

RAP: Retrieval-Augmented Personalization for Multimodal Large Language Models

Supplementary Material

A. Appendix Overview

- Section **B**: Additional evaluations of our models.
- Section **C**: More experimental details.
- Section **D**: More details of the RAP dataset.
- Section **E**: Additional demonstrations.
- Section **F**: Analysis of limitations of our work.
- Section **G**: Analysis of potential ethics issues.
- Section **H**: Examples of the personalized database.

B. Additional Evaluation Results

B.1. Ablation Study

Dataset Composition. We conduct ablation experiments on the question answering and recognition benchmark, experimental results are present in Table 7. The results further demonstrate that our data augmentation and the constructed negative samples also contribute to the model’s performance.

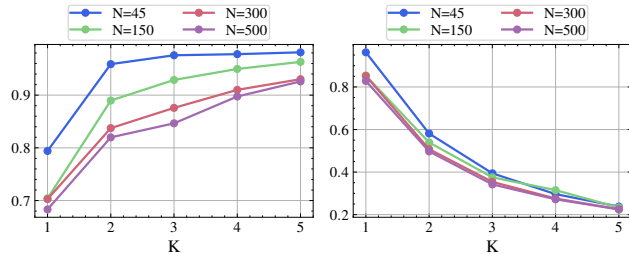


Figure 5. **Retriever’s Top-K Recall** under varying database size N.

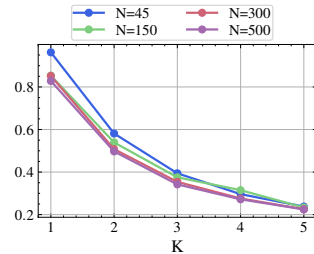


Figure 6. **Retriever’s Top-K Precision** under varying database size N.

Retriever. The retriever’s performance is crucial for a RAG system. We assess the retriever’s performance on the personalized captioning dataset. We use the detection model to identify potential concepts and retrieve the K concepts with the highest similarity from the database. Figure 5 and Figure 6 show the Top-K recall and precision for different values of K and database sizes N. The results indicate that as the database size increases, the retriever’s performance declines. While a larger K generally enhances recall, it also introduces more irrelevant concepts, leading to a drop in precision. Notably, even with 500 personal concepts to remember, the Top-5 recall rate can still exceed 90%, which guarantees the effectiveness of our RAP framework.

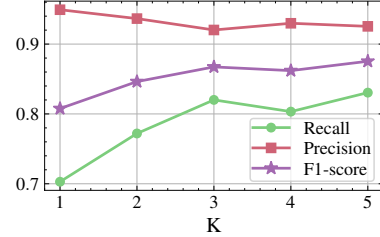


Figure 7. **Captioning Performance** with varying number of retrieved concepts.

Impact of Retrieving Different Numbers of Concepts.

We train a new model, RAP-LLaVA-OneVision (0.5B) [22], to analyze the impact of the number of retrieved concepts in detail. As shown in Figure 7, retrieving more concepts from a database of 300 concepts improves the recall metric of image captioning, but also introduces more noise, leading to a drop in precision. This reflects the trade-off between supporting a larger set of personalized concepts and maintaining generation precision.

Table 6. **Evaluation on Knowledge-intensive Multimodal Benchmarks.** KB: Knowledge Base.

Method	MMMU [54]	InfoSeek [6]
LLaVA [28]	0.364	0.205
LLaVA-LoRA [17]	0.359	0.205
RAP-LLaVA	0.361	0.218
RAP-LLaVA(With KB)	0.369	0.344

Multimodal Benchmark. We also evaluate our model’s performance on several traditional multimodal benchmarks, including MMMU [54] and InfoSeek [6]. We assess our models’ performance both with and without external knowledge base. For MMMU [54], we use 30K images paired with corresponding captions from Wikipedia as the external knowledge base. During testing, we retrieve the three most similar images based on the question’s image and incorporate only the textual knowledge to the input. For InfoSeek [6], we randomly sample 5K questions from the validation set and construct a knowledge base containing 50K entities from Wikipedia database provided by the authors, which includes all relevant entities associated with the questions. For each question, we retrieve the most similar entity and add only the textual knowledge to the input.

We evaluate on the validation set of MMMU, and 5K questions sampled from the validation set of InfoSeek. We use the official scripts to get the results, which are pre-

Table 7. Ablation studies on **Question Answering and Visual Recognition**. Weighted results are computed as arithmetic means.

Method	Question Answering			Visual Recognition		
	Visual	Text	Weighted	Positive	Negative	Weighted
RAP-LLaVA	0.935	0.938	0.936	0.979	0.982	0.980
- Data augmentation	0.924 (-0.011)	0.918 (-0.020)	0.921 (-0.015)	0.943 (-0.036)	0.988 (+0.006)	0.965 (-0.015)
- Negative samples	0.918 (-0.017)	0.933 (-0.005)	0.925 (-0.011)	0.958 (-0.021)	0.985 (+0.003)	0.971 (-0.009)
- Text information	0.935	0.908 (-0.030)	0.921 (-0.015)	0.988 (+0.009)	0.930 (-0.052)	0.959 (-0.021)

sented in Table 6. From the results, our RAP-LLaVA retains most general knowledge of the original LLaVA [28]. It also equips the MLLM with the ability to retrieve information from an external knowledge base, demonstrating superior performance in knowledge intensive tasks.

