

Saliency-R1: Enforcing Interpretable and Faithful Vision-language Reasoning via Saliency-map Alignment Reward

Supplementary Material

In this supplementary material, we first give the derivation of the saliency map techniques when the normalization layer is involved (Sec. A). Then we provide statistics and curation details of our training data (Sec. B). We also give detailed implementation details (Sec. C) as well as brief introductions of the benchmarks utilized (Sec. D). Finally, we show the complete numerical results (Sec. E).

A. Normalization Layers

In this section, we outline the modifications required to incorporate normalizations layer (usually RMSNorm) into our method. Normalizations are applied in two key locations: before the projection layer (on the output of the transformer) and before each layer in the transformer block (on the layer input). We detail how the attribution techniques should be adjusted accordingly.

A.1. Pre-unembedding Normalization

In typical language models, a normalization layer, usually RMSNorm, is applied to the transformer’s output before the unembedding layers. Formally, the final logits is given by $\text{RMSNorm}(\mathbf{h}_i^L)\mathbf{E}_u$, and the normalization layer can be written as:

$$\text{RMSNorm}(x) = \gamma \times \frac{x}{\sigma_l^2 + \epsilon}, \quad (2)$$

where $x \in \mathbb{R}^d$ is the input token, $\sigma_l \in \mathbb{R}$ is the standard deviation, and $\gamma \in \mathbb{R}^d$ is learned vectors. As the normalization is a linear operation, when calculating the contribution of each visual tokens, we can directly apply the same operation to each token, i.e.,:

$$c_p = \sum_{l=1}^L \sum_{j=1}^H \text{RMSNorm}(\alpha_{i,j,p}^l \mathbf{W}_{o,j}^l \mathbf{W}_{v,j}^l \mathbf{h}_p^{l-1}) \mathbf{E}_u. \quad (3)$$

One major difference is that in this case, the standard deviation $\hat{\sigma}$ should be calculated from the final representation \mathbf{h}_p^L , i.e., $\hat{\sigma} = \text{s.d.}(\mathbf{h}_p^L)$.

A.2. MSAs Input Layer Normalizations

In the main paper, we do not describe the normalization layers that are applied to each input of the MSA block in the model. More accurately, the complete representation of the

MSA block are:

$$\begin{aligned} \tilde{\mathbf{h}}_p^{l-1} &= \text{RMSNorm}(\mathbf{h}_p^{l-1}), \\ \mathbf{A}_i^l &= \sum_{j=1}^H \sum_{p=1}^i \alpha_{i,j,p}^l \mathbf{W}_{o,j}^l \mathbf{W}_{v,j}^l \tilde{\mathbf{h}}_p^{l-1}, \\ \alpha_{i,j,p}^l &= \text{softmax}(\mathbf{W}_{q,j}^l \tilde{\mathbf{h}}_i^{l-1} \cdot \mathbf{W}_{k,j}^l \tilde{\mathbf{h}}_p^{l-1}). \end{aligned} \quad (4)$$

Therefore, by the linearity of the normalization layer, we can rewrite the token-wise contribution to incorporate the normalization:

$$c_p = \sum_{l=1}^L \sum_{j=1}^H \text{RMSNorm}(\alpha_{i,j,p}^l \mathbf{W}_{o,j}^l \mathbf{W}_{v,j}^l \tilde{\mathbf{h}}_p^{l-1}) \mathbf{E}_u. \quad (5)$$

B. Description of Dataset

In this section, we introduce the details of the dataset we used. The training of Saliency-R1 involves the use of two dataset, i.e., Saliency-R1-cold and Saliency-r1-8k.

B.1. Saliency-R1-cold

Saliency-R1-cold is a subset of the Vision-R1-cold [31], which is a composition of LLaVA-CoT dataset (100K) [79] and Mulberry dataset (200K) [82]. For samples with multi-round conversation, we split it and construct multiple single-round conversation sample. This is due the computational constraints that we cannot process text with extremely long contexts. Additionally, we filter out samples whose number of tokens in the response is greater than 1024. In the end, we have 272,881 samples.

