, H., Doermann, D., Tian, Y.: Personalized large vision-language models. arXiv preprint arXiv:2412.17610 (2024) 4
29. Pi, R., Zhang, J., Han, T., Zhang, J., Pan, R., Zhang, T.: Personalized visual instruction tuning. In: ICLR (2025) 4
30. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.M.A., Botvinick, M.: Machine theory of mind. In: ICML (2018) 5
31. 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: ICML (2021) 8

32. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 4
33. Samuel, D., Ben-Ari, R., Levy, M., Darshan, N., Chechik, G.: Where's waldo: Diffusion features for personalized segmentation and retrieval. In: NeurIPS (2024) 4
34. 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) 2
35. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 4
36. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. In: CVPR (2024) 4
37. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. In: NeurIPS (2020) 4
38. Sundaram, S., Chae, J., Tian, Y., Beery, S., Isola, P.: Personalized representation from personalized generation. In: ICLR (2025) 4
39. Takmaz, E., Giulianelli, M., Pezzelle, S., Fernandez, R., et al.: Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In: ACL (2023) 5
40. Tang, L., Jiang, P.T., Xiao, H., Li, B.: Towards training-free open-world segmentation via image prompt foundation models. IJCV (2024) 4
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 1
42. Yang, T., Li, Z., Cao, J., Xu, C.: Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In: ICLR (2025) 11, 10
43. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. In: ICLR (2024) 4
44. Zheng, H., Zhang, Z.: Modality bias in lvlms: Analyzing and mitigating object hallucination via attention lens. arXiv preprint arXiv:2508.02419 (2025) 11, 10
45. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: ICLR (2024) 10