

Figure A.1: Visualization of a more complex and diverse multi-concept image captioning result using our proposed RePIC. The generated query image is sourced from InstantFamily [8].

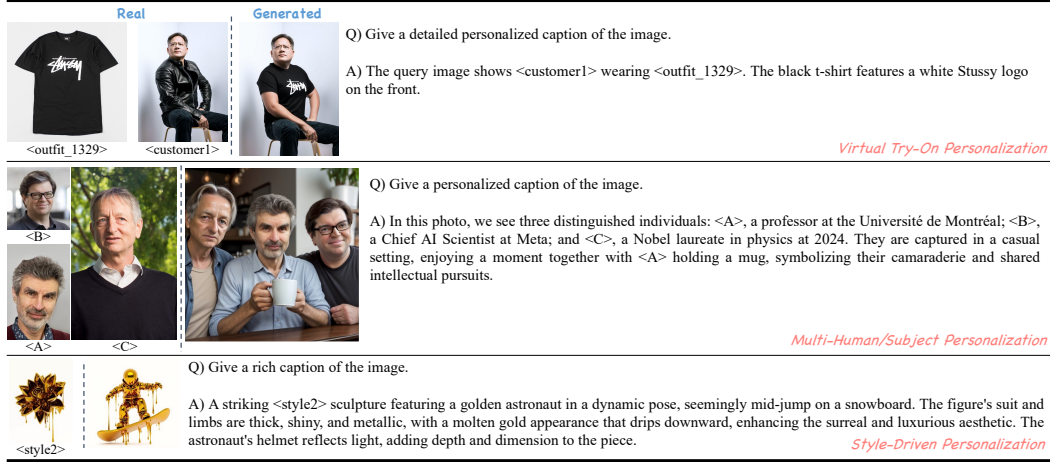


Figure A.2: Visualization of RePIC results on various personalized image captioning tasks.

A Additional Qualitative Results

A.1 More Challenging Multi-Concept Setting

In this section, we further present the results of our method in the 7-concept setting, as illustrated in Figure A.1. Notably, our approach generates faithful descriptions of the query image while accurately referencing the corresponding reference images and their associated information. In Figure A.2, we demonstrate that our proposed method can faithfully perform image captioning on synthetic images across various tasks, including virtual try-on [3], multi-human or subject personalization [6], and style-driven personalization [2], which is compatible with the state-of-the-art personalized image generation benchmarks. These results highlight the superiority of our approach in handling diverse personalized image captioning tasks with MLLM.

A.2 Visualizations on Image Captioning Quality

In Figure A.3, we present qualitative comparisons of image captioning quality in both single and multi-concept settings, highlighting the effectiveness of our proposed method. Note that the RAP-

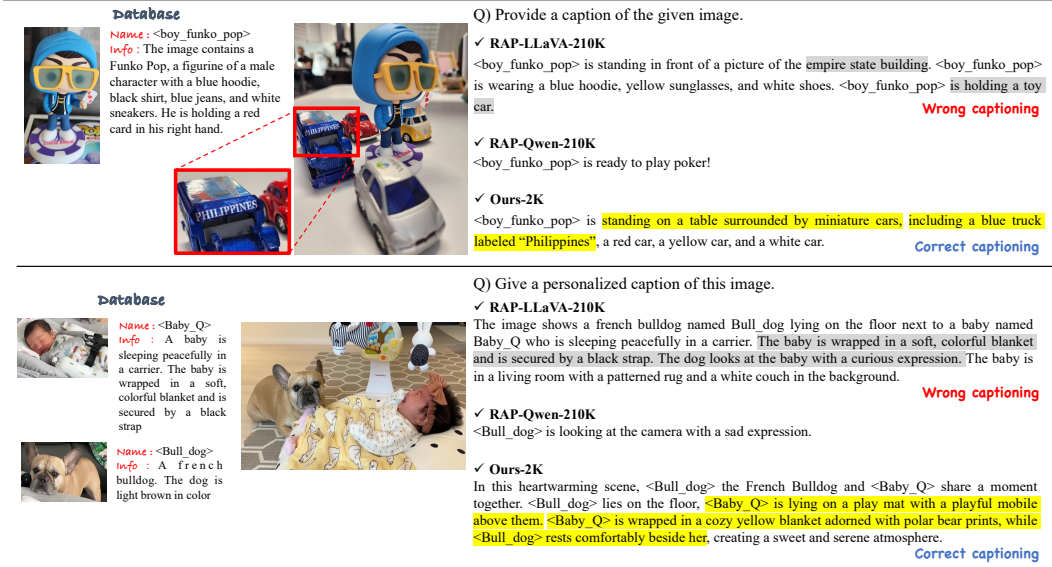


Figure A.3: Examples of generated captions on single and 2-concept personalized image captioning tasks.

