# Explainable Saliency: Articulating Reasoning with Contextual Prioritization

Nuo Chen Ming Jiang Qi Zhao
Department of Computer Science and Engineering
University of Minnesota
{chen9029, mjiang, qzhao}@umn.edu

## Abstract

*Deep saliency models, which predict what parts of an image capture our attention, are often like black boxes. This limits their use, especially in areas where understanding why a model makes a decision is crucial. Our research tackles this challenge by developing an explainable saliency (XSal) model that not only identifies what is important in an image, but also explains its choices in a way that makes sense to humans. We achieve this by using vision-language models to reason about images and by focusing the model's attention on the most crucial information using a contextual prioritization mechanism. Unlike prior approaches that rely on fixation descriptions or soft-attention based semantic aggregation, our method directly models the reasoning steps involved in saliency prediction, generating selectively prioritized explanations clarify why specific regions are prioritized. Comprehensive evaluations demonstrate the effectiveness of our model in generating high-quality saliency maps and coherent, contextually relevant explanations. This research is a step towards more transparent and trustworthy AI systems that can help us understand and navigate the world around us.*

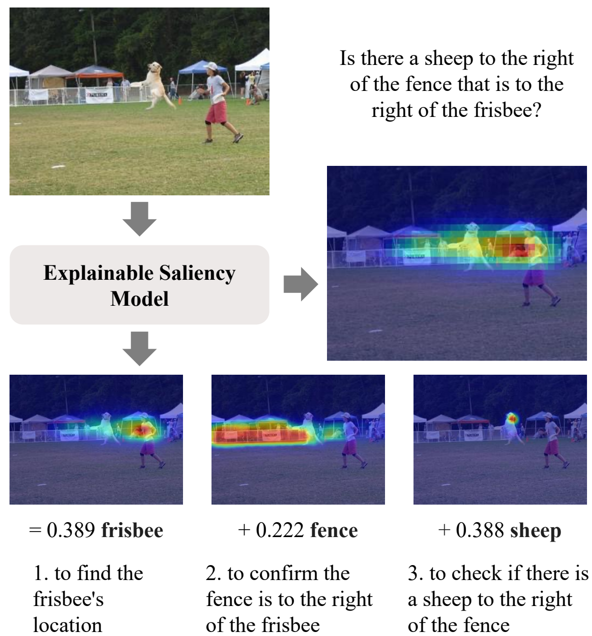


Figure 1. Our proposed XSal model provides step-by-step natural language explanations, semantic prototype visualization, as well as saliency weights to explain its prediction.

## 1. Introduction

Humans have this amazing ability to look at a scene and instantly know what is important. Saliency models, the cornerstone of computer vision, aim to mimic human gaze, identifying the most captivating regions within an image. These models have found wide application in image and video quality assessment [4, 54], virtual reality [42, 45], autonomous systems [8, 20], image captioning [13], and clinical diagnostics [11, 21]. However, while current saliency models [26, 36, 52] excel at replicating generic gaze patterns, they often operate as enigmatic black boxes. This lack of transparency hinders their application in scenarios where understanding why a region is deemed salient is as crucial as the prediction itself.

The development of explainable models that mimic human attention remains an underexplored research frontier. While some models have attempted to introduce explainability, they often rely on massive predefined vocabularies of semantics [10, 11] or simple descriptions of eye fixations without deep reasoning [15]. These methods, while offering some insights, lack true interpretability due to the lack of reasoning and prioritization within the context. This limitation is rooted in the inherent complexity of human attention. Human attention is a multifaceted phenomenon, influenced by a confluence of factors including visual cues, semantic understanding, individual preferences, and even cultural context. To generate meaningful explanations that unravel this interplay, a deep understanding of human cognition and perception is essential. Moreover, creating a model that can articulate its reasoning in a clear and concise manner poses a significant technical challenge.

To bridge the gap between saliency prediction and interpretability, we develop a saliency model that not only accurately predicts salient regions but also provides clear, human-understandable explanations. As shown in Fig. 1, our goal is to factorize saliency prediction with a weighted combination of spatial feature maps, namely semantic prototypes. Each semantic prototype represents the grounding of a semantic proposal (*e.g.*, frisbee, fence, and sheep) relevant to the visual task, and its relevance is explained with a natural language description. In this example, "frisbee" and "sheep" have higher weights in saliency prediction, because "to find the frisbee's location" and "to check if there is a sheep to the right of the fence" are essential reasoning steps to answer the question "Is there a sheep to the right of the fence that is to the right of the frisbee?"

To achieve this, we propose an explainable saliency (XSal) model that advances the state-of-the-art in saliency prediction and explanation. First, it integrates a vision-language model (VLM) to perform explicit reasoning for saliency prediction, generating step-by-step explanations and delivering natural language interpretations that clarify why specific regions are salient. Second, we introduce a contextual prioritization mechanism designed to handle the challenges of selecting truly important ones from a potentially large set of semantic explanations. This mechanism allows the model to focus on key elements, achieving interpretability without overcomplicating the representation. By integrating explicit reasoning with targeted selection, we create a cohesive framework that not only identifies and explains salient regions but also adapts dynamically to scene complexity.