B.2. Saliency-r1-8k

Saliency-r1-8k is a subset of the Visual-CoT dataset [59], which is a composition of TextVQA, TextCaps, DocVQA, DUDE, SROIE, CUB-200-2011, flickr30k, Visual7W, InfographicsVQA, VSR, GQA, and Open images. We randomly sample 8,080 images from the Visual-CoT dataset and construct our Saliency-r1-8k dataset. The proportion of each sub-dataset remains the same as the original Visual-CoT dataset.

C. Implementation Details

In this section, we provide more details of the implementation for bot cold start and the saliency-map alignment.

C.1. SFT

We use llama-factory [96] as the framework for conducting SFT. We conduct SFT on Saliency-R1-cold for 2 epochs. The maximum learning rate is 5×10^{-5} and then it is adjusted according to cosine annealing scheduler with warm-up ratio of 0.05. The batch size is 64. Lora with LoRA with rank 128 is used for optimization. Other settings are set to be the default values of llama-factory.

C.2. GRPO

GRPO is performed using TRL as the framework. We conduct GRPO on Saliency-r1-8k for 3 epochs. We utilize an original learning rate of 1×10^{-5} and batch size of 128. LoRA of rank 8 is performed on the query and value projection layers. The number of rollouts is set to be 8. Other settings are set to be the default values of TRL.

D. Introduction of benchmark

In this section, we provide a brief overview of the benchmarks used in this paper. Our experiments are conducted on well-established or recently introduced visual understanding benchmarks. These benchmarks were selected based on two key criteria: (1) The tasks must be vision-centric, meaning that the questions and answers should be derived from the content of the image, requiring attention to relevant regions within the image. Benchmarks that primarily focus on other aspects, such as mathematical reasoning or code generation, often depend heavily on the intrinsic reasoning abilities of the language model, which falls outside the scope of this paper. (2) The questions must be in a multiple-choice or True/False format, allowing us to directly evaluate the model’s performance without being affected by the variability of LLM evaluation.

MMMU-Pro. MMMU-Pro [88] is a strengthened version of the MMMU benchmark that more rigorously tests multimodal understanding and reasoning. It filters out questions solvable by text-only LLMs, expands multiple-choice options up to 10 to curb option-guessing, and introduces a vision-only setting where questions and options are embedded in photos/screenshots to require simultaneous “seeing” and “reading.”

MMBench. MMBench [43] is a comprehensive benchmark for evaluating multimodal models’ understanding and reasoning abilities across fine-grained tasks. It spans a wide range of skills: visual perception, commonsense and logical reasoning, detail recognition, spatial relations, text recognition, and cross-modal alignment, primarily through carefully curated multiple-choice questions. With rigorous human verification and de-duplication to mitigate training data contamination, MMBench aims to provide a fair and reproducible assessment.

POPE. POPE [39] is designed to evaluate object hallucina-

tion in vision–language models using binary-choice questions that verify the presence of specific objects in images. It reports metrics such as Accuracy, Recall, Precision, and F1 across three distinct sampling strategies, providing a robust and fine-grained assessment of hallucination tendencies. In our paper, we report the F1 score for POPE.

MME. The MME benchmark [10] evaluates multimodal models across 14 subtasks that span both perceptual processing and cognitive reasoning. By using carefully designed instruction–response pairs, MME reduces the likelihood of training data leakage, supporting a fair and rigorous assessment.

MME-RealWorld. MME-RealWorld [95] is a large, fully human-annotated benchmark for evaluating multimodal LLMs on high-resolution, real-world tasks. It comprises 13,366 images (avg $\sim 2000 \times 1500$) and 29,429 multiple-choice QA pairs across 43 subtasks in five domains: OCR in the wild, remote sensing, diagram/table understanding, video monitoring, and autonomous driving. Designed to stress fine-grained perception and reasoning, it includes an “E” option for unanswerable cases and enforces strict quality control by 25 annotators and 7 experts.