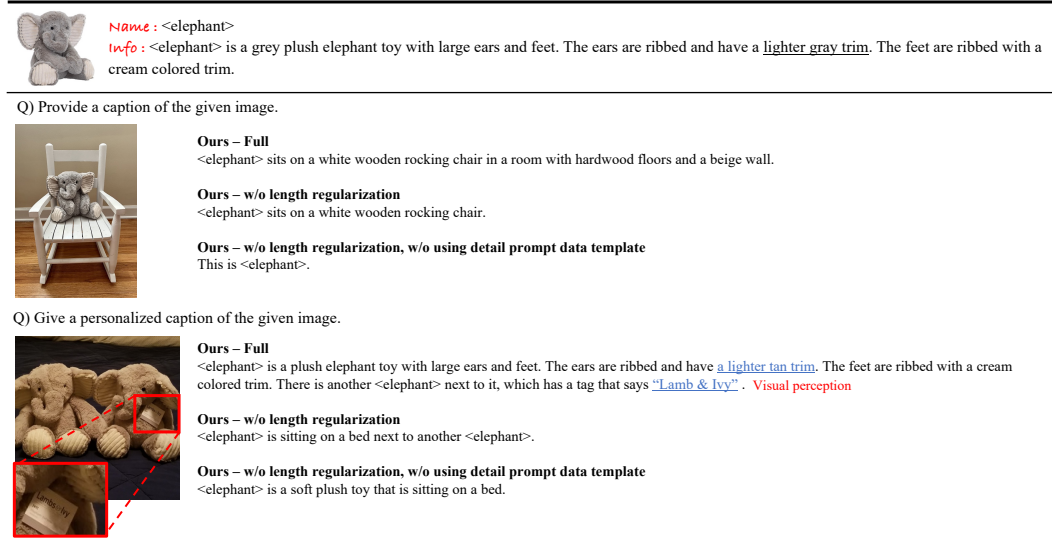


Figure A.4: Visualization of qualitative results for additional components used for our methods.

14 LLaVA often merely duplicates the retrieval information or generates visual hallucinations without
 15 considering vision perception, and RAP-Qwen severely fails to caption correctly. In contrast, only
 16 the proposed approach faithfully and concretely describes the given query image.

17 A.3 Effects of Additional Components for Preserving Captioning Quality

18 In Figure A.4, we present additional ablation results evaluating the impact of applying length
 19 regularization and incorporating detailed prompts in the training dataset. The results indicate that in-
 20 corporating length regularization and detailed prompts effectively mitigates generating uninformative
 21 captions for the query images.

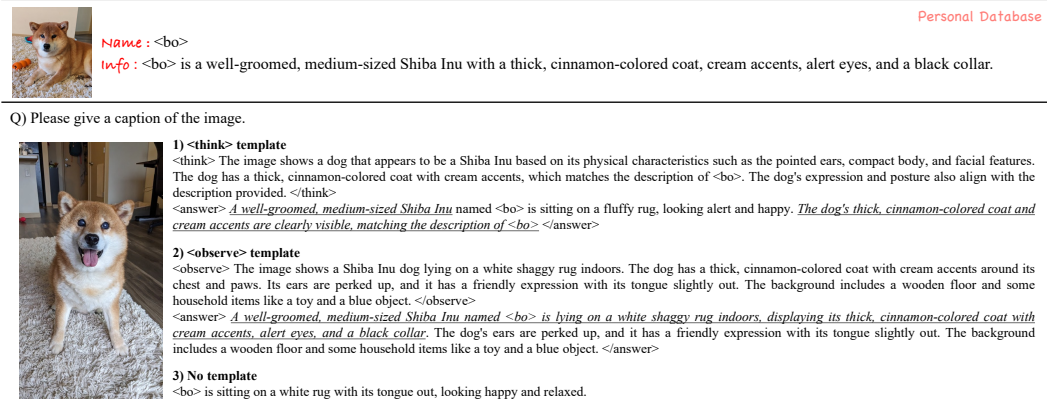


Figure A.5: Visualizations of output responses with and without the use of reasoning templates.

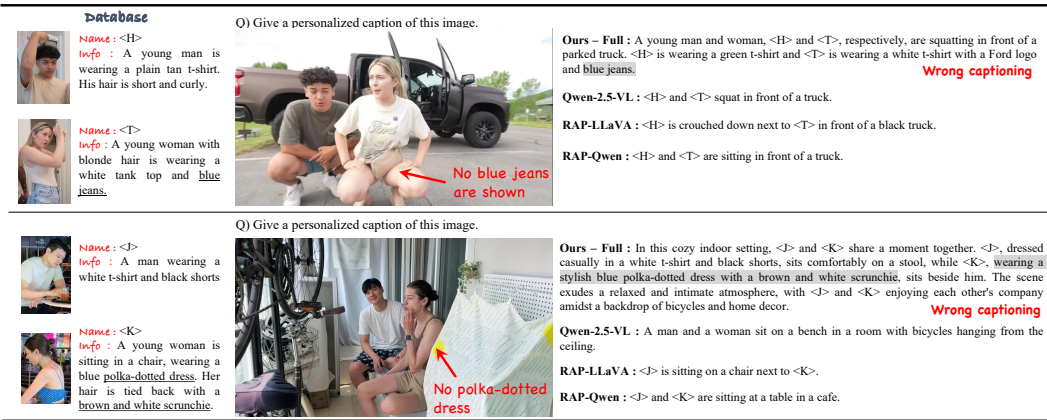


Figure A.6: Examples illustrating additional limitations of RePIC in 2-concept scenario.

