Explainable Saliency: Articulating Reasoning with Contextual Prioritization

Supplementary Material

1. Introduction

In the main paper, we introduced a novel framework for explainable saliency prediction that not only identifies important image regions but also provides human-understandable reasoning behind its predictions. Our approach leverages a vision-language reasoning mechanism and a contextual prioritization strategy to dynamically focus on semantically critical information, bridging the gap between predictive accuracy and interpretability. Unlike traditional saliency models that operate as black boxes, our method explicitly articulates the reasoning process, making it suitable for scenarios requiring high transparency and trustworthiness.

The supplementary materials provide further details and additional results to support these findings:

1) Sec. 2 investigates the effect of different top-K prototype selections on saliency prediction performance and identifies the optimal number of prototypes that balances computational cost and predictive accuracy.

2) Sec. 3 evaluates the impact of a range of backbone architectures, including variations of ResNet [25] and ViT [21], on saliency and faithfulness metrics.

3) Sec. 4 presents the model comparison results, analyzing the performance of MiniCPM-V 2.6 [58], LLaVA-1.5-7b [38], Llama-3.2-11B-Vision-Instruct [52], and GPT-4o [28] across saliency metrics (e.g., NSS [44], CC [42], AUC [3]) and faithful metrics (e.g., AUC-E [2], AOPC [6, 43], LOdds [46, 48]). The comparison highlights the strengths and limitations of each model in addressing saliency prediction.

4) Sec. 5 analyzes the limitations of our method, including failure cases where the model is misled by semantically similar objects and performance degradation on complex scenes and ambiguous questions.

By presenting these additional details and results, this supplementary document enhances the understanding of our framework, showcasing its robustness, interpretability, and generalizability.

2. Optimal Top-K Prototype Selection for Explainable Saliency

The selection of the optimal number of prototypes (top-K) represents a crucial design choice in our framework, balancing the competing needs of predictive accuracy and model interpretability. While a larger K value allows for a more comprehensive representation of the input image's semantics, it can lead to increased computational cost and reduced clarity in the explanations. Conversely, a smaller K value simplifies interpretation but may sacrifice predictive accuracy by neglecting important semantic information.

Our approach incorporates both theoretical considerations and empirical analysis to determine the optimal K. Theoretically, the AiR dataset's focus on task-driven saliency suggests a relatively small number of key objects or regions typically contribute to the final prediction. The model's Explicit Reasoning capability further refines this by prioritizing semantically relevant proposals. Based on these observations, we set an upper bound of K=10 for our experiments.

The experimental results for intermediate values of K (see Tab. 4) reveal that performance exhibits a peak at moderate K values (e.g., 3 and 5). Specifically, saliency metrics like NSS and AUC improve from 1.924 and 0.858 at K=2 to 1.946 and 0.860 at K=7. However, as K increases further to 10, these metrics slightly decline to 1.935 and 0.858, respectively. This can be attributed to the introduction of redundant or irrelevant prototypes at larger K values, which dilutes the focus of the model and increases noise in the explanation. Conversely, smaller K values like K=2 lack sufficient coverage of key semantic elements, resulting in reduced saliency accuracy and interpretability.

Notably, K=3 shows a good balance between accuracy and faithfulness, achieving near-peak performance in NSS (1.942) and AUC (0.858) while maintaining high values in AUC-E (0.713) and LOdds (-5.157). Based on these findings, we select K=3 as the optimal value for our framework. This choice is supported by the consistent high performance across both saliency and faithfulness metrics and the improved interpretability afforded by a smaller number of prototypes. While larger K values may offer marginal performance gains, they come at the cost of reduced explainability and increased computational burden, outweighing their potential benefits. The results show the effectiveness of our approach in balancing performance and interpretability, a crucial consideration for explainable AI systems.

3. Backbone Comparison

The choice of backbone architecture significantly impacts both the predictive performance and the explainability of our saliency model. To ensure a fair comparison, we carefully adapted ResNet [25] and ViT [21] architectures for our task, addressing their inherent architectural differences. This section details these adaptations and analyzes the resulting performance trade-offs.

Table 4. Performance comparison across different top-K prototype configurations on the AiR dataset. Metrics include CC, AUC, NSS, AUC-E, AOPC, and LOdds. The best value in each column is highlighted in blue, and the second-best value is highlighted in green.