MMStar. MMStar [11] is a vision-indispensable multimodal benchmark of 1,500 human-curated, multiple-choice samples designed to fairly evaluate Large Vision-Language Models. It targets six core capabilities (coarse/fine-grained perception, instance/logical reasoning, science & technology, mathematics) across 18 axes, with strict visual dependency, minimized data leakage, and balanced difficulty.

ChartQA. ChartQA [48] is a benchmark for testing how effectively multimodal models answer questions about charts, emphasizing both visual comprehension and logical reasoning. It covers a range of chart types—such as bar charts and line graphs—and combines human-authored and automatically generated questions to probe complex reasoning. Importantly, ChartQA is highly OCR-dependent and demands fine-grained visual understanding, as models must extract textual information from charts and reason over it.

IllusionVQA. IllusionVQA [58] is a two-part visual question answering benchmark designed to test Vision-Language Models on challenging optical illusions. It comprises 435 multiple-choice comprehension instances across 12 illusion categories curated from diverse real images, and a 1,000-sample “soft localization” task where models identify which of two side-by-side objects is geometrically impossible. In this paper, we mainly focus on the soft localization task.

ScienceQA. ScienceQA [56] is a large-scale multimodal multiple-choice benchmark of about 21k K–12 science questions spanning natural, social, and language sciences, many with text and/or image contexts. Uniquely, it annotates most questions with “lectures” (background knowledge) and “explanations” (reasoning), enabling evaluation

and training of chain-of-thought reasoning. The dataset covers 26 topics, 127 categories, and 379 skills, includes both diagrams and natural images, and supports both few-shot prompting and fine-tuning studies.

SalBench. SalBench [16] is a vision-centric benchmark for Large Vision-Language Models that tests human-obvious visual saliency—finding an “odd-one-out” among distractors based on low-level features. Built by augmenting the P3 (synthetic) and O3 (natural) datasets with language instructions, it evaluates three tasks: Odd-One-Out Detection, Referring Odd-One-Out (text box coordinates), and Visual Referring Odd-One-Out (highlighted box). Features span color, orientation, size (synthetic) and extend to focus, shape, location, pattern (natural).

E. Additional Experimental Results

E.1. Faithfulness Experiments

In the main paper, we evaluate the faithfulness of the proposed saliency-map techniques on COCO Caption and OpenPSG datasets. Here we give additional experimental results on GranDf dataset (see Table 4). The results further support the claim that the proposed saliency-map generation techniques achieve state-of-the-art faithfulness compared with previous saliency-map generation techniques.

In addition to the reported deletion and insertion metric, we conduct counterfactual analysis to show that Saliency-R1 would focus more on the foreground pixels compared with baselines. To do so, we inject the Gaussian noise of varied σ to the foreground and background images of test set of saliency-r1-8k, and report the final accuracy. We compare the performance drop of the base model and Saliency-R1 in Table 5. The results demonstrate that compared with the base model, the Saliency-R1 show larger performance drop when Gaussian noise is injected to the foreground pixels. On the other hand, with noisy background, the performance drop of Saliency-R1 is smaller than the base model. This means Saliency-R1 relies more on the foreground pixels compared with the base model.

E.2. Robustness of Saliency-R1

We examine the robustness of Saliency-R1 on several additional dataset. 1) Noisy data: we inject Gaussian noise with different intensity into POPE and MME images, and evaluate the accuracy of the models with the noisy images as inputs 2) OOD data: we use VQA-RAD [35], a medical VQA dataset with significant domain gap with the saliency-r1-8k, to show that Saliency-R1 can generalize well to the medical domain. 3) ReasonVQA [69]: a multi-hop benchmark that requires the model to attend to multiple distinct regions across multiple image areas to correctly answer the question. Due to the huge size of the original validation set, we only sample 500 questions from the test set. We com-

Table 4. **Faithfulness experiment results.** We use Qwen2.5-VL-3B-Instruct for experiments. Our saliency map technique achieves comparable or better faithfulness to SOTA methods regarding the deletion and insertion metrics. The best metric is **bold** and the second best metric is underlined.