B.2. Standard Image Captioning Metrics

We further evaluate using standard image captioning metrics. The results are shown in Table 8. Despite the additional concept information is offered, LLaVA’s performance declines due to its inability to effectively utilize the information. Although our RAP-LLaVA is not trained on these concepts, it still achieves comparable performance on general image captioning tasks while enabling personalization.

B.3. Robustness to Retrieval Errors

As described in Section 3.2, we construct negative samples to enhance models’ robustness to retrieval errors. Figure 5 and 6 show the retriever’s recall and precision under different database sizes. As the database size grows, the retriever’s recall and precision decrease, which means a higher likelihood of retrieving irrelevant concepts. Despite this, our models maintain outstanding performance, as shown in Figure 4, which demonstrates their robustness to retrieval errors. Additionally, results in Table 5 show that our models perform even better when the retrieval process is error-free. In Table 11, we provide examples to show model’s robustness to retrieval errors. Thanks to negative samples in the training dataset, RAP-MLLMs can distinguish irrelevant concepts and generate accurate responses.

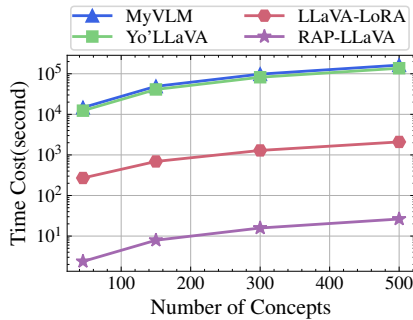


Figure 8. **Time Cost of Personalization**. We conduct experiment with 2 A800 GPUs.

B.4. Time Cost of Personalization

We also evaluate the time cost associated with different methods for learning a set of user’s concepts. The results are presented in Figure 8. MyVLM has to train an external recognition model for each concept and learn an embedding to adjust the model’s outputs. Similarly, YoLLaVA needs to learn new tokens for each concept. During the optimization process, both approaches necessitate multiple forward and backward pass of the MLLM, resulting in significant time consumption. In contrast, our RAP only requires time for encoding the image and adding its embedding to the database, which can be accomplished in just a few seconds. This significantly enhances the convenience and practicality of our models in practical applications.

C. More Experimental Details

Implementation Details. We utilize YOLO-Worldv2-X [7] as the detection model, setting detection classes to include all categories stored in the database to reduce the interventions from unrelated objects. We construct a multimodal retriever using Facebook AI Similarity Search (FAISS) [18], employing a pre-trained CLIP ViT-L/14-336 [34] as the visual encoder. Each key in the database is generated by inputting a concept’s image into the CLIP visual encoder, resulting in a 768-dimensional vector. Considering the restriction of context length of the backbone language model, we retrieve the 2 most similar images from the database for each region of interest. Then we select 2 and 3 different concepts with the highest similarity among all as supplementary inputs for RAP-LLaVA and RAP-Phi3-V, respectively.

Baselines. For MyVLM, we find that when the training data is very limited, it is quite hard for the classification head to work effectively. Therefore, we use data augmentation to help improve its performance. Specifically, we crop the single image into several pieces containing the target concept to improve the accuracy of classification heads. To distinguish between multiple different concepts that may appear in the image, we use $\langle \text{sks1} \rangle$, $\langle \text{sks2} \rangle \dots$ as concept identifiers. For YoLLaVA, we present its experimental results reported in the original paper [32]. For GPT-4V, reference images and descriptions are provided as additional prompts, thus it can generate responses about the target concept.

Table 8. **Quantitative Evaluation with Standard Image Captioning Metrics.** We compute standard image captioning metrics for personalized captions generated by RAP-LLaVA. For each image, we treat all 5 augmented captions as the set of ground truth captions. The results are averaged across all three evaluations and all concepts.

Model	B1	B2	B3	B4	METEOR	ROUGE.L	CIDEr	SPICE
LLaVA [28]	0.177	0.100	0.057	0.032	0.138	0.240	0.428	0.136
LLaVA [28] + Retriever	0.087	0.038	0.014	0.005	0.074	0.132	0.025	0.020
MyVLM-LLaVA [2]	0.195	0.103	0.061	0.036	0.213	0.263	0.435	0.145
RAP-LLaVA	0.178	0.094	0.058	0.037	0.215	0.189	0.388	0.108

Multi-concept Data Collection. We collect videos from YouTube and sample frames from them. For each video, we detect multiple concepts and use Gemini [12] to find frames that contain both concepts in a given pair. For each pair, there are 8 to 13 images used for testing. Examples of these concept pairs are shown in Table 2. We generate five captions for each image to enhance the concept learning of baseline methods.

D. Details of Dataset

D.1. Dataset Composition

- We provide a summary of the composition of our dataset in Figure 9, which visually represents the distribution of different components.
- Table 9 presents detailed numerical data for each part.
- In Table 10, we specify the sources for each component of our dataset.

D.2. Instructions

In this section, we present the instruction templates used to create our dataset:

- Table 22 contains instructions for visual grounding and recognition.
- Table 23 includes example instructions for image captioning.
- Table 24 presents example instructions for image description.
- Table 25 presents example questions used for question answering synthesis.

E. Additional Demonstrations

In this section, we provide more qualitative results obtained by various models.

- In Table 12, we demonstrate how our models achieve real-time editing of concepts by modifying the database.
- In Table 13, we demonstrate the real-time addition of new concepts by updating the database.
- In Table 14, we present qualitative results on personalized conversation of RAP-LLaVA.