In summary, this paper makes three key contributions to the field of saliency prediction and explainable AI:

1. This paper presents the first saliency model to incorporate a vision-language model for explicit reasoning. This integration allows the model to not only identify salient regions but also articulate its reasoning process, providing clear and human-understandable explanations.
2. We introduce a novel contextual prioritization mechanism that prioritizes the most relevant information within the context of saliency prediction and explanation. This results in a more interpretable model by focusing on key relevant semantics and simplifying the representation.
3. We conduct extensive evaluations of our model, demonstrating its effectiveness in generating high-quality saliency maps while providing coherent, contextually relevant explanations. This comprehensive assessment highlights the model's ability to effectively bridge the gap between prediction accuracy and interpretability.

## 2. Related Work

**Saliency Prediction.** Saliency prediction models have traditionally focused on bottom-up approaches, relying on low-level visual features in free-viewing conditions [5, 25, 27, 29] . However, with the rapid progress of deep learning methods, deep saliency models have emerged, achieving promising accuracy in predicting both task-free bottom-up saliency [26, 34, 52] and task-driven top-down attention [1, 30, 48]. A key strength of deep saliency models is their ability to extract high-level semantic features. Numerous deep neural network-based methods have boosted saliency prediction performance by implicitly encoding semantics using different approaches [46, 52]. However, such models often lack transparency and trustworthiness due to explicitly learned representations. While there has been growing interest in understanding saliency model mechanisms, existing approaches fall short of providing comprehensive explanations. For instance, Chen et al. [10, 11] analyze model behaviors using semantic prototypes and soft attention weights, but these approaches offer limited insight into the model's decision-making process, particularly when dealing with thousands of prototypes. Similarly, GazeXplain [15] generates plain natural language descriptions of fixation points, lacking a reasoning or prioritization mechanism to identify crucial factors driving its predictions. These limitations highlight the need for a more robust methodology that can uncover the underlying rationale behind saliency model behavior.

Our proposed method addresses these limitations by introducing a novel framework that combines explicit vision-language reasoning with a hard attention mechanism. This ensures that only the most critical elements are selected for saliency prediction and explanation, which not only improves the accuracy of saliency predictions but also provides a more transparent and insightful understanding of the model's behavior.

**Vision Language Models.** Our work builds upon the growing field of explainable AI (XAI) within vision-language models (VLMs). Recent advancements in VLMs [13, 14, 17, 22, 31, 35, 47, 50, 51] have revolutionized multimodal understanding, enabling these models to seamlessly integrate visual and textual information. This progress has opened new avenues for developing XAI systems capable of providing interpretable insights into complex decision-making processes. For instance, researchers have explored using VLMs to generate natural language explanations for tasks such as image classification [23, 38], abusive language detection [19], and autonomous driving [7], allowing users to understand the rationale behind model predictions. Additionally, VLMs have been employed to enhance the interpretability of visual scanpath models, describing image features driving human attention to a sequence of fixation points [15].

Different from these studies, by integrating textual context and introducing reasoning and prioritization mechanisms, we aim to provide deeper insights into the factors

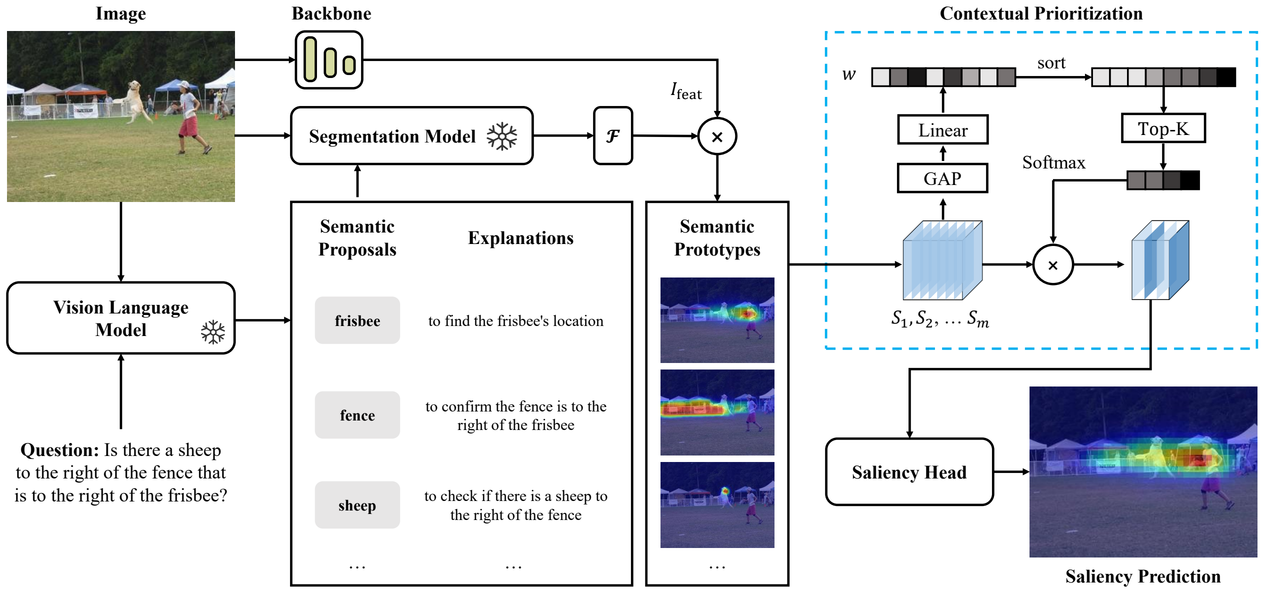


Figure 2. The proposed XSal architecture begins with a vision-language model generating semantic proposals from the input image, and uses a segmentation model to generate their corresponding semantic prototypes. A contextual prioritization module then selects the top-K relevant semantic prototypes, which are processed by a final saliency head to predict a focused saliency map.

driving saliency predictions, improving the transparency and interpretability of saliency models.