Method	GranDf					
	Deletion ↓			Insertion ↑		
	5%	15%	30%	5%	15%	30%
CAM [97]	81.71	78.21	74.14	-4.80	3.86	22.96
Grad-CAM [57]	81.71	78.21	74.14	-4.80	3.86	22.96
Raw Attention	79.72	73.02	66.58	-5.58	-2.06	15.91
Attention Rollout [1]	83.56	70.17	58.88	-4.94	1.36	25.98
ATTN-LRP [2]	70.97	57.06	44.57	-13.54	-1.58	30.38
TAM [40]	78.40	73.45	66.89	4.01	14.92	36.91
Ours	68.91	56.95	<u>49.22</u>	<u>-0.59</u>	<u>11.80</u>	<u>36.64</u>

Table 5. **Counterfactual test results.** We inject Gaussian noise with different σ to the foreground and background of the images, and prompt the model to answer visual questions.

Method	Accuracy				
	Clean	Foreground ↓		Background ↑	
	-	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$
Qwen2.5-VL-7B	70.8	60.8	63.9	50.5	55.4
Saliency-R1	74.4	58.0	66.4	48.9	59.0

Table 6. **Robustness Benchmark.** We compare Saliency-R1 with base model on several benchmarks that show the robustness and generalizability of the method. We inject Gaussian noise with different σ to the images of POPE and MME.

Method	POPE	POPE	MME	MME	VQA-RAD	ReasonVQA
	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$		
Qwen2.5-VL-7B	85.1	81.1	2317	2223	73.1	42.2
Saliency-R1	86.9	83.2	2360	2252	75.4	56.9

Table 7. **Throughput analysis.** We report the average time to generate one saliency map using the saliency-r1-8k dataset.

Method	CAM	Attention Rollout	ATTN-LRP	TAM	Ours
s/samples	2.29	3.89	4.52	5.56	2.59

pare the performance of Saliency-R1 with base model, the results are shown in Table 6. Saliency-R1 performs better than base model on these datasets, showing the robustness of the proposed method.

E.3. Throughput Analysis

In Table 7, we report the saliency-map generation time (including the inference time). All evaluations are conducted on a single NVIDIA A6000 GPU. The results demonstrate that our method maintains high training efficiency.

Table 8. Comparison with hallucination-reduction baselines.

Method	POPE
MFP-3B	86.5
RE-ALIGN-7B	87.1
MFCD-7B	87.4
Saliency-R1-3B	87.5
Saliency-R1-7B	88.1

E.4. Comparison with Additional Hallucination-reduction Baselines.

We compare Saliency-R1 with three additional hallucination-reduction baselines on the POPE benchmark, namely RE-ALIGN [78], MFCD [41], and MFP [38]. The results are shown in Table 8. Saliency-R1 performs favorably, achieving the highest performance among the evaluated methods.

E.5. Human Evaluation

To further show the improved interpretability of Saliency-R1, we conduct a small-scale user study. Specifically, we collect 10 VQA examples with the corresponding saliency maps, and ask 12 volunteers to score them on a scale of 1 to 5 regarding the quality of the interpretation. A saliency map of high quality should be visually coherent, relevant to the asked questions, and related to the thinking process and final answers. The average score of the base model is 3.6 and that of Saliency-R1 is 4.5. Wilcoxon signed-rank test shows the $p < 0.05$. This further validates that Saliency-R1 presents better interpretability.

E.6. Qualitative Examples

In this section, we show more examples of the saliency maps generated to reflect the reasoning process of the model (see Fig. 5). Our saliency map techniques can faithfully highlight the critical regions relevant to the questions.

