Training-Free Personalization via Retrieval and Reasoning on Fingerprints

Supplementary Material

8. Supplementary Overview

In this supplementary, we provide a detailed statistics of the PerVA dataset in Section 8.1, qualitative example R2P demonstrating our proposed Attribute Focused CoT and Pairwise Image Matching in Section 8.2, details on prompting in Section 8.3 and analysis of the proposed R2P on MyVLM and Yo'LLaVA datasets in Section Sec. 8.4.

8.1. PerVA Dataset Overview and distribution

The PerVA dataset comprises 21 object categories with a total of 67,482 images distributed across 329 unique concepts. The dataset is structured to support training-free retrieval-augmented generation (RAG) approaches, with each concept represented by multiple training images and a smaller set of test images.

The dataset follows a balanced concept distribution across train and test splits, with identical concept coverage (329 concepts each) but different image quantities per concept. The training split contains 59,392 images with an average of 180.5 images per concept, while the test split contains 8,090 images with an average of 24.6 images per concept.

For our RAG-based approach, we constructed a reference database containing exactly one representative image per concept (329 images total), ensuring complete concept coverage across all categories.

Category	Concepts	Train Images	Test Images	DB Images
decoration	62	12,935	2,011	62
retail	69	12,678	1,750	69
clothe	30	5,437	873	30
bag	25	4,534	426	25
veg	20	3,419	604	20
plant	17	2,971	485	17
toy	13	2,440	136	13
book	12	2,018	140	12
cup	11	1,691	396	11
pillow	9	1,584	174	9
tumbler	8	1,496	296	8
bowl	7	1,186	83	7
towel	7	1,138	83	7
remote	7	1,047	79	7
plate	7	1,013	127	7
bottle	7	978	102	7
tie	6	885	53	6
tro_bag	5	912	169	5
umbrella	3	536	65	3
headphone	2	259	22	2
telephone	2	235	16	2
Total	**329**	**59,392**	**8,090**	**329**

Table 6. Category-wise distribution of the PerVA dataset

The dataset exhibits natural category imbalance reflecting real-world object distributions, with decoration and retail categories containing the most concepts (62 and 69, respectively), while telephone and headphone categories contain the fewest (2 each). This distribution provides a realistic testbed for evaluating retrieval-based recognition systems across diverse object categories with varying levels of intra-category diversity.

8.2. Additional qualitative example

In Figure 5, we qualitatively demonstrate the Concept Inference with Retrieval-Reasoning of R2P. Given a query about a personalized object (in this case, a cartoon cat) and the Top-K retrieved concepts along with their descriptions, R2P first performs attribute-focused Chain-of-Thought (CoT) reasoning on the fingerprint attributes of each retrieved concept. The model is instructed to compare the unique attributes between the query image and each description, then reason over them to identify the correct concept name. However, in this case, the model hallucinated the attribute 'pink bow on head' for the query, which is a fingerprint attribute for option A (i.e., marie-cat), leading to misclassification. Since attribute verification failed here, R2P proceeds to perform more computationally expensive pairwise reasoning, correctly identifying the concept name, which it then uses to generate a personalized caption.

8.3. Prompting

In this section, we define the prompts used for our personalization task. Figure 6 shows the prompt template for creating the personal database. Here, we provide the model with the image, object category, and concept name, prompting it to identify the distinct attributes that make the object unique, compared to similar objects in the same category.

Figure 7 illustrates the prompt template for attribute-focused CoT reasoning. The model is shown the query image along with the retrieved concepts from the database. For each retrieved concept, the VLM is asked to compare the query image with its description and list the attributes that are common between the two. If no shared attributes are found, the model may omit mentioning any. Based on the matching attributes, the model carefully selects the option that best describes the query image.