22 A.4 Effect of Reasoning Templates

23 We consider the post-tuned model trained with reasoning templates such as <think> and <observe>
 24 to verify the effectiveness of visual reasoning in personalized tasks, which has become a prevalent
 25 choice [9, 7, 13] for MLLM post-training with RL. However, in Figure A.5, we observe that using
 26 reasoning templates often results in longer outputs that fail to faithfully describe the query image.
 27 In contrast, omitting templates leads to more concise yet accurate and faithful image descriptions.
 28 Furthermore, as demonstrated in our main analysis, reasoning templates negatively impact personal
 29 grounding performance in personalized image captioning tasks.

30 A.5 Further Limitations of RePIC

31 We illustrate the limitations of our RePIC model on the personalized image captioning task in
 32 Figure A.6. In the first row, RePIC incorrectly captions the image with blue jeans, despite no such
 33 item being present. A similar issue is observed in the second row, where the model references a
 34 polka-dotted dress that does not appear in the query image. These examples show a limitation
 35 of RePIC in generating accurate personalized captions, primarily due to insufficient fine-grained
 36 visual perception. For instance, it struggles when objects are not visibly present (e.g., no blue jeans
 37 appear) or when the reference and query images differ significantly (e.g., back view vs. front view),
 38 making it difficult to recognize them as the same person or the same object. We expect that these
 39 limitations can be mitigated either by constructing a high-quality database for each concept—avoiding
 40 the use of personal information based solely on the visual appearance of the image, and ensuring the
 41 reference image clearly shows a front view of the object—or by leveraging an MLLM equipped with
 42 an advanced vision encoder and a more powerful backbone LLM, such as Qwen-3 [14].

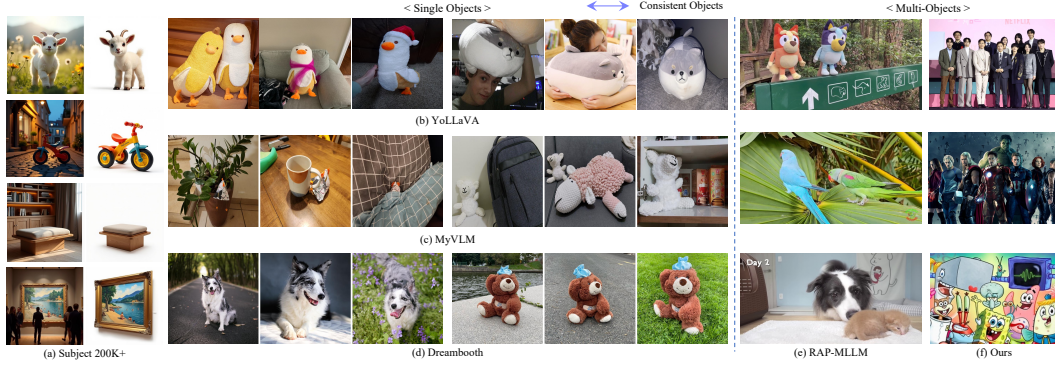


Figure A.7: Datasets used for training and evaluation. Note that the Subject200K+ dataset (a) was used for training, while all real datasets (b) to (f) were used only for evaluation.



Figure A.8: Visualization of the DreamBooth database constructed in this work.

43 B Additional Experimental Configurations

44 B.1 Experimental Details

45 Our implementation is based on the open-source codebase¹. To train our model, we set LoRA rank as
 46 64, LoRA alpha as 128, and use the number of generations per prompt as 8.

47 B.2 Used Datasets for Evaluation

48 The data configuration used for both training and evaluation in our experiments is detailed in
 49 Figure A.7. Notably, the Subject200K+ dataset was used exclusively for post-training and was not
 50 included in the evaluation. All other real-image benchmarks were used for evaluation purposes.
 51 In Figure A.8, we present the configuration of our curated DreamBooth [12] database used for
 52 single-concept captioning evaluation in our experiment.