## 3. Method

The human visual attention process is a captivating blend of explicit reasoning and selective focus, guiding us to prioritize relevant elements while navigating complex scenes. While existing saliency models have made strides, they often fall short of capturing this nuanced behavior, relying instead on dense representations or lacking a mechanism for explaining their choices. This incongruence between model behavior and human attention motivates us to introduce a novel approach that leverages VLMs to perform explicit reasoning and prioritizing the reasoning outcomes based on contextual relevance for saliency prediction and explanation.

### 3.1. Explicit Reasoning

At the heart of our approach lies a vision-language model (VLM), specifically MiniCPM-V 2.6 [53], which enables the model to identify and explain salient elements across task-driven or task-free contexts. Specifically, the VLM produces a set of semantic proposals $\mathbf{C}$ and associated explanations $\mathbf{R}$:

$$\mathbf{C}, \mathbf{R} = \mathrm{VLM}(\mathbf{I}, \mathbf{Q}) \tag{1}$$

Here, $\mathbf{I}$ is the input image, and $\mathbf{Q}$ is the question guiding attention. $\mathbf{C} = \{P_1, P_2, \ldots, P_n\}$ represents semantic proposals, each focusing on specific objects within the scene, while $\mathbf{R}$ provides rationale for why each element was

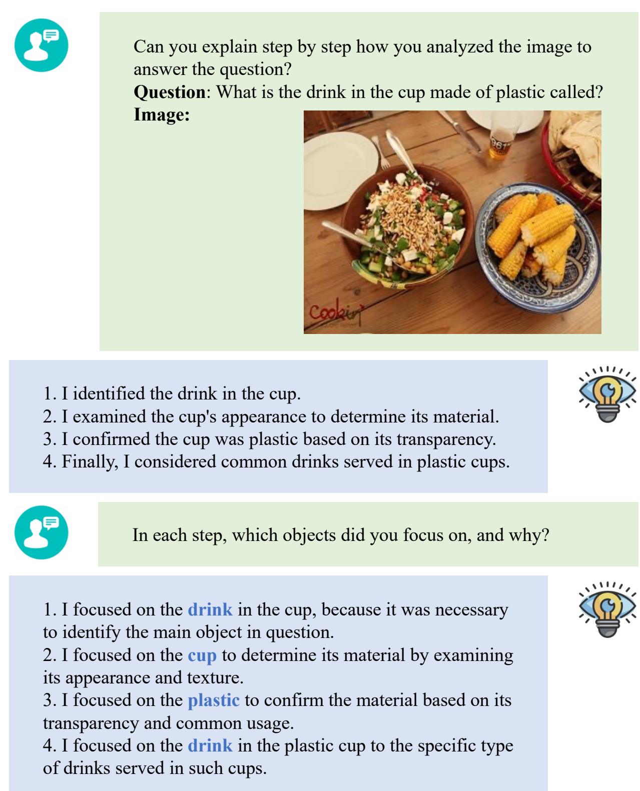


Figure 3. The two-stage prompting method to obtain semantic proposals and explanations from the VLM.

selected, ensuring interpretability through alignment with human attention patterns. This integration supports a reasoning process that mirrors human problem-solving, where decisions are informed by both visual cues and a structured understanding of why certain elements are relevant. To obtain the semantic proposals and explanations from the VLM, as shown in Fig. 3, we employ a two-stage prompting method:

In the initial planning stage, we ask the VLM to describe "how you analyzed the image to answer the question" (for top-down task-driven cases) or "important objects in the image" (for bottom-up free-viewing cases). The model responds with a coherent sequence of actions of reasoning steps to follow when answering the question. For example, to answer the question "What is the drink in the cup made of plastic called?", the reasoning steps include identifying the drink in the cup, examining the cup's appearance to determine its material, confirming the cup was plastic based on its transparency, and considering common drinks served in plastic cups.

Following this planning, the VLM is asked to explain what specific objects it focused on in each step, and explain the reason. It then proceeds to generate the set of semantic proposals $\mathbf{C}$ (such as drink, cup, and plastic) and their corresponding explanations $\mathbf{R}$. Finally, to ensure precision in the final output, the VLM is also prompted to verify that the identified semantic proposals are concrete and relevant, addressing potential limitations in model consistency.

This layered approach replicates human selective attention and explanatory reasoning, ensuring the model's focus and reasoning align closely with human attention patterns.