K	NSS	CC	AUC	AUC-E ↓	AOPC ↑	LOdds ↓
2	1.924	0.689	0.858	0.695	0.756	-5.710
3	1.942	0.701	0.858	0.713	0.748	-5.157
5	1.938	0.693	0.859	0.730	0.736	-5.059
7	1.946	0.698	0.860	0.731	0.735	-5.050
10	1.935	0.691	0.858	0.727	0.732	-5.049

Table 5. Comparison of backbone architectures on saliency prediction tasks using both saliency and faithfulness metrics on the AiR dataset.

Backbone	NSS	CC	AUC	AUC-E ↓	AOPC ↑	LOdds ↓
ResNet50	1.942	**0.701**	0.858	**0.713**	**0.748**	-5.157
ResNet101	**1.972**	0.696	**0.860**	0.725	**0.748**	-5.144
ViT-B/16	1.741	0.638	0.839	0.720	0.667	-5.167
ViT-B/32	1.723	0.636	0.838	0.729	0.660	**-5.189**

3.1. Adapting Backbones for Saliency Prediction

To facilitate a direct comparison, we implemented several modifications to ensure consistent feature resolution and representation on the ResNet [25] and ViT [21] backbones.

ResNet Adaptation. We employed a dilated ResNet architecture, inspired by DINet [57], known for its effectiveness in low-level vision tasks. This involved replacing standard convolutional layers in layers 3 and 4 with dilated convolutions (dilation rates of 2 and 4, respectively). The stride in these layers was set to 1 to maintain high-resolution feature maps, crucial for capturing fine-grained spatial details essential for accurate saliency prediction. A 1x1 convolutional adapter was then used to standardize the dimensions of the output channel. We evaluated the ResNet-50 and ResNet-101 architectures to explore the impact of network depth.

ViT Adaptation. Unlike ResNet, ViTs employ a patch-based approach, making dilated convolutions impractical. Therefore, we adapted ViT by resizing the input image to a standard resolution and applying bilinear upsampling to the output features to match the resolution of the ResNet outputs. A 1x1 convolutional adapter was similarly used for channel standardization. We evaluated both ViT-B/16 and ViT-B/32 to explore the impact of patch size.

3.2. Results

Tab. 5 presents the performance of each backbone architecture, evaluated using both saliency metrics (NSS, CC, AUC) and faithfulness metrics (AUC-E, AOPC, LOdds). The results reveal a complex interplay between model capacity and explanation quality.

ResNet-101, the deepest ResNet architecture, achieves the highest saliency performance (highest NSS and AUC scores). However, its faithfulness metrics (AUC-E and LOdds) are comparatively lower than ResNet-50. This suggests that its superior feature extraction capabilities may overshadow the contribution of our reasoning modules, making the model less transparent. The more powerful backbone essentially "solves" the problem more independently, leaving less for the explanation to clarify.

ResNet-50 provides a more balanced performance, achieving strong results in both saliency and faithfulness metrics. This highlights the importance of finding an appropriate level of backbone complexity to effectively capture both high-level semantics and low-level visual details without compromising explanation quality.

ViT models, despite the modifications, exhibit consistently lower performance than ResNets. This is likely attributed to the information loss inherent in the upsampling process necessary to achieve comparable resolution. This limitation highlights the importance of considering architectural limitations when selecting backbones for tasks requiring fine-grained spatial information. The superior performance of ViT-B/16 over ViT-B/32 is attributed to its smaller patch size, retaining more spatial information.

In conclusion, the choice of backbone architecture significantly impacts the trade-off between saliency prediction performance and the fidelity of explanations. ResNet-50 emerges as a favorable choice, striking a strong balance between accuracy and interpretability. The results highlight the need to carefully consider not only predictive power but also the explainability of the chosen architecture when developing explainable AI systems.

Table 6. Comparison of vision language models on saliency prediction tasks using both saliency and faithfulness metrics on the AiR dataset. The average number of semantic proposals per sample is also included. The average number of semantic proposals per sample is also included.