E.7. Complete Response in Fig. 3

In Fig. 3 of the main text, we present several examples of generated responses along with their corresponding saliency maps. Due to space limitations, we could not include the full thinking processes for some cases. Here, we provide the complete versions of these reasoning processes. From the examples, it is evident that the original Qwen2.5-VL model struggles with reasoning, often producing trivial content. After applying SFT, the model’s reasoning capability improves significantly. However, we observe that the reasoning process often becomes excessively long, containing a substantial amount of irrelevant and unnecessary content. This irrelevant information dilutes the attention, causing the saliency map to highlight unrelated regions of

the image. With saliency-map alignment, the reasoning becomes more concise and targeted. This not only improves the model’s reasoning efficiency but also enhances its overall accuracy.

Row 1.

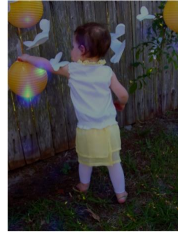
- **Qwen2.5-VL-7B.** The primary topic of the book is Cisco routers and switch management, as indicated by the title "Cisco Routers for the Desperate 2nd Edition." It suggests that the book is designed to help readers manage router and switch systems effectively, possibly in an easy or simplified manner.
- **Saliency-R1-CI-7B.** Okay, let’s see. The user provided a book cover titled "Cisco Routers for the Desperate: 2nd Edition." The subtitle displays router and switch management made easy. The author is Michael W. Lucas. The image has hands working on a Cisco router. The question is asking for the primary topic of the book based on the title. First, the main title is "Cisco Routers for the Desperate." The word "Cisco" here refers to the company known for networking equipment, specifically routers. The phrase "for the Desperate" suggests that the book is aimed at people who might be struggling or need help with their Cisco routers. The subtitle reinforces this by mentioning router and switch management. So putting it all together, the primary topic would be teaching readers how to manage Cisco routers and switches effectively. Even though there’s an illustration of hands working on the router, the title itself points directly to Cisco routers. The subtitle adds more detail about the content being about management, but the main subject is still Cisco routers. Therefore, the primary topic is managing Cisco routers.
- **Saliency-R1-7B.** Okay, let’s see. The user provided a book cover titled "Cisco Routers for the Desperate: 2nd Edition." The question is asking for the primary topic of the book based on the title. Looking at the title, the main subject is clearly "Cisco Routers." The subtitle shows management, but the primary topic is the routers themselves. The image also displays the cover has an illustration of hands working on a router, reinforcing that the focus is on Cisco routers. So the primary topic is Cisco Routers.

Row 2.

- **Qwen2.5-VL-7B.** The chairs appear to be arranged around tables in the foreground of the image. They are positioned facing towards the front where the speaker is standing, suggesting that this is a setup for a presentation or lecture. The chairs are likely located on the left side of the image, as they are visible in the foreground and seem to be part of the audience seating area.
- **Saliency-R1-CI-7B.** Okay, let’s see. The question is asking where the chairs are located based on the image provided. First, I need to recall the details from the image.



Question: Can you describe what the boat in the water is filled with?
<think> Okay, let's see. The user wants me to describe what the boat in the water is filled with based on the image provided. Looking back at the image, the boat has various items. The person is holding a basket of green vegetables, which are prominently displayed. There are also other baskets and containers with more produce. The boat is filled with these items, and the person is selling them. So the main things in the boat are the green vegetables in the basket, other baskets and containers with produce. The answer should list those items as mentioned. **</think>**
Answer: The boat in the water is filled with various items, including a basket of green vegetables and other baskets and containers with produce. "



Question: What kind of toy is the little girl playing with at the party?
<think> Okay, let's see. The question is asking what kind of toy the little girl is playing with at the party. From the image provided, the image shows a child holding a yellow paper lantern. The image displays illustrates that the child is interacting with the lantern, which is part of the party decorations. The key detail here is the yellow paper lantern. The answer should be straightforward since the image explicitly depicts "yellow paper lantern" as the toy she's playing with. **<think>**
Answer: The little girl is playing with a yellow paper lantern at the party.