¹<https://github.com/om-ai-lab/VLM-R1>

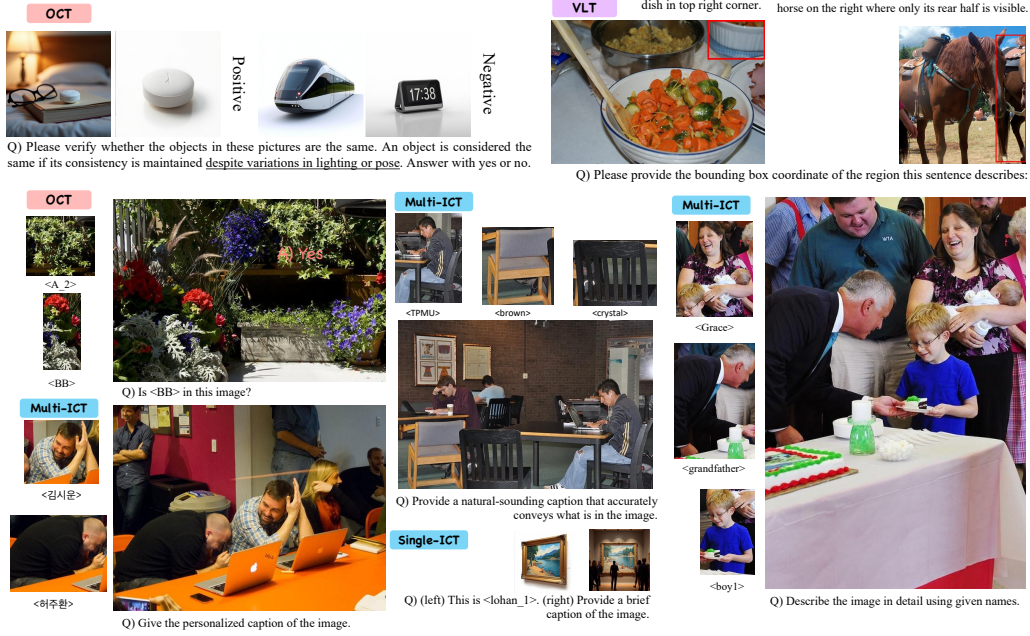


Figure A.9: Visualization of data templates used for MLLM post-training, including examples of OCT, VLT, single-ICT, and multi-ICT.

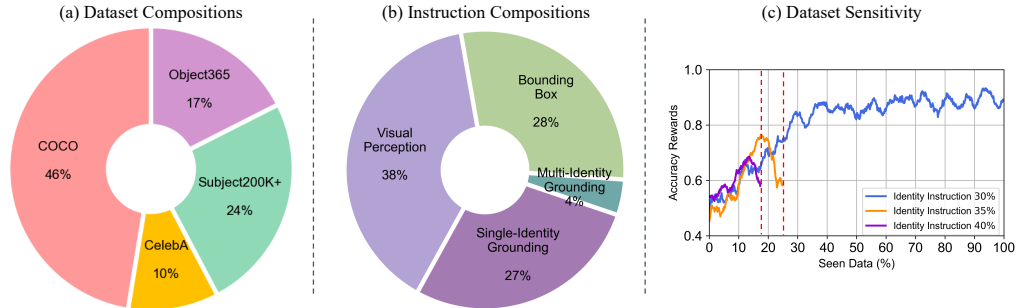


Figure A.10: (a) Dataset composition, (b) instruction composition, and (c) the sensitivity to the proportion of identity grounding instructions within the overall training set.

Table A.1: Used reasoning templates

<think> Reasoning Template:

- First output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags.

<observation> Reasoning Template:

- First, observe carefully and enclose the observation process in <observe> </observe> tags and then output the final answer in <answer> </answer> tags.

B.3 Used Data Templates

In Figure A.9, we illustrate the data and instructions for verifiable rewards of OCT, VLT, and ICT used for post-training.

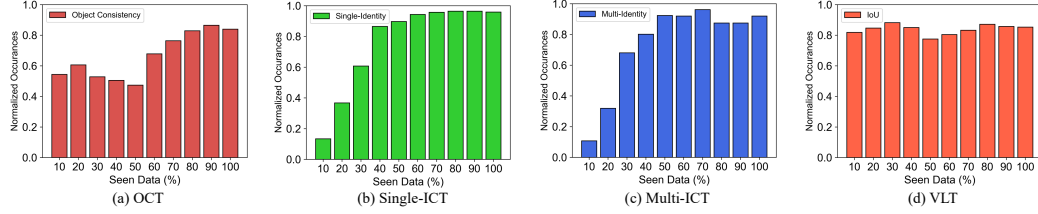


Figure A.11: Distributions of mean verifiable rewards during training for each task: (a) OCT, (b–c) ICT, and (d) VLT.

B.4 Dataset Compositions

For generating `<name>`, we use a random name generator² to sample human or object names in an on-the-fly manner. Table A.1 shows the two reasoning templates of using `<observe>` and `<think>` tokens used for our ablation studies involving special tokens. These templates were appended directly before each captioning query.

Figure A.10 illustrates how we construct a high-quality dataset for personal grounding. In (a), we show that the dataset is composed of COCO, Objects365, CelebA, and Subject200K+. (b) visualizes the instruction composition, which includes OCT, VLT, single-ICT, and multi-ICT. We note that approximately 31% of the total training data is composed of single and multi-ICT samples. In (c), we highlight the dataset sensitivity for convergence. We observe that if the amount of ICT instruction in training data is too high, the RL training often fails. This demonstrates that while RL is inherently data-efficient, it is sensitive to data quality. Overall, our findings highlight the importance of a well-structured instruction dataset for effective RL-based post-training in MLLM personalization.