- In Table 15, we present qualitative results on personalized conversation of RAP-Phi3-V.
- In Table 16, we present additional image captions generated by RAP-LLaVA and other methods.
- In Table 17, we present additional image captions generated by RAP-Phi3-V and other methods.
- In Table 18, we provide demonstrations of image description generated by RAP-LLaVA and LLaVA.
- In Table 19, we provide demonstrations of image description generated by RAP-Phi3-V and Phi3-V.
- In Table 20 and 21, we provide results on visual recognition of RAP-LLaVA. It also has the ability to give precise bounding box of specific concept in the image.

F. Limitation

Our proposed RAP framework is a retrieval-based method. The limitations of RAP mainly concern the additional computational cost of generation and the precision of the retriever. While incorporating external information effectively generates more specific answers, it inevitably increases the context length for MLLMs, leading to additional computational overhead during the generation process. We will further explore ways to mitigate this computational burden. Another limitation is that the personalization performance of our RAP-MLLMs depends on the retriever’s capability. This proposes a need for a robust multimodal retriever that can discern intricate features to enhance retrieval precision. Despite these limitations, RAP offers a timely solution for MLLM personalization. By retrieving from a user’s specific database, RAP facilitates reliable and flexible personalized generation, which is valuable in practical applications.

G. Potential Ethics Issues

In our RAP framework, users can deploy and manage their own databases and personalized MLLMs locally. This design gives them greater flexibility to customize the model to fit their specific needs while ensuring privacy and security. However, improper use of personalized MLLMs can result in bias or other ethical concerns. For example, the model may unintentionally generate biased information or spread

Figure 9. Composition of our dataset.

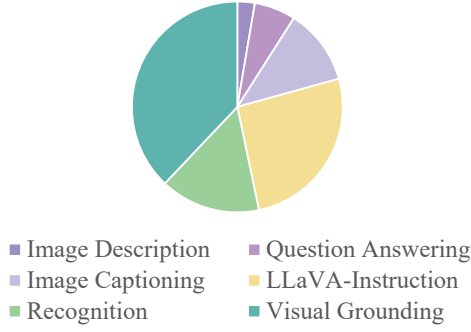


Table 10. Data source.

Type	Source Dataset
Visual Grounding	RefCOCO [20], TAO [9] ILSVRC2015-VID [39], Object365 [40]
Recognition	CustomConcept101 [21], CelebA [30]
Caption & Description	RefCOCO [20], TAO [9] Object365 [40], CustomConcept101 [21]
Question Answering	RefCOCO [20], TAO [9] Object365 [40], CustomConcept101 [21] CelebA [30]
LLaVA-Instruction	LLaVA-Instruct-665K [28]

Table 11. **Examples of Robustness to Retrieval Errors.** Our RAP-MLLMs can distinguish noisy concepts and generate accurate answers.




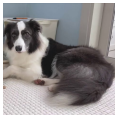
Image	Retrieved Concepts
	 my dog  dog1
	 dog*
<p>Question: Give a caption of this image. RAP-Phi3-V: my dog is looking out the window.</p> <p>Question: Can you see dog* in this image? RAP-Phi3-V: No, dog* is not visible in this picture.</p>	

Table 9. Statistics of our dataset.

Type	Size
Visual Grounding	100K
Recognition	40K
Caption & Description	37K
Question Answering	16K
LLaVA-Instruction	67K
Total	260K

H. Examples of the Personalized Database

We give some visualized examples of our database in Table 26. For each concept in the database, users need to provide an image with its name and optional text information. During inference, the images, names and other information of retrieved concepts are integrated into the input for the MLLM. Users have the flexibility to define the name and personalized description based on their preferences, and our RAP-MLLMs will generate answers according to the provided information.

misleading or harmful content. To mitigate these risks, it is crucial to implement effective mechanisms for bias detection, fairness evaluation, and responsible model usage.

Table 12. **Examples of Concept Editing.** Based on the information recorded in the database, our RAP-LLaVA can provide reliable and accurate answers.



Image	Concept	Information in the Database
	my dog	1. A white and gray dog with long fur. He has black eyes. His favorite food is <u>chicken</u> . 2. A white and gray dog with long fur. He has black eyes. His favorite food is <u>beef</u> .
		Question: What is \langle my dog \rangle 's favorite food? 1. RAP-LLaVA: <u>Chicken</u> . 2. RAP-LLaVA: His favorite food is <u>beef</u> .

Table 13. **Examples of Concept Updating.** The first caption is generated when toy2 not yet stored in the database. Once the new concept is added, RAP-LLaVA can recognize both toy1 and toy2.




Image	Concept	Information in the Database
	toy1	A plush toy. It is orange with a yellow belly and a brown nose.
	toy2	This is a plush toy of the bluey character. It is a light blue color with a purple patch on its head, and its ears are yellow.
		Question: Give a caption of this image. 1. RAP-LLaVA: \langle toy1 \rangle is ready for bed! He's snuggled up with his friend, a blue and yellow dog plushie. They're both looking so cozy and ready for a good night's sleep! (Add new concept \langle toy2 \rangle to the database) 2. RAP-LLaVA: \langle toy1 \rangle and \langle toy2 \rangle are ready for bed! They are both snuggled under the covers with their favorite stuffed animals.

Table 14. Examples of personalized conversations obtained by RAP-LLaVA.