After identifying semantic proposals through the VLM, we utilize an open-vocabulary segmentation model, *ClipSeg* [37], to extract the visual grounding of these proposals on the image. *ClipSeg*'s ability to handle open-vocabulary concepts is crucial as the language model may describe elements that don't perfectly align with traditional segmentation categories. This step bridges the semantic reasoning process with visual representations, allowing for visual validation of the model's reasoning. While *ClipSeg* effectively outlines entire object contours, human attention typically focuses on smaller, more localized regions. Therefore, instead of using the segmentation results directly for saliency prediction, we use them as binary masks to filter the image features, and apply a convolutional layer $\mathcal{F}$ to obtain fine-grained localization. This results in a sequence of focused spatial maps, namely semantic prototypes, where each prototype $S_i$ aligns with a semantic proposal $P_i$, localizing its representations in the image. This process pinpoints focused regions that naturally attract human attention, associating them with natural language explanations.

## 3.2. Contextual Prioritization Mechanism

Our approach utilizes a contextual prioritization mechanism to extract the most relevant semantic prototypes for saliency prediction. Different from existing saliency models that often integrate hundreds to thousands of spatial feature maps with linear weights, humans naturally focus on a limited number of features, while other elements, though present, are not prioritized to the same extent. Our proposed contextual prioritization is a highly selective approach, which only allows attention to be directed by the most relevant features. It introduces a dynamic approach that adapts to varying numbers of semantic prototypes. This flexibility allows the model to handle scenes of different complexities effectively.

Our contextual prioritization mechanism adjusts importance by applying global average pooling on the extracted semantic prototypes, followed by a linear layer to generate initial weights. These weights are then sorted, and only the top $K$ and their corresponding features are selected. A softmax operation is applied to normalize the selected weights, yielding a refined representation:

$$\mathbf{w} = \mathrm{Linear}(\mathrm{GAP}(\mathbf{S})), \tag{2}$$

$$\mathbf{w}_{\mathrm{topK}} = \mathrm{Softmax}(\mathrm{TopK}(\mathbf{w})), \tag{3}$$

$$\mathrm{O} = \sum_{i=1}^{K} \mathbf{w}_{\mathrm{topK},i} \cdot \mathbf{S}_{\mathrm{topK},i} \tag{4}$$

where $\mathbf{S}$ represents a collection of semantic prototypes, $\{S_1, S_2, \ldots, S_n\}$. The global average pooling (GAP) and subsequent linear transformation extract initial weights $\mathbf{w}$ for each prototype, allowing the model to assign varying importance to each semantic element. By selecting and normalizing the top $K$ weights, we obtain $\mathbf{w}_{\mathrm{topK}}$ and the corresponding $\mathbf{S}_{\mathrm{topK}}$. Here, $\mathbf{S}_{\mathrm{topK}}$ corresponds to the semantic prototypes associated with the top $K$ weights in $\mathbf{w}$, determined by sorting. This ensures that the model prioritizes the most significant semantic elements, aligning its focus with the most contextually important regions, and discarding less relevant elements. Finally, the aggregated saliency feature O, the weighted combination of the top $K$ semantic prototypes, is processed by a saliency head to produce the final saliency prediction. This saliency head consists of only a single linear layer (with bias set to 0) that simply compresses the channel dimension, acting as a weighted sum without introducing any complex black-box behavior.

To ensure adequate feature optimization and prevent limited feature selection from hindering convergence, we introduce a *Progressive Top-K Reduction* strategy. Rather than starting with a fixed, small set of key features, this strategy initially preserves a larger number of semantic elements, enabling comprehensive feature optimization in early training. As training progresses, the $K$ value is gradually reduced, guiding the model to focus on the most significant features. The progression of $K$ across epochs is defined as

$$K_t = K_{\text{init}} - \frac{(K_{\text{init}} - K_{\text{final}}) \times t}{T - 1} \quad (5)$$

where $K_t$ is the $K$ at epoch $t$, $K_{\text{init}}$ and $K_{\text{final}}$ are the initial and final values of $K$, and $T$ is the total number of epochs. This progressive reduction allows the model to gradually shift from broad feature optimization to a selective focus on the most critical elements. With this strategy, the contextual prioritization mimics the adaptive prioritization seen in human attention, flexibly weighing elements based on their relevance and adapting to scenes with different numbers of semantic prototypes.

# 4. Experiments

This section presents a comprehensive evaluation of our proposed model, examining its performance and explainability across both task-driven and task-free eye-tracking datasets. We conduct an ablation study to further investigate the contributions of individual components, and present qualitative examples to illustrate the model's capabilities in generating interpretable saliency maps and explanations. Additional experimental results and analyses are provided in the supplementary material for further reference.

## 4.1. Experimental Setup

**Datasets.** Our experiments are conducted on two eye-tracking datasets: The *AiR* dataset [9], derived from the GQA dataset, comprises images and questions paired with eye-tracking data from 20 participants answering those questions. This setup provides a realistic scenario for evaluating saliency models, as it captures human attention in the context of a specific task (question answering). The *OSIE* [49] dataset is a free-viewing dataset featuring multiple salient objects competing for attention in the same image context, suitable for developing and evaluating saliency models that consider object- and semantic-level information. In our experiments, we split both datasets into training, validation, and test sets following [12, 16].

**Compared Models.** We compare our model's performance against several state-of-the-art saliency prediction models including SAM [18], SALICON [26], DINet [52], as well as explainable models such as Chen et al. [10] and GazeXplain [15]. Note that for GazeXplain [15] we aggregate fixation points from its predicted scanpaths into saliency maps for comparison. By comparing our model with these diverse methods, we aim to demonstrate its performance in various aspects, including general saliency prediction and explainability.