B.5 Details On Retrieval Setting

Following the previous work [5], for a retrieval setting, we first utilize a database of image, text pairs representing user-specific concepts. Given database images, each image is first processed using a pre-trained CLIP [11] encoder to obtain visual embeddings. Then, for a given query image and its corresponding textual instructions, YOLO-World [1] is employed to detect regions of interest. Thus, cropped images are encoded into embeddings, and by computing Euclidean distances between these embeddings and pre-stored embeddings, the most relevant reference images are retrieved from the database.

C Additional Analyses

C.1 How Efficiently Does RL Maximize Rewards During Post-Training?

Figure A.11 illustrates how efficiently our proposed method achieves personalization of the model. To analyze this, we divide the sections with the criterion of seen data during training into bins and count the number of responses with a verifiable reward of 1 within each bin. These counts are then normalized by the total number of responses that include both rewards of 0 and 1, which we call this score as normalized occurrence. The results show a clear upward trend in performance across both OCT, single and multi-ICT, once the proportion of seen data exceeds 50%. Here, the total number of seen data is 2K. Notably, ICTs both begin with a low occurrence rate of approximately 0.2 but show a sharp emergence towards 1.0 once the seen data surpasses 50% (*i.e.*, 1K samples). These results suggest that our method effectively guides MLLM personalization in a data-efficient and effective manner, armed with our carefully designed verifiable rewards, data construction, and instruction compositions. Note, VLT shows relatively stable performance regardless of the amount of seen data.

²<https://faker.readthedocs.io/en/master/>

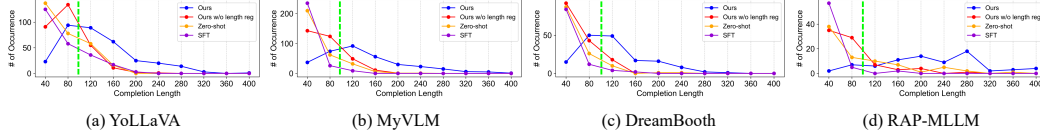


Figure A.12: Ablation studies on output length distributions of image captioning across single and multi-concept evaluation datasets.

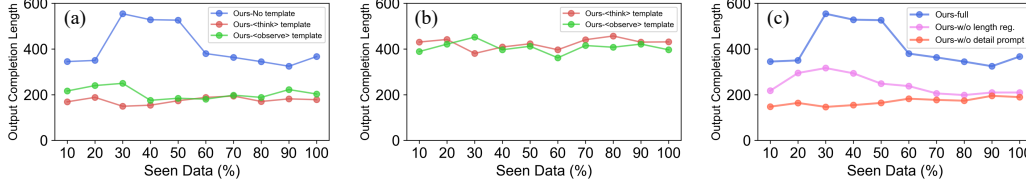


Figure A.13: Visualization of results: (a) measured output response length (e.g. between `<answer>` and `</answer>` tokens), (b) output length measured within the reasoning template (e.g. between `<think>` and `</think>` tokens), and (c) ablation studies.

90 C.2 Can Length Regularization Reward Guides To Prolong Output Completions?

91 Figure A.12 presents ablation results on output completion lengths of the image captioning task across
 92 the evaluation datasets. In all cases, applying our length regularization proves as a simple yet effective
 93 strategy for increasing output lengths, consistently yielding longer completions, surpassing those
 94 generated by both zero-shot and SFT (i.e., RAP-Qwen) baselines, which often generate uninformative
 95 captions such as ‘This is <name>’.

96 C.3 Does Reasoning Template Matter for Personalization?

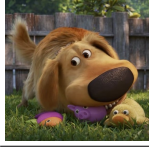
97 We further conduct an experiment to investigate whether the reasoning templates have a meaningful
 98 impact on personalization. Specifically, we examine the effects of using special tokens and measure the
 99 output completion length when no reasoning template is used. Interestingly, as shown in Figure A.13,
 100 our ablation results reveal the following: (a) Even without a reasoning template, the model is capable
 101 of producing sufficiently long and informative answer responses. (b) In contrast, when using a
 102 reasoning template, we observe a similar tendency toward overthinking [10], which is a decline in
 103 accuracy as the average reasoning time increases, where the model focuses primarily on the reasoning
 104 process at the expense of informative answers. Thus, in the same context as our experiments in the
 105 main paper, for post-training with RL, eliminating the reasoning template contributes to enhancing
 106 the personal grounding.

107 C.4 Does RePIC Enhance The General Image Captioning Ability of MLLM?

108 In this section, we compare the captioning performance of our proposed method with the zero-shot
 109 baseline on a general image captioning task. The evaluation does not consider both skip-retrieval
 110 and retrieval settings, as it focuses solely on captioning a single query image using general prompts
 111 without any reference images.