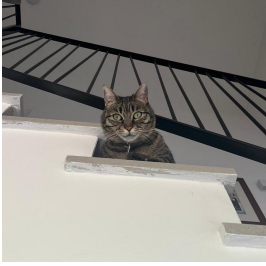




Image	Concept	Information in the Database
	my cat	A tabby cat with its eyes closed is laying on a light grey surface. The cat's fur is a mix of light and dark brown stripes.
		<p>Question: Where is \langlemy cat\rangle? RAP-LLaVA: \langlemy cat\rangle is on a staircase.</p> <p>Question: What is the type of \langlemy cat\rangle? RAP-LLaVA: \langlemy cat\rangle is a tabby cat.</p> <p>Question: What is the color of \langlemy cat\rangle? RAP-LLaVA: \langlemy cat\rangle is a tabby cat with a mix of light and dark brown stripes.</p>
Image	Concept	Information in the Database
	A	\langle A \rangle is a smiling Asian man with short black hair, wearing glasses and a grey sweater.
		<p>Question: Can you see \langleA\rangle in this photo? RAP-LLaVA: Yes, \langleA\rangle is captured in this image.</p> <p>Question: What color is \langleA\rangle's shirt? RAP-LLaVA: \langleA\rangle is wearing a purple shirt.</p> <p>Question: What is \langleA\rangle's right hand holding? RAP-LLaVA: Chopsticks.</p>
Image	Concept	Information in the Database
	dug	\langle dug \rangle is a golden retriever from the movie Up. He has a brown coat and a big, goofy smile, wearing a collar with some round objects.
(Text-only conversation)		<p>Question: Can you describe \langledug\rangle in detail? RAP-LLaVA: \langledug\rangle is a golden retriever from the movie Up. He has a brown coat and a big, goofy smile. He is wearing a collar with some round objects. He looks very happy and playful.</p>
Image	Concept	Information in the Database
	P	\langle P \rangle is a bald man with a red beard.
(Text-only conversation)		<p>Question: Can you describe \langleP\rangle's appearance in detail? RAP-LLaVA: \langleP\rangle is a bald man with a red beard. He is wearing orange shorts and a blue shirt. He is sitting in a black director's chair with his legs crossed.</p>

Table 15. Examples of personalized conversations obtained by RAP-Phi3-V.

Image	Concept	Information in the Database
	bo	<p>⟨bo⟩ is a well-groomed, medium-sized Shiba Inu with a thick, cinnamon-colored coat, cream accents, alert eyes, and a black collar.</p>
		<p>Question: What is ⟨bo⟩ doing?</p> <p>RAP-Phi3-V: ⟨bo⟩ is lying on its side on a fluffy, gray dog bed. It is resting its head on a white stuffed toy. ⟨bo⟩ has a thick, cinnamon-colored coat with cream accents, alert eyes, and a black collar. ⟨bo⟩ looks very relaxed and content.</p> <p>Question: Where do you think he likes to play?</p> <p>RAP-Phi3-V: It’s hard to say for sure, but it’s possible that ⟨bo⟩ enjoys going to the park. Many dogs like to go to the park to play and get exercise.</p>
Image	Concept	Information in the Database
	Anya	A young girl with pink hair and big green eyes.
	Bond	A white dog wearing a black bow tie.
		<p>Question: What are they doing, describe in detail.</p> <p>RAP-Phi3-V: ⟨Anya⟩ is feeding ⟨Bond⟩, a white dog wearing a black bow tie. ⟨Bond⟩ is eating out of a bowl. ⟨Anya⟩ is kneeling down and watching ⟨Bond⟩ eat. ⟨Bond⟩ is standing on a tiled floor and looking at the bowl. ⟨Anya⟩ is wearing a pink sweater and has pink hair. ⟨Bond⟩ is wearing a black bow tie. The scene is indoors, and there is a TV in the background.</p> <p>Question: What is ⟨Anya⟩ wearing?</p> <p>RAP-Phi3-V: ⟨Anya⟩ is wearing a black sweater.</p>
Image	Concept	Information in the Database
	C	<p>⟨C⟩ is a young woman with long blonde hair, wearing a white sleeveless shirt. She has bright red lipstick and light makeup.</p>
(Text-only conversation)		<p>Question: What color is ⟨C⟩’s hair?</p> <p>RAP-Phi3-V: ⟨C⟩’s hair is blonde.</p> <p>Question: What color of clothes does she like?</p> <p>RAP-Phi3-V: She likes white clothes.</p>

Table 16. Additional qualitative comparison on image captioning between RAP-LLaVA and other methods.