**Evaluation Metrics.** To comprehensively evaluate our model, we utilize a combination of metrics that assess saliency prediction performance and explainability. 1) *Saliency Prediction Metrics:* We employ standard metrics to evaluate the alignment between our model's predicted saliency maps and human attention patterns. These metrics are Area Under the ROC Curve (AUC) [3], Normalized Scanpath Saliency (NSS) [41], and Correlation Coefficient (CC) [39], which provide quantitative insights into the accuracy and relevance of the generated saliency maps. 2) *Explainability Metrics:* We evaluate the interpretability of saliency predictions adapting three XAI metrics: AUC [2] assesses model stability by calculating the area under the curve that represents changes in model performance as different proportions of semantic prototypes are progressively perturbed. Note that this AUC differs from the AUC used in saliency evaluation: here, it measures the model's sensitivity to perturbations, while saliency AUC measures alignment with human attention data. To facilitate distinction, we refer to this interpretability metric as AUC-E throughout this paper. Area Over the Perturbation Curve (AOPC) [6, 40] quantifies the sensitivity of the model to perturbations by measuring the change in NSS as semantic prototypes are sequentially altered. Log-odds Score (LOdds) [43, 44] evaluates the change in NSS as semantic prototypes are perturbed, focusing on the degree of reduction in NSS in response to prototype alterations. These metrics provide a comprehensive framework for assessing the quality of our model's explanations, enabling us to gauge not only where the model focuses, but also the robustness of its interpretative reasoning under perturbations.

**Loss Function.** To align the model's saliency predictions effectively with human attention patterns, we employ a combination of loss functions, including NSS, CC, and Kullback-Leibler Divergence (KLD) [33]. Our implementation follows [10]. These loss functions provide complementary signals that guide the model to capture nuanced attention details, helping to achieve balanced performance across various saliency prediction metrics.

**Implementation Details.** Our model employs a dilated ResNet-50 [24] backbone, following the setting used in DINet [52]. This architecture incorporates dilated convolutions to expand the receptive field without additional downsampling, which enhances the model's capacity to capture high-resolution spatial details essential for accurate saliency prediction. The model is optimized using the Adam optimizer [32] with an initial learning rate of 4e-4 and a weight decay of 1e-7. To ensure stable convergence, the learning rate is adaptively reduced by a decay factor of 0.1 every two epochs, while gradient clipping is applied with a threshold of 0.1 to prevent gradient explosion. Training is conducted over 10 epochs with a batch size of 10. To ensure stable training and prevent convergence issues, we employ the proposed *Progressive Top-K Reduction* strategy, which begins with a broader feature selection and gradually narrows the focus to the most significant features. Specifically, we set the initial $K$ value to 20, which progressively reduces to a final value of 3 over 3 epochs, enabling adaptive and selec-

| | AiR | | OSIE | |
|---|---|---|---|---|
| | NSS | CC | NSS | CC |
| SAM [18] | 1.65 | 0.62 | 2.70 | 0.65 |
| SALICON [28] | 1.70 | 0.63 | 2.75 | 0.63 |
| DINet [52] | 1.77 | 0.63 | 2.88 | 0.63 |
| *Chen et al.* [10] | 1.76 | 0.63 | 2.91 | 0.64 |
| GazeXplain [15] | 1.85 | 0.66 | 2.53 | **0.74** |
| XSal | **1.94** | **0.70** | **2.96** | 0.72 |

Table 1. Comparative results of saliency prediction across the AiR and OSIE datasets.

tive feature prioritization.

### 4.2. Saliency Prediction Results

Tab. 1 presents a quantitative comparison of saliency prediction performance across the AiR (task-driven) and OSIE (task-free) saliency datasets for our proposed model and several state-of-the-art methods. As shown in the table, our proposed model consistently outperforms other methods on the AiR dataset, achieving the highest NSS score of 1.94 and the highest CC score of 0.70. This suggests that our model effectively captures top-down human attention patterns in task-driven scenarios. Our model also achieves the highest NSS score of 2.96 on the OSIE dataset. Though our CC score (0.72) is slightly lower than GazeXplain (0.74), it still demonstrates strong performance in capturing the overall distribution of human attention in task-free image-viewing. The consistently high NSS scores across both datasets indicates that our model accurately predicts the locations where humans are mostly likely to fixate their gaze, demonstrating a strong accuracy of predicting the most salient regions. However, our model might be slightly less effective at capturing the overall distribution of human attention in images, because image regions with low saliency (low agreements across human observers) are less explainable and likely to be discarded by the contextual prioritization mechanism. Overall, the quantitative results demonstrate that our proposed method significantly improves saliency prediction performance compared to existing approaches. The incorporation of explicit reasoning and selective attention mechanisms contributes to more accurate and robust models that effectively capture the complexity of human visual attention.