Finally, Figure 8 shows the prompt for multimodal pairwise reasoning. The model is provided with the query image, the concept reference image, and the associated description. The model is prompted to compare the two images and determine whether they match. The description of the reference image is included to help the model identify relevant attributes in the query image, enabling a more informed deci-

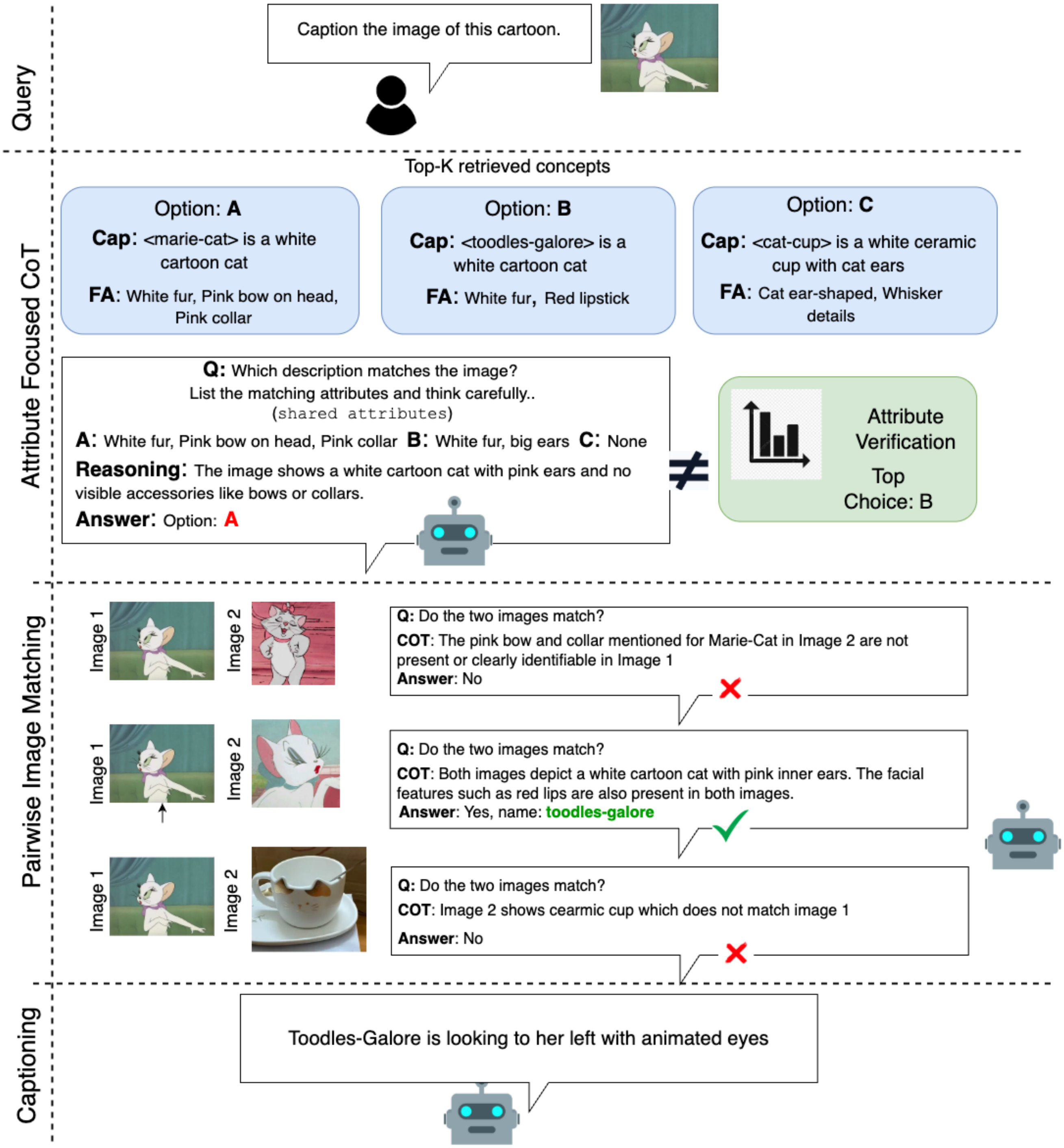
Query
Caption the image of this cartoon.
Attribute Focused CoT
Top-K retrieved concepts
Option: A
Cap: <marie-cat> is a white cartoon cat
FA: White fur, Pink bow on head, Pink collar
Option: B
Cap: <toodles-galore> is a white cartoon cat
FA: White fur, Red lipstick
Option: C
Cap: <cat-cup> is a white ceramic cup with cat ears
FA: Cat ear-shaped, Whisker details
Q: Which description matches the image?
List the matching attributes and think carefully..
(shared attributes)
A: White fur, Pink bow on head, Pink collar B: White fur, big ears C: None
Reasoning: The image shows a white cartoon cat with pink ears and no visible accessories like bows or collars.
Answer: Option: A
≠
Attribute Verification
Top Choice: B
Pairwise Image Matching
Image 1
Image 2
Q: Do the two images match?
COT: The pink bow and collar mentioned for Marie-Cat in Image 2 are not present or clearly identifiable in Image 1
Answer: No
Image 1
Image 2
Q: Do the two images match?
COT: Both images depict a white cartoon cat with pink inner ears. The facial features such as red lips are also present in both images.
Answer: Yes, name: toodles-galore
Image 1
Image 2
Q: Do the two images match?
COT: Image 2 shows cearmic cup which does not match image 1
Answer: No
Captioning
Toodles-Galore is looking to her left with animated eyes