Image	Caption
 <p>doll*</p>	<p>LLaVA: A blue coffee cup sits next to two dolls on a window sill.</p> <p>LLaVA-LoRA: ⟨doll*⟩ positioned on a windowsill with a blue mug containing pencils</p> <p>MyVLM: ⟨doll*⟩ is found sitting beside a window ledge with a cup and toothbrushes nearby. In the image, ⟨doll*⟩ appears to be peacefully resting, possibly enjoying the view or taking a break.</p> <p>RAP-LLaVA(Ours): ⟨doll*⟩ and her friend are ready for a fun day at the office! They're ready to take on any challenge that comes their way.</p>
 <p>mug*</p>	<p>LLaVA: A wooden shelf holds several potted plants, including a large clay pot and a small vase. The shelf is positioned near a wall, and the plants are arranged in a visually appealing manner.</p> <p>LLaVA-LoRA: ⟨mug*⟩ on a wooden shelf next to a plant and two potted plants.</p> <p>MyVLM: ⟨mug*⟩ on a blue and white tiled floor next to indoor plants, a shelf with pots and a vase.</p> <p>RAP-LLaVA(Ours): A brown ceramic tiki mug with the face of a tiki head, ⟨mug*⟩, sits on a wooden shelf with two potted plants.</p>
 <p>my cat</p>	<p>LLaVA: A cat is sitting on a ledge near a staircase.</p> <p>LLaVA-LoRA: ⟨my cat⟩ sitting on a staircase, looking down.</p> <p>MyVLM: ⟨my cat⟩ in a typical pose, gripping the banister of a staircase. He is wearing a collar.</p> <p>RAP-LLaVA(Ours): ⟨my cat⟩ is relaxing on the stairs. Look at those eyes! What a beautiful cat.</p>
 <p>my cat</p> <p>teapot*</p>	<p>LLaVA: A cat is laying on a blanket on a couch, next to a colorful teapot.</p> <p>LLaVA-LoRA: ⟨my cat⟩ reclining on a chair with a ⟨teapot*⟩ beside</p> <p>MyVLM: ⟨teapot*⟩ on the couch near an orange, yellow, and blue teapot. The cat is laying on a blanket, and the teapot is placed on a table.</p> <p>RAP-LLaVA(Ours): ⟨my cat⟩ is taking a well-deserved nap next to ⟨teapot*⟩.</p>
 <p>H</p> <p>T</p>	<p>LLaVA: A man and a woman are walking down a street, with the man taking a selfie using his cell phone. They are both smiling as they walk, and the man is holding his phone ...</p> <p>LLaVA-LoRA: ⟨H⟩ and ⟨K⟩ are looking happy as they walk down the street together.</p> <p>MyVLM: ⟨H⟩ and ⟨T⟩ are looking very serious as they take in the scenery.</p> <p>RAP-LLaVA(Ours): ⟨T⟩ is walking down the street with her friend ⟨H⟩.</p>
 <p>parrot1</p> <p>parrot2</p>	<p>LLaVA: Two colorful parrots are perched on a palm tree, sitting next to each other.</p> <p>LLaVA-LoRA: ⟨parrot1⟩ and ⟨parrot2⟩ are looking very serious as they perch on a palm frond.</p> <p>MyVLM: ⟨parrot1⟩ and ⟨parrot2⟩ look so cute together.</p> <p>RAP-LLaVA(Ours): ⟨parrot1⟩ and ⟨parrot2⟩ are both beautiful birds. ⟨parrot1⟩ has a bright red beak and ⟨parrot2⟩ has a light blue head and pink feet. They are both perched on a branch.</p>

Table 17. Additional qualitative comparison on image captioning between RAP-Phi3-V and other methods.


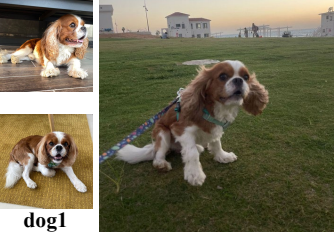

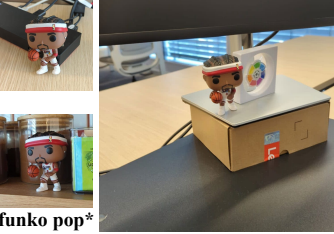
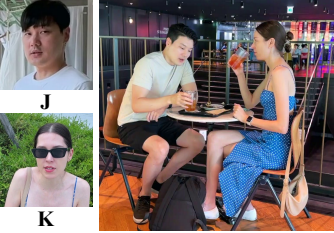

Image	Caption
 <p>sheep*</p>	<p>Phi3-V: A small white sheep figurine is sitting on a wooden table next to a plant.</p> <p>LLaVA-LoRA: <sheep*> positioned next to a potted plant on a wooden table.</p> <p>MyVLM: <sheep*> positioned within a square shape within a vase made of metal positioned on a table made of wood.</p> <p>RAP-Phi3-V(Ours): <sheep*> is ready to play on the shelf with the plant!</p>
 <p>dog1</p>	<p>Phi3-V: A brown and white dog is sitting on the grass with a leash attached to it.</p> <p>LLaVA-LoRA: <dog1> sitting on the grass with a look of determination.</p> <p>MyVLM: on a cheerful note, <dog1> restfully resides, surrounded by the soothing sounds of nature.</p> <p>RAP-Phi3-V(Ours): This is a picture of <dog1>, a cute dog with light brown and white fur. He looks happy and playful.</p>
 <p>bull*</p>	<p>Phi3-V: A bottle of wine is displayed next to a wooden cow statue.</p> <p>LLaVA-LoRA: <bull*> positioned next to a bottle of supreme cabernet sauvignon.</p> <p>MyVLM: <gold pineapple> pars the bottle of wine. A bottle of wine sits next to a bottle of wine.</p> <p>RAP-Phi3-V(Ours): <bull*> stands guard over a bottle of SUPREME wine.</p>
 <p>funko pop*</p>	<p>Phi3-V: A small figurine of a basketball player is placed on top of a box.</p> <p>LLaVA-LoRA: <funko pop*> sitting on a desk next to a box and a monitor.</p> <p>MyVLM: <funko pop*> stands out in a photo set against a dark background.</p> <p>RAP-Phi3-V(Ours): <funko pop*> is ready to dominate the court!</p>
 <p>J</p> <p>K</p>	<p>Phi3-V: A man and woman are sitting at a table in a restaurant, having a conversation and enjoying their time together.</p> <p>LLaVA-LoRA: <J> and <K> enjoying a meal together.</p> <p>MyVLM: <J> and <K> enjoying their time together.</p> <p>RAP-Phi3-V(Ours): <J> and <K> are enjoying a drink at a bar.</p>
 <p>cup*</p> <p>head*</p>	<p>Phi3-V: A small statue of a man wearing glasses is sitting next to a cup on a table.</p> <p>LLaVA-LoRA: A table hosts <cup*> and a sculpture of a head wearing glasses.</p> <p>MyVLM: A table with <cup*>, an espresso cup, and a ceramic mug depicting a face wearing glasses sits on a wooden surface.</p> <p>RAP-Phi3-V(Ours): <head*> and <cup*> are ready for their morning coffee.</p>

Table 18. Qualitative results of personalized image description obtained by RAP-LLaVA.