### 4.3. Saliency Explanation Results

Tab. 2 presents the quantitative results based on explainability metrics, comparing our model's performance with Chen et al. [10] on both the AiR and OSIE datasets. Notably, our model demonstrates significantly lower AUC-E, LOdds, and higher AOPC scores compared to Chen et al. Specifically, on the AiR dataset, our model significantly outperforms Chen et al. with a lower AUC-E (0.713 vs. 0.899) and a higher AOPC (0.748 vs. 0.241). The substantial improvement in LOdds (-5.157 vs. -0.143) also reflects the model's ability to focus on the most relevant elements, minimizing irrelevant attention. On the OSIE dataset, a similar trend is observed. Our model achieves a better AUC-E (0.685 vs. 0.810) and significantly higher AOPC (1.115 vs. 0.586) compared to Chen et al. Additionally, the considerable decrease in LOdds (-4.169 vs. -0.250) further supports our model's capability to highlight key semantic concepts more effectively, thereby improving interpretability in a task-free viewing context.

These results indicate that our model is more sensitive to changes in semantic prototypes, highlighting the stronger reliance on these prototypes for generating accurate saliency predictions. The lower AUC-E and LOdds scores suggest that perturbing prototypes leads to a more substantial decrease in performance for our model, highlighting the critical role of semantic understanding in our approach. Conversely, the higher AOPC score signifies a greater sensitivity to changes in prototypes, revealing a stronger correlation between the model's performance and the identified semantic elements. This reinforces the notion that our model's interpretability is closely tied to the semantic prototypes it extracts, providing a more transparent and interpretable reasoning process compared to methods that rely on complex aggregation of numerous features. Overall, these results emphasize the importance of explicit reasoning and selective attention in achieving a higher degree of interpretability and understanding in saliency prediction.

### 4.4. Ablation Study

Our ablation study, conducted on both the AiR and OSIE datasets, reveals the significant contributions of each component to our model's performance and explainability. As shown in Tab. 3, a simple baseline model, using only a dilated ResNet-50 backbone [24] with a saliency head, achieves reasonable performance in both settings, yielding an NSS of 1.777 on AiR and 2.835 on OSIE. The baseline model lacks explainability, with a LOdds of -0.131 and -0.246 on these datasets, respectively.

By incorporating the VLM for reasoning, our model links saliency with task-relevant semantic elements and significantly improves both accuracy and interpretability across both datasets. For instance, on the AiR dataset, the NSS increases to 1.917 and the CC score improves to 0.703, while the AUC remains relatively stable at 0.858. This highlights the crucial role of task-oriented alignment and semantic understanding in generating clear and contextual explanations for saliency predictions. This is further evidenced

| | AiR | | | OSIE | | |
|---|---|---|---|---|---|---|
| | AUC-E↓ | AOPC↑ | LOdds↓ | AUC-E↓ | AOPC↑ | LOdds↓ |
| Chen *et al.* [52] | 0.899 | 0.241 | -0.143 | 0.810 | 0.586 | -0.250 |
| XSal | **0.713** | **0.748** | **-5.157** | **0.685** | **1.115** | **-4.169** |

Table 2. Comparative results of explainability metrics across the AiR and OSIE datasets.

| Dataset | Baseline | Reasoning | ContPri | NSS | CC | AUC | AUC-E↓ | AOPC ↑ | LOdds ↓ |
|---|---|---|---|---|---|---|---|---|---|
| AiR | ✓ | | | 1.777 | 0.640 | 0.849 | 0.910 | 0.225 | -0.131 |
| | ✓ | ✓ | | 1.917 | **0.703** | 0.858 | **0.700** | **0.749** | -5.077 |
| | ✓ | ✓ | ✓ | **1.942** | 0.701 | **0.858** | 0.713 | 0.748 | **-5.157** |
| OSIE | ✓ | | | 2.835 | 0.681 | 0.885 | 0.823 | 0.567 | -0.246 |
| | ✓ | ✓ | | 2.904 | 0.717 | **0.897** | 0.765 | 0.876 | -1.738 |
| | ✓ | ✓ | ✓ | **2.965** | **0.726** | 0.895 | **0.685** | **1.115** | **-4.169** |

Table 3. Ablation study on the impacts of Baseline, Reasoning, and Contextual Prioritization (ContPri).

by the significant improvements in interpretability metrics – the AUC-E drops to 0.700 and the AOPC increases to 0.749, suggesting our model's ability to provide more coherent explanations for task-driven saliency.

Further, the contextual prioritization mechanism exhibits a more pronounced impact, especially in the OSIE task-free setting. For example, on OSIE, the NSS increases to 2.965, the CC improves to 0.726, and the AUC remains relatively stable at 0.895. This suggests that prioritizing the most salient elements plays a crucial role in handling complex scenes with multiple potential salient objects, ultimately leading to more concise and clear attention maps. This is further evidenced by the significant improvements in interpretability metrics – the AUC-E decreases to 0.685, the AOPC increases to 1.115, and the LOdds decreases to -4.169. This greater impact on the OSIE dataset could be attributed to the presence of a large number of potential semantic prototypes in task-free settings, making selective focus particularly valuable for distilling relevant information and improving interpretability.

These findings demonstrate that the combination of explicit reasoning and contextual prioritization mechanisms, implemented in our model, is essential for achieving superior performance and explainability in both task-driven and task-free saliency prediction.