Question: Can you describe the attire of the man who is busy on the phone?
<think> Okay, let's see. The question is asking to describe the attire of the man who is busy on the phone. From the image provided, the man is wearing a black jacket over a white shirt. He also has a striped tie with red, white, and black colors. There's a lanyard around his neck with an ID badge. So, the key elements here are the black jacket, white shirt, striped tie, and lanyard with an ID badge. I need to make sure I include all those details. The answer should be concise, just the attire. **<think>**
Answer: The man is wearing a black jacket over a white shirt, along with a striped tie that has red, white, and black colors. He also has a lanyard around his neck with an ID badge.



Question: Can you describe the position of the players in relation to the volleyball net?
<think> Okay, let's see. The question is asking about the position of the players in relation to the volleyball net based on the image provided. From the image, both players are jumping above the net. The player in blue is slightly behind the net, and the one in green is closer to the net but still above it. They're both mid-air, so their positions relative to the net would be that they're both above the net while jumping. The key points are that they're both above the net, with the blue player behind and the green one closer but still above. So the answer should describe their positions as being above the net. **<think>**
Answer: Both women are jumping above the volleyball net.



Question: Is there someone guiding the group of children and if so, can you describe them?
<think> Okay, let's see. The question is asking if there's someone guiding the group of children and, if so, describing them. From the image provided, the image shows a group of children walking on a road. In front of them, there's an adult holding a rainbow flag. The adult is wearing a dark coat and carrying a bag. The image displays illustrates that the adult appears to be guiding the children. So, yes, there is someone guiding them. The adult is illustrated as wearing a dark coat and has a bag. The key points here are the presence of the adult in front with the flag, which indicates they're leading. The answer should just state the adult is guiding them, and the image gives details about their appearance. **<think>**
Answer: Yes, there is an adult woman guiding the group of children. She is holding a rainbow flag and walking in front of the children.



Question: What's the setting behind the people who are gathered together?
<think> Okay, let me try to figure this out. The question is asking about the setting behind the people who are gathered together. From the image given, there's a narrow alleyway with buildings on both sides. The buildings have various colors like red, yellow, and white. There are some green plants and a gate in the background. The ground is wet with puddles, so it's been raining. The people are near the gate, which is part of the buildings. So the setting behind them would be the buildings and the gate. The answer should mention the buildings and the gate. **<think>**
Answer: Buildings

Figure 5. Additional examples of the saliency maps generated by our proposed saliency map techniques, and the corresponding questions and responses. The examples are generated using Saliency-R1-7B.

The image shows a conference room or meeting space. There's a stage with a podium where someone is speaking. In front of the stage, there are rows of tables arranged in an auditorium-style setup. Each table has chairs around them. The chairs are positioned directly in front of each table, facing the stage. So, the key points here are that the chairs are around the tables, which are set up in rows facing the stage. The image explicitly displays that the chairs are "positioned directly in front of each table, facing the stage." That means when people sit in the chairs, they're facing towards the stage where the speaker is standing. Therefore, the chairs are placed in front of the tables, which themselves are arranged in rows facing the stage. The answer should be straightforward since the image clearly illustrates their location relative to the tables and the stage.

- **Saliency-R1-7B.** Okay, let's see. The question is asking

where the chair is at based on the image provided. The image shows displays that the chairs are arranged around tables in rows, and they're positioned near the tables. So the chairs are at the tables. The answer should be straightforward from the given details.

Row 3.

- **Qwen2.5-VL-7B.** The image shows a group of people gathered outdoors near a waterfront. There is no visible instrument or device to the left of any girl in the image. The focus seems to be on the people and their activities rather than any specific objects or devices.
- **Saliency-R1-CI-7B.** Okay, let's break this down. The question is asking: "What kind of device is to the left of the instrument that the girl is to the right of?" First, I need to parse the image given. The image has a group near a waterfront with modern buildings. A girl is wearing a helmet and holding a guitar case. To her left, there's a device

Table 9. **Full Results of Ablation Studies.** The best metric is **bold** and the second best is underlined.