112 As a result, in Figure A.14, our method consistently generates more faithful and accurate descriptions
 113 for the image compared to the zero-shot model under general query settings. In Figure A.15, we
 114 further compare results using detailed query prompts. In this case, the results of both our method
 115 and the zero-shot model show nearly equivalent performance in caption generation. This suggests
 116 that RL-based post-training does not enhance a model’s ability to perform detailed image captioning
 117 beyond what the zero-shot model can already achieve. Rather, our RL-based post-training method
 118 reinforces the frequency of more faithful and preferable captions in the output under general query
 119 prompt settings. These observations align with the results reported in concurrent studies [15, 4], and
 120 quantitative results for the preference evaluations with GPT4o across the MyVLM, YoLLaVa, and
 121 DreamBooth datasets are presented in Figure A.16.

Q) Can you provide a caption of this photo?



Ours

A cheerful animated dog with a big smile and expressive eyes lies on a grassy lawn, surrounded by colorful toys, [including a purple pig and a yellow duck](#). The scene exudes warmth and playfulness.

Zero-Shot

A cheerful animated dog enjoys some playtime with colorful toys in a grassy backyard setting.

YoLLaVA Dataset

Q) Give a caption of the image.



Ours

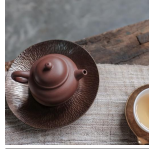
A playful and quirky scene inside a car, [featuring a small figurine with glasses sitting on the dashboard](#), looking out through the windshield. The interior of the vehicle includes a modern steering wheel with a [Kia logo](#), and the view outside shows a mix of greenery and a textured wall.

Zero-Shot

Steering into adventure with a quirky companion behind the wheel!

MyVLM Dataset

Q) Give a caption of the image.



Ours

A serene tea setting featuring [a traditional clay teapot and saucer](#), accompanied by a delicate cup filled with tea, all resting on a rustic wooden surface. The scene evokes a sense of calm and tradition.

Zero-Shot

A serene tea setting featuring a traditional clay teapot and saucer, accompanied by a delicate cup of tea on a rustic wooden surface.

DreamBooth Dataset

Q) Provide a caption of the given image.



Ours

A vibrant scene featuring three parrots perched on a railing in a lush, green outdoor setting. The birds include a [bright yellow parrot on the left](#), a [green parrot with a pink beak in the center](#), and a [blue parrot on the right](#). The background showcases a dense forest, adding depth and natural beauty to the image.

Zero-Shot

Three colorful parrots perched on a railing, showcasing vibrant hues against a lush green backdrop.

RAP-MLLM Dataset

Figure A.14: Visualization of image captioning results for general query prompts.

Q) Describe the image in detail.



Ours

The image features a person standing outdoors, smiling warmly at the camera. They are wearing a [light blue cap with a small logo on the front and glasses](#). The individual is dressed in a [white, short-sleeved button-up shirt](#) with a relaxed fit. [A gray jacket is casually draped over their left arm](#). The background showcases [vibrant red flowers, likely salvia, with green foliage](#), suggesting a garden or park setting. The lighting appears natural, [indicating daytime](#), and the overall atmosphere is cheerful and serene.

Zero-Shot

The image features a person standing outdoors, smiling at the camera. They are wearing a [light blue cap with a small logo on the front and glasses](#). The individual is dressed in a [white, short-sleeved, button-up shirt](#) that appears to be made of a lightweight fabric. [They have a gray jacket draped over their left arm](#). The background is lush with [vibrant red flowers, likely salvia, and green foliage](#), suggesting a garden or park setting. The lighting indicates [it might be daytime](#), possibly late afternoon given the softness of the light. The overall atmosphere is cheerful and relaxed.

YoLLaVA Dataset

Q) Describe the image in detail.



Ours

The image depicts an anime-style scene featuring two characters. On the left, there is a [large, fluffy white bear with a friendly expression, showing its snout and part of its body](#). The bear's fur appears soft and detailed, with a slightly darker shade around its mouth area. On the right, there is a [young girl with vibrant pink hair and green eyes](#). She has a cheerful expression, with a wide smile and bright, lively eyes. Her hair is styled in a [bob cut](#), and she is wearing a school uniform that includes [a brown jacket with gold accents, a white shirt, and a red tie](#). The background suggests they are indoors, possibly in a hallway or room with wooden paneling visible behind the characters. [The overall atmosphere of the image is warm and friendly.](#)

Zero-Shot

The image features two characters from an anime-style illustration. [On the left, there is a large, fluffy white bear with a friendly expression, showing its snout and part of its body.](#) The bear's fur appears soft and detailed, with a slightly darker shade around its mouth area. On the right, there is a [young girl with vibrant pink hair and green eyes](#). She has a cheerful expression, with a wide smile and bright, lively eyes. Her hair is styled in a [bob cut](#) with bangs that frame her face. She wears a school uniform consisting of [a brown jacket with gold trim, a white shirt, and a red ribbon or sash](#). The background suggests they are indoors, possibly in a hallway or room with wooden paneling. [The overall tone of the image is warm and inviting,](#) with soft lighting enhancing the characters' expressions.

RAP-MLLM Dataset

Figure A.15: Visualization of image captioning results for detail query prompts.