### 4.5. Qualitative Analysis

Qualitative analysis of our model reveals its ability to effectively integrate task-driven guidance and contextual prioritization to generate accurate saliency predictions and explanations. As illustrated in Fig. 4, the model aligns its attention with the goals of each question, selectively focusing on essential objects while ignoring irrelevant elements. For example, in questions about the position of a skateboard and a red cone (row 1), the model prioritizes these objects over traditionally salient regions like the skateboarder, demonstrating the advantage of explicit reasoning. The model also adapts well to abstract concepts, such as interpreting "modern" style in a question about a faucet (row 2). Finally, the model excels in context prioritization by isolating key elements in complex scenes, highlighting the person, bag, and cars in a question about cars (row 3) while filtering out other visual elements. These qualitative findings demonstrate the model's ability to produce explanations that are closely aligned with the model's internal reasoning, ensuring outputs that are both interpretable and reliable.

Beyond its interpretability, our model shows significant performance advantages in accuracy and task relevance. In Fig. 5, our model achieves the highest accuracy among state-of-the-art methods, aligning closely with ground truth. For example, in the question "What is the drink in the cup made of plastic called?", it focuses on the plastic cup, unlike other models that are distracted by irrelevant food items (row 1), showcasing the strength of our explicit reasoning approach. Additionally, our model demonstrates strong context prioritization, effectively concentrating on relevant elements even in complex scenes with multiple distractions. For instance, in the question "Is the fire hydrant to the right or to the left of the garbage can?", our model isolates the hydrant and garbage can, while other models diffuse attention across the scene (row 2), highlighting its ability to prioritize contextually important elements. Furthermore, in dynamic scenes, our model reliably focuses on salient objects. In the question "Are both the helmet and the bat the same color?", it highlights the helmet and bat specifically, unlike other models that diffuse attention (row 3), demonstrating

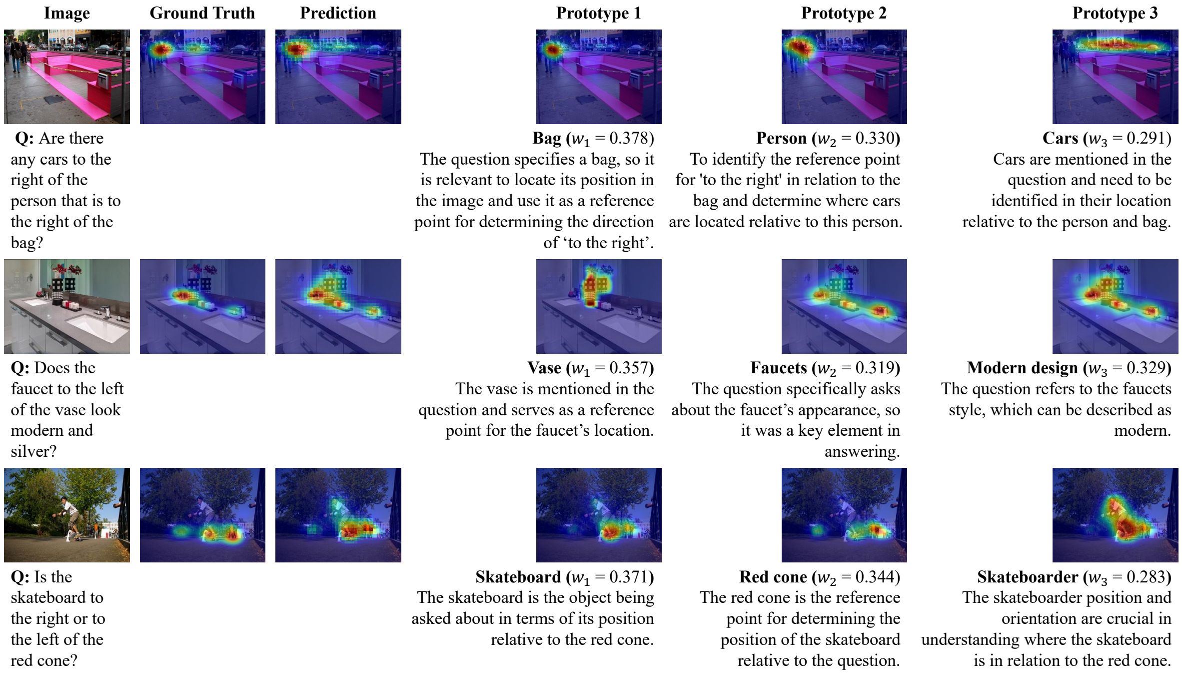


Figure 4. Qualitative examples from our methods. Each semantic prototype is shown along with their saliency weights and explanations.

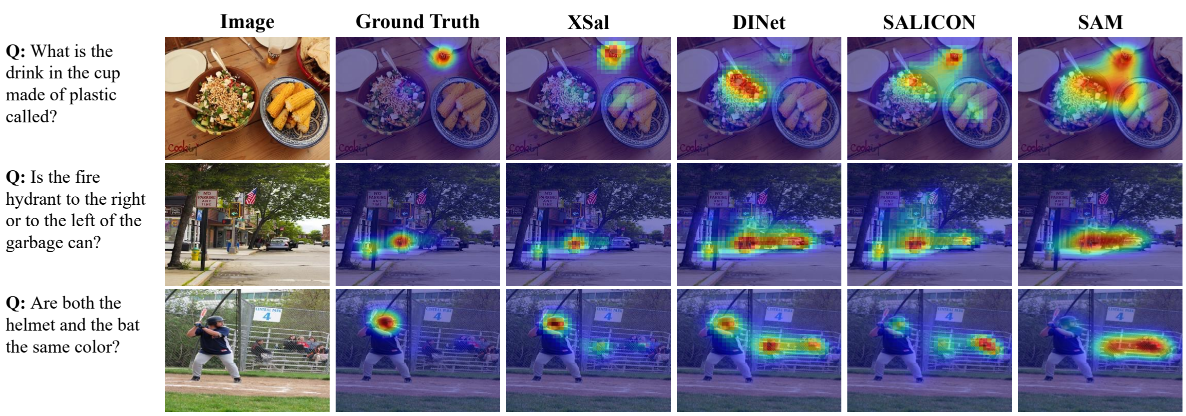


Figure 5. Qualitative comparison of saliency predictions across models (XSal, DINet, SALICON, SAM). Each row shows an input question and image, with the corresponding ground truth attention map and model outputs.