Image	Description
 <p>my cat</p>	<p>LLaVA: The image features a large, fluffy cat sitting on a countertop, possibly in a kitchen. The cat is positioned near a sink and appears to be resting or sleeping. The countertop is covered with a towel, providing a comfortable surface for the cat to lay on . . .</p> <p>RAP-LLaVA(Ours): The image shows my cat, a tabby cat with its eyes closed, laying on a light grey surface. The cat’s fur is a mix of light and dark brown stripes. The cat is relaxed and comfortable.</p>
 <p>toy3</p>	<p>LLaVA: The image features a small metal robot figure, possibly a toy, standing on a table or counter. The robot is positioned in the center of the scene, surrounded by a variety of leaves scattered around it. The leaves are of different sizes and . . .</p> <p>RAP-LLaVA(Ours): The image shows a toy figurine of Iron Man, a superhero from Marvel comics. He is wearing a silver and red suit of armor and is standing on a surface covered in leaves. The toy is called toy3.</p>
 <p>pillow*</p>	<p>LLaVA: The image features a chair with a pillow on it, which has a picture of a sheep on it. The chair is positioned in front of a desk with a laptop and a monitor on it. The laptop is placed on the left side of the desk, while the monitor is on the right side . . .</p> <p>RAP-LLaVA(Ours): The image shows a chair with a pillow* on it. The pillow* is a pattern of flowers, leaves, and a sheep wearing glasses with a flower crown. There is also a laptop and a computer monitor in the background.</p>
 <p>toy4</p>	<p>LLaVA: The image features a bookshelf filled with various books. On the shelf, there are two Rubik’s cubes, one green and one white, sitting next to each other. The bookshelf is well-stocked with books, occupying most of . . .</p> <p>RAP-LLaVA(Ours): The image shows a bookshelf with books on it. There are two toys on the shelf. One is a toy4, which looks like a cat with orange fuzzy ears and a brown cord attached. The other is a Rubik’s cube.</p>
 <p>toy1 toy2</p>	<p>LLaVA: The image features a bed with two stuffed animals, a dog and a cat, lying on it. The dog is positioned on the left side of the bed, while the cat is on the right side. The stuffed animals are placed close to each other, giving the impression of a cozy and . . .</p> <p>RAP-LLaVA(Ours): The image shows two plush toys, toy1 and toy2, lying in bed together. toy2 is a light blue plush with a purple patch on its head and yellow ears. toy1 is an orange plush with a yellow belly and a brown nose. They are both looking at the camera.</p>
 <p>Anya Bond</p>	<p>LLaVA: The image features a young girl hugging a large white dog in a room. The girl is wearing a black dress and appears to be enjoying her time with the dog. The dog is positioned close to the girl, and they seem to be sharing a tender moment together . . .</p> <p>RAP-LLaVA(Ours): Anya is a young girl with pink hair and big green eyes. She is wearing a black dress and black shoes. She is standing in a room with wooden floors and is hugging a white dog named Bond. Bond is wearing a black bow tie. The dog is standing on its hind legs and is looking at the camera. The room is decorated with a blue rug and a wooden cabinet . . .</p>

Table 19. Qualitative results of personalized image description obtained by RAP-Phi3-V.

Image	Description
 <p>sheep*</p>	<p>Phi3-V: The image features a display of various animal figurines on a shelf. There are three giraffes, two of which are standing and one sitting, placed next to each other. A sheep figurine is also present on the shelf, located between the giraffes . . .</p> <p>RAP-Phi3-V(Ours): The image shows a white ⟨sheep*⟩ sitting on a shelf next to a giraffe figurine. The ⟨sheep*⟩ is made of felt and has black eyes and a black nose. The giraffe figurine is made of metal and has a colorful necklace around its neck. The shelf is white . . .</p>
 <p>mug*</p>	<p>Phi3-V: The image features a wooden shelf with two wine bottles and two wine glasses. One of the wine bottles is positioned towards the left side of the shelf, while the other is located more towards the center. The wine glasses are placed on the right side of the . . .</p> <p>RAP-Phi3-V(Ours): The image shows a wooden shelf with three glassware items on it. The top glassware item is a ⟨mug*⟩ with a tiki head design. The other glassware items are wine glasses. There is a sign on the wine bottle that says "SUPREME". There is also a plant in a pot in the background.</p>
 <p>my dog</p>	<p>Phi3-V: The image features two dogs playing together in a grassy yard. One dog is located on the left side of the yard, while the other dog is on the right side. They are both running and enjoying their time outside. In the background, there are several chairs . . .</p> <p>RAP-Phi3-V(Ours): The image shows two dogs playing in a grassy backyard. The dog on the left is a white and gray dog with long fur. The dog on the right is a black and brown dog with long fur. The dogs are both running and playing on the grass. The dog on the left is named ⟨my dog⟩.</p>
 <p>toy4</p>	<p>Phi3-V: The image features a white shelf with several decorative items on it. There are three chicken-shaped decorations hanging from the shelf, each with a different design. Additionally, there are three more chicken decorations placed on the shelf, making . . .</p> <p>RAP-Phi3-V(Ours): The image shows three handmade stuffed toys. The toys are all shaped like chickens and are hanging on a white shelf. The toys are all different colors and have different designs on them. The toy on the left is called ⟨toy4⟩.</p>
 <p>character1</p> <p>character2</p>	<p>Phi3-V: The image features a group of three cartoon characters, two of which are wearing hats. They are standing together on a sidewalk, with one character positioned towards the left side of the frame, another in the center, and . . .</p> <p>RAP-Phi3-V(Ours): The image shows ⟨character1⟩ and ⟨character2⟩ standing in front of a brick wall. ⟨character1⟩ is a cartoon fox wearing a yellow button-down shirt and a purple tie. He has big eyes, fluffy fur, and a sly grin on his face. ⟨character2⟩ is a cartoon rabbit with large ears, a police officer's hat, and a red and yellow safety vest. She has a sweet, innocent look on . . .</p>
 <p>baby*</p> <p>dog2</p>	<p>Phi3-V: The image features a baby lying on a bed, with a dog resting on the floor nearby. The baby appears to be playing with a toy, possibly a puzzle, while the dog watches attentively . . .</p> <p>RAP-Phi3-V(Ours): The image shows a baby, ⟨baby*⟩, sleeping in a carrier. The baby is wrapped in a colorful blanket and is secured by a black strap. A french bulldog, ⟨dog2⟩, is sleeping next to the baby. The baby and dog are both sleeping peacefully.</p>