Backbone	NSS	CC	AUC	AUC-E ↓	AOPC ↑	LOdds ↓	Avg. Proposals
MiniCPM-V 2.6	1.942	0.701	0.858	0.713	0.748	-5.157	2.47
LLaVA-1.5-7b	1.967	0.706	0.862	0.758	0.707	-4.684	3.31
Llama-3.2-11B-Vision-Instruct	2.009	0.720	**0.863**	0.742	0.736	-4.694	3.42
GPT-4o	**2.027**	**0.729**	**0.863**	**0.708**	**0.775**	**-5.196**	**2.28**

4. Vision-Language Model Comparison

In this section, we compare the performance of three Vision-Language Models (VLMs) on the AiR dataset: MiniCPM-V 2.6 [58], LLaVA-1.5-7b [38], Llama-3.2-11B-Vision-Instruct [52], and GPT-4o [28]. This comparison evaluates both saliency prediction metrics and faithful metrics, highlighting the trade-offs between semantic proposal focus, generality, and model performance.

4.1. Compared Models

MiniCPM-V 2.6: An 8-billion parameter model designed for efficient on-device deployment, demonstrating strong performance in OCR, high-resolution image understanding, and multilingual support.
LLaVA-1.5-7b: A 7-billion parameter model optimized for instruction-based multimodal tasks. It leverages the Vicuna architecture with a CLIP-ViT [47] vision encoder and an MLP cross-modal connector. It was fine-tuned on various visual instruction datasets.
Llama-3.2-11B-Vision-Instruct: An 11-billion parameter instruction-tuned model that excels in visual recognition, image reasoning, and question answering.
GPT-4o: A multimodal model with undisclosed parameter count, though presumably larger than the other models compared. It is optimized for instruction-following tasks through end-to-end training across text, vision, and audio, demonstrating strong performance in visual reasoning, semantic comprehension, and multimodal question answering.

To ensure meaningful and format-compliant semantic proposals, a consistent baseline prompt structure was employed. However, minor adjustments were made to account for variations in model size and instruction-following capabilities.

4.2. Results

Tab. 6 summarizes the performance of the three VLMs. Our analysis reveals a trade-off between saliency performance and faithfulness, likely influenced by the number and specificity of generated proposals.

While larger models like Llama-3.2-11B-Vision-Instruct and GPT-4o demonstrate superior saliency prediction performance (NSS: 2.009, 2.027; CC: 0.720, 0.729; AUC: 0.863, 0.863) compared to smaller models like MiniCPM-V 2.6 and LLaVA-1.5-7b, this advantage introduces a potential trade-off with faithfulness. Llama-3.2-11B-Vision-Instruct generates significantly more proposals per sample (3.42) than MiniCPM-V 2.6 (2.47, a 28% increase) and LLaVA-1.5-7b (3.31, a 25% increase), while GPT-4o produces the most selective outputs (2.28), yet maintains superior faithfulness (AUC-E: 0.708, LOdds: -5.196). This suggests that GPT-4o's focused semantic proposal generation enables it to balance saliency and faithfulness more effectively than the other models.

The results suggest a trade-off between saliency and faithfulness. Among the open-source models, MiniCPM-V 2.6's higher faithfulness stems from its more focused, task-relevant proposal generation. Its smaller, more concise output prioritizes critical concepts for saliency prediction. Conversely, while the larger LLMs exhibit stronger general instruction-following capabilities and generate more comprehensive proposals, leading to improved saliency, these broader outputs may include less essential concepts, thus reducing faithfulness. However, GPT-4o distinguishes itself by achieving strong saliency scores while maintaining the best faithfulness metrics, indicating that it effectively balances comprehensive semantic coverage with selectivity. The difference in faithfulness scores likely reflects variations in how each model handles task-specific semantic proposals. This underscores the importance of carefully balancing the breadth of semantic information captured with task-specific objectives when designing saliency prediction models. Future research should investigate strategies to optimize this trade-off, potentially through techniques that enhance the selectivity of larger models or improve the task-specificity of smaller ones.

5. Limitations

Despite the strong performance of our approach, there are still limitations that need to be addressed. In this section, we analyze two key areas where our model faces challenges: (1) failure cases where the model struggles due to semantic similarity, and (2) performance degradation on harder sub-

sets with increased scene complexity and ambiguous questions.

5.1. Failure Cases

In complex scenes requiring precise attention allocation and reasoning, current vision-language models (VLMs) still exhibit notable limitations. As shown in Figure 6, when given the question "Which object is to the left of the mug?", the VLM is misled at the visual level, incorrectly focusing on the bowl and utensils while failing to identify the plate as the correct answer.