122 Importantly, these results also demonstrate that our RePIC does not degrade the original model's
123 general captioning capabilities after post-tuning. Unlike SFT approaches, our GRPO-based RL
124 training maintains the model's generalization ability. This is achieved by applying KL-divergence
125 regularization between the reference and target models during training, ensuring that the target
126 model remains close to the reference. Thus, by maximizing a verifiable reward while preserving
127 instruction-following ability through KL-divergence, RePIC generates the preferable personalized
128 image captions without compromising the original model's zero-shot capabilities.

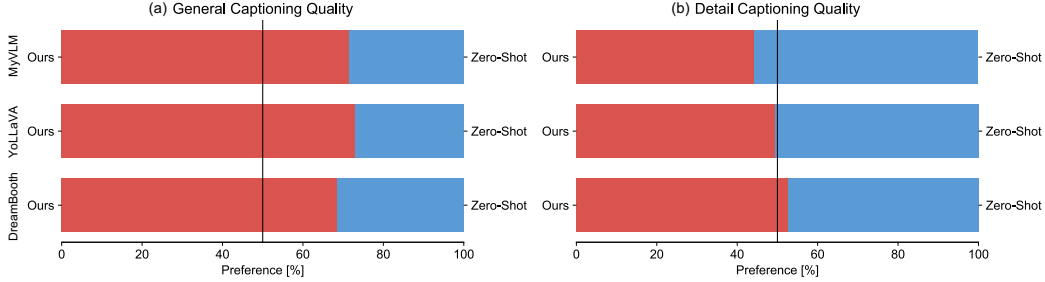


Figure A.16: Quantitative results of preference evaluations for the single-image captioning task without reference images, using (a) general query prompts and (b) detailed query prompts. Note that RePIC outperforms the zero-shot model in (a), and achieves comparable results in (b).

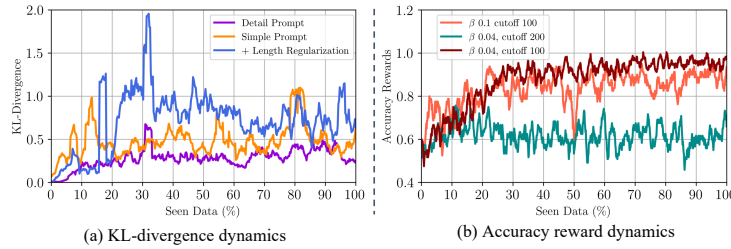


Figure A.17: Visualization of KL-divergence and accuracy reward plots on the seen data.

Table A.2: Visualization of recall scores (%) for 2-concept personal grounding.

Models	2-Concept		AVG
	Skip-Ret.	Ret.	
Ours-2K w/o multi-ICT	43.9	42.7	43.3
Ours-2K w/ multi-ICT	98.8	92.7	95.8

129 C.5 Why Multi-ICT is Necessary?

130 We present the recall scores in the 2-concept settings to verify the need to contain multi-ICT in
 131 the training data. As shown in Table A.2, models trained only with single-ICT fail to perform well
 132 in a multi-concept setting. This highlights the necessity of our proposed multi-ICT for improving
 133 multi-concept personal grounding performance.

134 C.6 Analysis on Hyperparameter Sensitivity

135 Figure A.17 presents the results of various ablation studies. In (a), we compare three settings: using
 136 only simple prompts, incorporating detailed prompts, and further applying length regularization based
 137 on a verifiable reward function. This reward assigns a value of 1 only when the output response
 138 length exceeds a predefined cutoff. We set the cutoff length to 100, as the average length of personal
 139 information in our database is approximately 100 tokens—roughly equivalent to at least one complete
 140 sentence. To encourage more informative image captions, we regularize the model to generate outputs
 141 of at least this length. In (b), we investigate how the expected reward changes with different values
 142 of the KL-divergence regularization weight β_{KL} . We also observe that the convergence behavior is
 143 influenced by the cutoff length used for length regularization. Our results indicate that the combination
 144 of $\beta_{KL} = 0.04$ and a cutoff length of 100 yields the best performance.

145 D Used Templates

146 D.1 Evaluation Templates

In Table A.3, we present the evaluation prompts used for personalized image captioning.

Table A.3: Prompts used for evaluating the personalized image captioning experiments.

General Caption Template:

- Give a personalized caption of this image.
- Give a caption of the image.
- Can you provide a personalized caption for this photo?
- Provide a caption of the given image.

147

148 D.2 Preference Evaluation Templates

149 The template used for our preference evaluation is shown below. Rather than favoring captions that
150 merely duplicate retrieved content, we instructed the model to evaluate preferred captions that convey
151 meaningful and accurate information to satisfy the following criteria:

- 152 1. **Reference Similarity:** Measures how closely the generated caption matches the retrieved
153 reference sentence. A higher similarity indicates potential redundancy, and thus a lower
154 preference score is assigned.
- 155 2. **Captioning Faithfulness:** Assesses how accurately the generated caption describes the
156 visual content of the input image.