Table 20. Qualitative results of personalized concept recognition obtained by RAP-LLaVA. We use **green** rectangle to show the bounding box in the image.

Image	Conversation
 <p>head*</p>	<p>Question: Is $\langle \text{head}^* \rangle$ in this image?</p> <p>RAP-LLaVA: Yes, $\langle \text{head}^* \rangle$ is in the photo.</p> <p>Question: Give $\langle \text{head}^* \rangle$'s bounding box in the image.</p> <p>RAP-LLaVA: [0.57, 0.67, 0.68, 0.84]</p>
 <p>head*</p>	<p>Question: Is $\langle \text{head}^* \rangle$ in this image?</p> <p>RAP-LLaVA: No, $\langle \text{head}^* \rangle$ is not part of this photograph.</p>
 <p>toy*</p>	<p>Question: Can you identify if $\langle \text{toy}^* \rangle$ is captured in this picture?</p> <p>RAP-LLaVA: Yes, $\langle \text{toy}^* \rangle$ is visible in this picture.</p> <p>Question: Please provide $\langle \text{toy}^* \rangle$'s bounding coordinates in the image.</p> <p>RAP-LLaVA: [0.42, 0.19, 0.54, 0.39]</p>
 <p>toy*</p>	<p>Question: Can you identify if $\langle \text{toy}^* \rangle$ is captured in this picture?</p> <p>RAP-LLaVA: $\langle \text{toy}^* \rangle$ is not captured in this image.</p>
 <p>A</p>	<p>Question: Can you confirm if $\langle A \rangle$ appears in this photo?</p> <p>RAP-LLaVA: Yes, $\langle A \rangle$ is in this photo.</p> <p>Question: Describe $\langle A \rangle$'s position in the image.</p> <p>RAP-LLaVA: [0.24, 0.12, 0.72, 0.98]</p>
 <p>A</p>	<p>Question: Can you confirm if $\langle A \rangle$ appears in this photo?</p> <p>RAP-LLaVA: $\langle A \rangle$ is not present in this photograph.</p>

Table 21. Qualitative results of personalized concept recognition obtained by RAP-LLaVA. We use **green** rectangle to show the bounding box in the image.






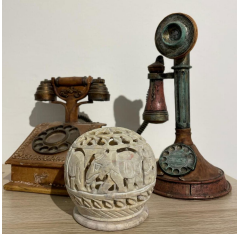












Image		Conversation
  statue1		<p>Question: Is $\langle \text{statue1} \rangle$ included in this shot?</p> <p>RAP-LLaVA: Yes, $\langle \text{statue1} \rangle$ is included in this shot.</p>
		<p>Question: Give $\langle \text{statue1} \rangle$'s bounding box in the image.</p> <p>RAP-LLaVA: [0.36, 0.16, 0.59, 0.41]</p>
  statue1		<p>Question: Is $\langle \text{statue1} \rangle$ included in this shot?</p> <p>RAP-LLaVA: $\langle \text{statue1} \rangle$ is not captured in this picture.</p>
		<p>Question: Specify $\langle \text{my dog} \rangle$'s bounding box in the photograph.</p> <p>RAP-LLaVA: [0.33, 0.23, 1.0, 0.69]</p>
  my dog		<p>Question: Is $\langle \text{my dog} \rangle$ captured in this image?</p> <p>RAP-LLaVA: Yes, I recognize $\langle \text{my dog} \rangle$ in this picture.</p>
		<p>Question: Is $\langle \text{my dog} \rangle$ captured in this image?</p> <p>RAP-LLaVA: No, $\langle \text{my dog} \rangle$ is not included in this shot.</p>
  my dog		<p>Question: Is $\langle \text{my dog} \rangle$ captured in this image?</p> <p>RAP-LLaVA: No, $\langle \text{my dog} \rangle$ is not included in this shot.</p>
		<p>Question: Is $\langle \text{teapot}^* \rangle$ visible in this image?</p> <p>RAP-LLaVA: Yes, $\langle \text{teapot}^* \rangle$ is visible in this picture.</p>
  teapot*		<p>Question: Describe $\langle \text{teapot}^* \rangle$'s position in the image.</p> <p>RAP-LLaVA: [0.52, 0.69, 0.73, 0.91]</p>
		<p>Question: Is $\langle \text{teapot}^* \rangle$ visible in this image?</p> <p>RAP-LLaVA: No, $\langle \text{teapot}^* \rangle$ is not visible in this picture.</p>
  teapot*		<p>Question: Is $\langle \text{teapot}^* \rangle$ visible in this image?</p> <p>RAP-LLaVA: No, $\langle \text{teapot}^* \rangle$ is not visible in this picture.</p>

Table 22. Instructions for visual grounding and recognition.