Figure 5. Qualitative example of the Concept Inference with Retrieval-Reasoning of R2P

sion. Please note that the prompts shown are for demonstration purposes only. In practice, we populate the descriptions of retrieved objects directly from the database.

8.4. Additional Results

In this section we report additional ablation analysis on existing personalization datasets in the literature, namely MyVLM [4] and Yo'LLaVA [29].

Tab. 7 and Tab. 9 reports the ablation analysis of the pro-

Describe the $< \ g_i \ >$ in the image identified by the concept-identifier $< c_i >$ and highlight what makes it unique.
Your response MUST be in valid JSON format and must follow EXACTLY the format below:
{
general: a brief description of the object in one sentence.,
category: category of the object,
distinct features: [List of distinct features that makes the object unique],
}
IMPORTANT:
- List only the most distinguishing features that set this object apart.
- Avoid generic descriptions that apply to every object in this category.
- Do not include any extra text or commentary. Any deviation from this format will be considered incorrect.

Figure 6. Prompt Template for Personal Database Creation

Pairwise Reasoning	Fingerprint Attributes	Reasoning CoT	Recognition Wtd	Captioning Recall
✗	✗	✗	96.1	87.4
✗	✗	✓	94.2	88.8
✓	✗	✗	**98.7**	88.2
✓	✗	✓	<u>97.6</u>	<u>89.6</u>
✓	✓	✓	97.4	**91.5**
✓	privileged	✓	97.8	91.1

Table 7. Ablation on pairwise reasoning, fingerprint attributes, and the use of CoT reasoning for R2P in terms of weighted recognition metrics (↑) and captioning recall (↑) on MyVLM [4]. We include a privileged version of our approach with human pre-defined fingerprint attributes for concepts.

posed approach on the use of pairwise-reasoning, fingerprint attributes and CoT reasoning of the underlying VLM. Results are mostly consistent with the main paper results on PerVA. Notably, we observe how in both the considered settings R2P outperforms the privileged approach relying on human-knowledge pre-defined attributes, showcasing how the VLM effectively relies on its knowledge to predict a discriminative attribute fingerprint.

Tab. 8 and Tab. 10 report the analysis on different verification strategies, following the experimentation reported in the main paper. MyVLM [4] and Yo'LLaVA [29] datasets are considered, respectively. Results show that R2P multimodal verification step outperforms the other compared strategies.

Finally, in Tab. 11 (MyVLM [4]) and Tab. 12 (Yo'LLaVA [29]) consistent observations on the effectiveness of the proposed multimodal concept retrieval are reported.