Method	MMMU-Pro	MMBench	POPE	MME	MME-RW	MMStar	ChartQA	IllusionVQA	ScienceQA	SalBench
	test	en_test	test	test	test	test	test	loc_test	img_test	p3_test
Qwen2.5-VL-7B [6]	36.2	<u>82.8</u>	86.7	2302	58.7	62.4	84.0	37.5	88.2	49.1
Saliency-R1-zero	37.8	82.9	86.6	2279	60.6	60.0	75.2	35.7	86.6	53.6
Vision-R1	36.3	80.2	87.3	2371	62.3	62.0	87.9	37.8	94.0	56.5
Saliency-R1-attn	35.9	81.7	88.3	2377	62.6	62.1	87.6	38.4	93.7	60.3
Saliency-R1-pure	37.5	80.9	87.9	2390	<u>62.7</u>	62.5	88.2	36.7	94.0	<u>61.5</u>
Saliency-R1-think	36.1	82.3	82.4	2328	<u>62.5</u>	61.1	86.6	35.3	94.5	58.5
Saliency-R1-answer	36.5	80.3	87.9	2378	62.0	60.6	87.8	35.7	93.1	55.2
Saliency-R1	<u>37.6</u>	81.8	<u>88.1</u>	<u>2385</u>	62.9	62.6	88.2	38.4	<u>94.3</u>	63.7

on the ground that looks like a camera or phone. The key here is understanding the spatial relationships. The girl is to the right of an instrument. The instrument mentioned is likely the guitar she’s holding, since she’s holding a guitar case. So the instrument in question is the guitar. Now, the device to the left of this guitar would be the one displayed as being on the ground to her left. According to the image, that device is a camera or phone. The question is asking for the type of device. Since both options are possible (camera or phone), but the image doesn’t specify which, just depicts “camera or phone,” the answer should reflect that ambiguity. However, the question might expect the more common association. But the original image shows displays “device resembling a camera or phone.” So the correct answer would be either camera or phone. But since the question uses “device” without specifying, maybe the answer is “camera” or “phone.” Wait, the problem illustrates “what kind of device,” so the answer should be the specific type. But the image only specifies “camera or phone.” So perhaps the answer is “camera” or “phone.” But the question is phrased as “what kind of device,” implying a singular answer. However, the image doesn’t specify which one; it’s either a camera or a phone. Therefore, the answer must be that it’s a camera or phone. But the way the question is structured, maybe the answer is “camera” because that’s more commonly associated with devices found on the ground at events. Alternatively, maybe the answer is “phone” if the context suggests a mobile device. But based strictly on the image provided, the answer is “camera or phone.” However, the question might be expecting the most likely one. But since the instruction shows to provide the final answer without explanation, and the image shows displays “camera or phone,” the answer should be those two possibilities. But the question is phrased as “what kind of

device,” so the answer is “camera or phone.”

- **Saliency-R1-7B.** Okay, let’s break this down. The question is asking what kind of device is to the left of the instrument that the girl is to the right of. First, from the image, there’s a girl with a guitar on her back. To the left of the guitar (which she’s carrying), there’s a device. The image shows illustrates a camera or similar device. So the device to the left of the guitar is a camera. The girl is to the right of the guitar, so the device is to the left of the guitar. The answer should be camera.

E.8. Complete Experimental Results of Ablation Studies

Here, we present the complete results of the ablation study. In addition to the experiments discussed in the main text, we conducted an additional experiment to evaluate the effectiveness of the proposed attention-rollout mechanism. Similar to the approach of Saliency-R1-think, we also tested a version of saliency maps that directly aggregates the answer tokens’ saliency maps without applying attention rollout, referred to as Saliency-R1-answer. The results of this experiment are shown in Table 9. Among all the ablation variants, the full version of Saliency-R1 demonstrates the best overall performance, highlighting the effectiveness of our design choices.