Visual recognition:

- Is ⟨V⟩ visible in this picture?
- Is ⟨V⟩ in this image?
- Do you see ⟨V⟩ in the photo?
- Is ⟨V⟩ present in this photograph?
- Can you identify if ⟨V⟩ is captured in this picture?
- Is ⟨V⟩ depicted in this image?
- Does the picture feature ⟨V⟩?
- Can you confirm if ⟨V⟩ appears in this photo?
- Is ⟨V⟩ included in this shot?
- Is ⟨V⟩ shown in this image?
- Can you tell if ⟨V⟩ is part of this photograph?
- Is there any sign of ⟨V⟩ in this picture?
- Can you detect ⟨V⟩ in the photo?
- Is ⟨V⟩ captured in this image?
- Do you recognize ⟨V⟩ in this picture?

Visual grounding:

- Give ⟨V⟩'s bounding box in the image.
- Describe ⟨V⟩'s position in the image.
- Please provide the coordinates of the bounding box for ⟨V⟩ in the given image.
- Specify the rectangular boundaries of ⟨V⟩ in the image.
- Give ⟨V⟩'s position in the following image.
- Please provide ⟨V⟩'s bounding coordinates in the image.
- Indicate the bounding box for ⟨V⟩ in the image.
- Show the bounding box for ⟨V⟩ in the picture.
- Specify ⟨V⟩'s bounding box in the photograph.
- Mark ⟨V⟩'s bounding box within the image.

Table 23. Instructions for image captioning.

Image caption:

- Give a caption of the image.
- Give a personalized caption of this image.
- Provide a brief caption of the image.
- Summarize the visual content of the image.
- Create a short caption of the image.
- Offer a short and clear interpretation of the image.
- Describe the image concisely.
- Render a concise summary of the photo.
- Provide a caption of the given image.
- Can you provide a personalized caption of this photo?
- Could you describe this image concisely?

Table 24. Instructions for image description.

Image description:

- Describe the image.
- Give a description of the image.
- Give a description of the image in detail.
- Give a short description of the image.
- Describe the image in detail.
- Please provide a description of the image.
- Can you give me details about the image?
- Could you explain what's shown in the image?

Table 25. Seed questions used for question answering synthesis.

Person:

- What is $\langle H \rangle$'s hair color?
- What is $\langle H \rangle$'s height (estimated)?
- What is $\langle H \rangle$'s skin tone?
- What is $\langle H \rangle$'s eye color?
- What style of clothing is $\langle H \rangle$ wearing?
- Does $\langle H \rangle$ have any visible tattoos?
- Does $\langle H \rangle$ wear glasses or contact lenses?
- Does $\langle H \rangle$ have any facial hair?
- What is $\langle H \rangle$'s approximate age?
- What is $\langle H \rangle$'s build or body type?
- What is $\langle H \rangle$ doing?






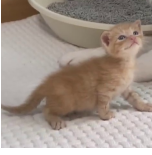

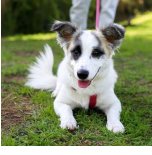

Object:

- What color is $\langle O \rangle$?
- What pattern is on $\langle O \rangle$?
- What shape does $\langle O \rangle$ have?
- What size is $\langle O \rangle$?
- What is the texture of $\langle O \rangle$?
- Is $\langle O \rangle$ shiny or matte?
- What material is $\langle O \rangle$ made of?
- Does $\langle O \rangle$ have any patterns or designs on it?
- Is $\langle O \rangle$ new or worn?
- Does $\langle O \rangle$ have any visible brand or logo?
- Is $\langle O \rangle$ functional or decorative?

Multi-concept question:

- What do $\langle C_1 \rangle$ and $\langle C_2 \rangle$ have in common?
- What activity are $\langle C_1 \rangle$ and $\langle C_2 \rangle$ engaged in?
- Where could $\langle C_1 \rangle$ and $\langle C_2 \rangle$ be located?
- What is the most noticeable difference between $\langle C_1 \rangle$ and $\langle C_2 \rangle$?
- What are they doing?

Table 26. Examples of our database. A concept should be provided with an image and its personalized description.

Image	Concept	Information
	Anya	A young girl with pink hair and big green eyes.
	doll*	This is a cute figurine of a girl wearing a pink and blue dress, holding a white bubble.
	toy1	A plush toy. It is orange with a yellow belly and a brown nose.
	toy2	This is a plush toy of the bluey character. It is a light blue color with a purple patch on its head, and its ears are yellow.
	statue*	This is a figurine of a cat. The cat has a blue body with yellow, red, and green stripes and a long tail that is also striped.
	cat*	A small ginger kitten with bright blue eyes looks up at the camera.
	H	A young man is wearing a plain tan t-shirt. His hair is short and curly.
	my dog	A white and gray dog with long fur. He has black eyes.
	T	A young woman with blonde hair is wearing a white tank top and blue jeans.