You are a helpful AI agent specializing in image analysis and object recognition.
Your task is to analyze and compare a query image with three provided descriptions.
Below are the description(s)

A. Name: $< c_{i_1} >$,
Info: {general: A generic description about $< c_{i_1} >$,
category: category of $< c_{i_1} >$,
distinct features: [distinct feature 1, distinct feature 2, ...]}

B. Name: $< c_{i_2} >$,
Info: {general: A generic description about $< c_{i_2} >$,
category: category of $< c_{i_2} >$,
distinct features: [distinct feature 1, distinct feature 2, ...]}

C. Name: $< c_{i_3} >$,
Info: {general: A generic description about $< c_{i_3} >$,
category: category of $< c_{i_3} >$,
distinct features: [distinct feature 1, distinct feature 2, ...]}

Your Task:
- Compare the query image with each description and answer the following question:
Which description matches the subject in the image?
Answer in A, B, C.
- List shared attributes between the image and each description very concisely
- If no attributes match for a certain, generate None
- Provide a brief reasoning for your final answer.
- Respond strictly in the following JSON format:
{
"A": "[Matching attributes for option A]",
"B": "[Matching attributes for option B]",
"C": "[Matching attributes for option C]",
"Reasoning": "<Brief justification>",
"Answer": "<one of A, B, C>"
}

Any deviation from this format will be considered incorrect.
Do not output any additional text.

Figure 7. Prompt template for Attribute-focused CoT reasoning

Method	Captioning Recall
Pairwise-reasoning	91.3
No estimation	90.5
Abstention	90.4
Logits-based	90.8
Attr. Verification **(Ours)**	**91.5**

Table 8. Ablation on verification strategies for R2P on MyVLM [4] dataset. Performance is evaluated based on captioning recall (↑).

You are a helpful AI agent specializing in image analysis and object recognition.
You are given two images: **Image 1** and **Image 2**.
Additionally, the name and a textual description of the object in **Image 2** is also provided below:

1. Name: $< c_i >$,
Info:
{general: A generic description about $< c_i >$,
category: category of $< c_i >$,
distinct features: [distinct feature 1, distinct feature 2, ...]}
Task:
- Compare the two images and answer the following question.
Can you see $< c_i >$ in this Image 1?
Answer with a single word, either yes or no.
- Provide your reasoning based on the two images and the given description.
- Generate your response with JSON format:
{
"Reasoning": "<Your reasoning in 2-3 sentences.>",
"Answer": "<yes or no>"
}
Output only the JSON response. DO NOT output any additional text.

Figure 8. Prompt Template for Image based pairwise comparison

Pairwise Reasoning	Fingerprint Attributes	Reasoning CoT	Recognition Wtd	Captioning Recall
✗	✗	✗	92.4	73.8
✗	✗	✓	89.1	76.8
✓	✗	✗	93.5	83.4
✓	✗	✓	**95.4**	81.5
✓	✓	✓	94.4	**87.1**
✓	privileged	✓	96.7	84.3

Table 9. Ablation on pairwise reasoning, fingerprint attributes, and the use of CoT reasoning for R2P in terms of weighted recognition metrics (↑) and captioning recall (↑) on Yo'LLaVA [29]. We include a privileged version of our approach with human pre-defined fingerprint attributes for concepts.

Method	Captioning Recall
Pairwise-reasoning	88.1
No estimation	82.0
Abstention	85.6
Logits-based	83.4
Attr. Verification **(Ours)**	**91.5**

Table 10. Ablation on verification strategies for R2P on Yollava [29] dataset. Performance is evaluated based on captioning recall (↑).

.

Embedding	H@1	H@3	H@5	H@10
DINOv2	64.7	80.3	86.5	93.8
CLIP-Image	88.2	96.2	98.2	99.4
CLIP-Text	80.5	95.6	98.1	98.7
Multi-modal (2-step)	92.2	98.2	98.8	99.4
R2P **(Ours)**	**93.5**	**98.5**	**99.4**	**100**

Table 11. Performance with different retrieval strategies evaluated in terms of HIT@K (↑) on MyVLM [4] dataset.

Embedding	H@1	H@3	H@5	H@10
DINOv2	67.3	83.5	88.3	94.3
CLIP-Image	82.9	93.4	94.9	100
CLIP-Text	64.8	85.8	94.7	99.4
Multi-modal (2-step)	**92.2**	98.2	98.8	99.4
R2P **(Ours)**	84.7	**99**	**99.6**	**100**

Table 12. Performance with different retrieval strategies evaluated in terms of HIT@K (↑) on Yo'LLaVA [29] dataset.