its adaptability to diverse scenarios. These examples highlight our model's superior performance in generating accurate and contextually relevant saliency predictions.

## 5. Conclusion

This paper has presented a novel approach for saliency prediction that combines explicit reasoning and contextual prioritization, generating both accurate and explainable saliency predictions. We have demonstrated that our method, powered by a VLM, outperforms existing models in terms of accuracy and explainability. This is achieved by leveraging an explicit reasoning process that allows the model to provide clear explanations for its predictions and a contextual prioritization mechanism that focuses on a few truly important elements, mimicking human attention behavior. Our approach not only achieves state-of-the-art saliency prediction accuracy but also offers valuable insights into the decision-making processes of saliency models. This advancement contributes to the development of more trustworthy and reliable saliency models.

## Acknowledgments

This work is supported by NSF Grant 2143197 and Fasikl Incorporated.

## References


[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[2] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 5

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019. 5

[4] Patrick Le Callet and Ernst Niebur. Visual attention and applications in multimedia technologies. *Proceedings of the Institution of Electrical Engineers*, 2013. 1

[5] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009. 2

[6] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 2020 Conference on Association for Computational Linguistics (ACL)*, 2020. 5

[7] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14093–14100. IEEE, 2024. 2

[8] Nuo Chen, Jin Xie, Jing Nie, Jiale Cao, Zhuang Shao, and Yanwei Pang. Attentive alignment network for multispectral pedestrian detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3787–3795, 2023. 1

[9] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[10] Shi Chen, Ming Jiang, and Qi Zhao. What do deep saliency models learn about visual attention? In *Advances in Neural Information Processing Systems*, pages 9543–9555. Curran Associates, Inc., 2023. 1, 2, 5, 6

[11] Shi Chen, Ming Jiang, and Qi Zhao. Deep learning to interpret autism spectrum disorder behind the camera. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–12, 2024. 1, 2

[12] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10876–10885, 2021. 5

[13] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging human attention in novel object captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 1, 2

[14] Xianyu Chen, Ming Jiang, and Qi Zhao. Self-distillation for few-shot image captioning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[15] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 5, 6

[16] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[17] Xianyu Chen, Jinhui Yang, Shi Chen, Louis Wang, Ming Jiang, and Qi Zhao. Every problem, every step, all in focus: Learning to solve real-world problems with integrated attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2024. 2

[18] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *TIP*, 27(10):5142–5154, 2018. 5, 6

[19] Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection. In *Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, 2024. 2

[20] Ning Ding, Ce Zhang, and Azim Eskandarian. Saliendet: A saliency-based feature enhancement algorithm for object detection for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 9(1):2624–2635, 2023. 1

[21] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *ACM Multimedia Systems Conference (MMSys)*, 2019. 1

[22] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[23] Songhao Han, Le Zhuo, Yue Liao, and Si Liu. Llms as visual explainers: Advancing image classification with evolving visual descriptions. *arXiv preprint arXiv:2311.11904*, 2023. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[25] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007. 2

[26] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015. 1, 2, 5

[27] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions*

*on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 2

[28] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[29] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 2

[30] Christopher Kanan, Mathew H Tong, Lingyun Zhang, and Garrison W Cottrell. Sun: Top-down saliency using natural statistics. *Visual cognition*, 17(6-7):979–1003, 2009. 2

[31] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 2

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[33] Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959. 5

[34] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *ICLR workshop*, 2015. 2

[35] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[36] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*, 494:455–467, 2022. 1

[37] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 4

[38] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 2

[39] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007. 5

[40] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference on North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. 5

[41] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005. 5

[42] Thammathip Piumsomboon, Gun Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *IEEE Symposium on 3D User Interfaces (3DUI)*, 2017. 1

[43] Yao Qiang, Deng Pan, Chengyin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining transformers via attentive class activation tokens. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 5

[44] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 5

[45] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. 1

[46] Yubao Sun, Mengyang Zhao, Kai Hu, and Shaojing Fan. Visual saliency prediction using multi-scale attention gated network. *Multimedia Systems*, 28(1):131–139, 2022. 2

[47] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[48] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 451–466. Springer, 2016. 2

[49] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 5

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2

[51] Jinhui Yang, Xianyu Chen, Ming Jiang, Shi Chen, Louis Wang, and Qi Zhao. VisualHow: Multimodal problem solving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[52] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *TMM*, 22(8):2163–2176, 2020. 1, 2, 5, 6, 7

[53] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3

[54] Wei Zhang and Hantao Liu. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, 26(3):1275–1288, 2017. 1