Preference Evaluation Template

Retrieval-based Preference Evaluation: You are an evaluation expert. Your task is to determine which answer best describes the given image accurately. Carefully analyze the options and select the most appropriate one as your final choice.

Input: <Image>

The name of the object in this image is: {Name}.

The additional information for the given image is: {Info}.

The preferable caption is one that is not merely a duplication of the given information but provides a meaningful and accurate description.

Which one is more preferable caption to the {Name}?

Options:

A: {string1}

B: {string2}

Output the final answer by choosing one of the options with a single alphabet.

Answer: A, B

157

158 D.3 Instruction Templates

159 We further present the system prompts used for OCT and ICTs. In the following, in Tables A.4, A.5,
160 and A.6, we present the full instruction templates used for OCT, ICT, and VLT, respectively. Note,
161 we augment the instructions using GPT-4o.

Table A.4: Instruction templates used for OCT in training data.

Object Consistency Tuning (OCT) Template:

- Please verify whether the objects in these pictures are the same. An object is considered the same if its consistency is maintained despite variations in lighting or pose.
- Is <name> visible in this picture?
- Is <name> in this image?
- Do you see <name> in the photo?
- Is <name> present in this photograph?
- Can you identify if <name> is captured in this picture?
- Is <name> depicted in this image?
- Does the picture feature <name>?
- Can you confirm if <name> appears in this photo?
- Is <name> included in this shot?
- Is <name> shown in this image?
- Can you tell if <name> is part of this photograph?
- Is there any sign of <name> in this picture?
- Can you detect <name> in the photo?
- Is <name> captured in this image?
- Do you recognize <name> in this picture?

System Prompt for OCT

As an evaluation expert, your task is to verify whether the object identified as <name> in the first image is also present in the second image. Answer with yes or no. {Question}

162

System Prompt for Single-ICT

You are a captioning expert. Your task is to generate an accurate caption for the second image while referencing the first image. Both images contain the same object. The object in the first image is named <name>. {Question}

163

System Prompt for multi-ICT

You are a captioning expert. Your task is to generate an accurate caption for the last query image while referencing the given reference images. The reference images each contain an object, named respectively as <name1>, <name2>. {Question}

These are additional information about the given images except the last image: <name1>, <name2>, and <name3>. {Question}

Each object in the images not including the last image has a name: <name1>, <name2>. {Question}

Below is additional information about the object all images except the last one: <name1>, <name2>. {Question}

164

Table A.5: Instruction templates used for VLT in training data.

Visual Localization Tuning (VLT) Template:

- Please provide the bounding box coordinate of the region this sentence describes: <name>.
- Give <name>'s bounding box in the image.
- Describe <name>'s position in the image.
- Please provide the coordinates of the bounding box for <name> in the given image.
- Specify the rectangular boundaries of <name> in the image.
- Give <name>'s position in the following image.
- Please provide <name>'s bounding coordinates in the image.
- Indicate the bounding box for <name> in the image.
- Show the bounding box for <name> in the picture.
- Specify <name>'s bounding box in the photograph.
- Mark <name>'s bounding box within the image.

Table A.6: Instruction templates used for single and multi-ICT in training data.

Identity Consistency Tuning (ICT) Template:

- Give a caption of the image.
- Give a personalized caption of this image.
- Provide a general caption of the image.
- Summarize the visual content of the image.
- Create a detail caption of the image.
- Offer a rich and clear interpretation of the image.
- Describe the image in detail.
- Render a 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 faithfully?
- Generate a detailed and accurate description of the image.
- Write a caption that reflects the contents and context of the image.
- Compose a meaningful caption that truly represents the image.
- Describe the image in a personalized and context-aware manner.
- Provide a natural-sounding caption that accurately conveys what is in the image.
- Craft a caption that authentically describes the scene in the image.
- Create a caption that captures the essence of the image.
- Write a caption that reflects what’s visually happening in the photo.
- Generate a human-like description that accurately represents the image.
- Describe this image as if you were explaining it to a friend.
- Produce a relevant and truthful caption based on the image.
- Give a caption that matches the visual elements in the image.
- Summarize the visual content of this image in a natural way.
- Write an image-grounded caption that remains faithful to the content.
- Provide a descriptive sentence that corresponds closely to the image.

References

- [1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.
- [2] Jooyoung Choi, Chaehun Shin, Yeongtak Oh, Heeseung Kim, and Sungroh Yoon. Style-friendly snr sampler for style-driven generation. *arXiv preprint arXiv:2411.14793*, 2024.
- [3] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer, 2024.
- [4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [5] Haoran Hao, Jiaming Han, Changsheng Li, Yu-Feng Li, and Xiangyu Yue. Remember, retrieve and generate: Understanding infinite visual concepts as your personalized assistant. *arXiv preprint arXiv:2410.13360*, 2024.
- [6] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024.
- [7] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [8] Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024.
- [9] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.
- [10] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.
- [13] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.
- [14] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [15] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.