This error occurs because the VLM relies heavily on semantic associations when determining relevant objects. During analysis, the model assigns higher importance to objects that frequently co-occur with the mug, such as the bowl, which is commonly present in kitchen scenes. At the same time, the presence of utensils as additional distractors further skews the model's attention. Instead of accurately identifying the most relevant object, the VLM overweights these semantically related items, leading to an incorrect attention distribution.

Moreover, the misinterpretation suggests that the model prioritizes conceptual relationships over distinguishing between visually distinct entities. The plate, despite satisfying the query condition, is overlooked, likely due to its weaker association with the mug in the model's learned priors. This case highlights the challenge of integrating linguistic and visual reasoning effectively, particularly in scenarios where multiple related objects exist.

While the model's inherent explainability enables us to trace the source of the misinterpretation, addressing such errors requires improving how the model balances semantic reasoning with visual grounding. With the advancement of vision-language models, more powerful architectures may help mitigate these issues and improve robustness in complex scenarios.

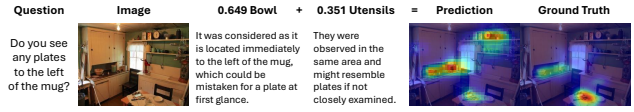

Figure 6. A failure case where the VLM is misled by semantic similarity, focusing on common kitchen items instead of correctly identifying the plate based on spatial positioning.

5.2. Hard Subset Analysis

To further evaluate the model's limitations, we assess its performance on two hard subsets:

- **Hard Images:** Images containing more than 30 objects.
- **Hard Questions:** Questions where human answer accuracy is below 70%.

Table 7 reports the results on these two subsets, comparing them to the overall model performance on the dataset.

Subset	NSS ↑	CC ↑	AUC ↑	AUC-E ↓	AOPC ↑	LOdds ↓
Overall (MiniCPM-V 2.6)	1.942	0.701	0.858	0.713	0.748	-5.157
Hard Images	1.793	0.700	0.857	0.712	0.694	-5.187
Hard Questions	1.732	0.682	0.843	0.734	0.642	-5.214

Table 7. Performance of our method on the AiR dataset, evaluated on the Hard Image and Hard Question subsets. The results show a decline in performance compared to standard evaluation, highlighting the increased difficulty of these subsets.

The results show a decline in both saliency prediction and faithfulness metrics, reflecting the increased difficulty of these cases.

For Hard Images, where the number of objects in the scene increases, the model's performance drops across all metrics. The NSS score decreases from 1.942 to 1.793, showing a significant reduction in alignment with human attention. Similarly, CC drops from 0.701 to 0.700, and AUC slightly decreases from 0.858 to 0.857, indicating that the model struggles to maintain high saliency coherence in cluttered environments. The faithfulness metrics also reflect a decline, with AOPC dropping notably from 0.748 to 0.694, revealing reduced effectiveness when removing high-saliency features. AUC-E remains nearly unchanged (from 0.713 to 0.712), indicating no meaningful difference in explanation consistency, while LOdds shifts slightly from -5.157 to -5.187, showing minimal numerical variation without a clear implication for attribution reliability.

For Hard Questions, where human agreement is lower, the performance drop is more pronounced. The NSS score declines from 1.942 to 1.732, reflecting a weaker correlation between model predictions and human gaze. CC also falls more significantly from 0.701 to 0.682, showing that the model's saliency maps deviate more from human attention patterns in ambiguous cases. AUC decreases from 0.858 to 0.843, further indicating reduced alignment in saliency estimation. On the faithfulness side, AUC-E increases from 0.713 to 0.734, pointing to greater inconsistency in explanation robustness, while AOPC drops from 0.748 to 0.642, suggesting that the removal of high-saliency regions has a lower impact on model predictions. LOdds moves from -5.157 to -5.214, but given its log-scale nature, this change remains relatively small without a clear impact on reliability trends.

These results indicate that both visual complexity and question ambiguity introduce significant challenges for saliency prediction and explanation reliability. The model struggles more in cluttered environments and when human attention patterns become less predictable, suggesting that it relies on contextual priors that may not generalize well to harder cases. As vision-language models continue to advance, handling such challenges will remain an important direction.