Panoptic Saliency Ranking

Zhenyu Wu[1] **Wei Wang**[2] **Min Li**[1*] **Chong Ma**[1]
Xun Gong[1] **Chenglizhao Chen**[3] **Aimin Hao**[4] **Shuo Li**[5]

School of Computing and Artificial Intelligence, Southwest Jiaotong University[1],
Harbin Institute of Technology (Shenzhen)[2], China University of Petroleum (East China)[3],
Beihang University[4], Case Western Reserve University[5]

Abstract

We introduce Panoptic Saliency Ranking (PSR), a new task that aims to estimate the visual saliency of all objects within a scene. Unlike salient object ranking, which focuses only on ranking salient regions, PSR extends the problem to the entire scene, providing a holistic and fine-grained understanding of visual saliency. To facilitate this task, we construct PSR18K, the first large-scale benchmark for panoptic saliency ranking. It contains 17,861 images with 146,460 annotated object instances spanning 13 superclasses and 376 fine-grained categories. In addition, PSR18K provides relation graph annotations to explicitly model semantic and spatial relationships among objects. We further propose an efficient Intrinsic Relation Graph Network (IRGNet), which reinterprets Transformer self-attention as an implicit relational matrix and seamlessly transforms it into an explicit relation graph, thereby obviating the need for additional relation modeling and significantly improving inference speed. Extensive experiments on PSR18K and three widely used SOR benchmarks demonstrate that our IRGNet achieves new state-of-the-art performance and faster inference with fewer parameters. Project Page: `https://sites.google.com/view/PSR18K`.

1 Introduction

Salient object ranking (SOR) is designed to model human visual attention, going beyond salient object detection by not only identifying salient regions but also capturing their relative saliency ranking. This capability is crucial for many high-level tasks, such as image captioning [29, 2], image compression [1], and autonomous driving [17].

Despite its success, SOR suffers from a fundamental limitation: it focuses exclusively on salient objects while treating all remaining regions as a homogeneous background. This coarse saliency ranking will limit its application in downstream tasks. For example, in the image compression task, assigning a uniform compression ratio to all non-salient regions inevitably leads to color distortion and blurring in large background regions (see Appendix E). Generally, most computer vision tasks require a holistic understanding of the saliency ranking of *all* objects in a scene rather than only salient objects. All of these call for a more fine-grained and holistic saliency ranking for comprehensive scene understanding.

In this paper, we introduce a new task termed *Panoptic Saliency Ranking (PSR)*. Unlike SOR, which focuses solely on ranking salient objects, PSR aims to predict the saliency ranking of *all objects* in the scene, as illustrated in Fig. 1(a). This key distinction allows PSR to provide a more comprehensive representation of visual saliency, thereby enabling more holistic scene understanding. To facilitate research on this new task, we introduce **PSR18K**, the first large-scale benchmark dataset

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

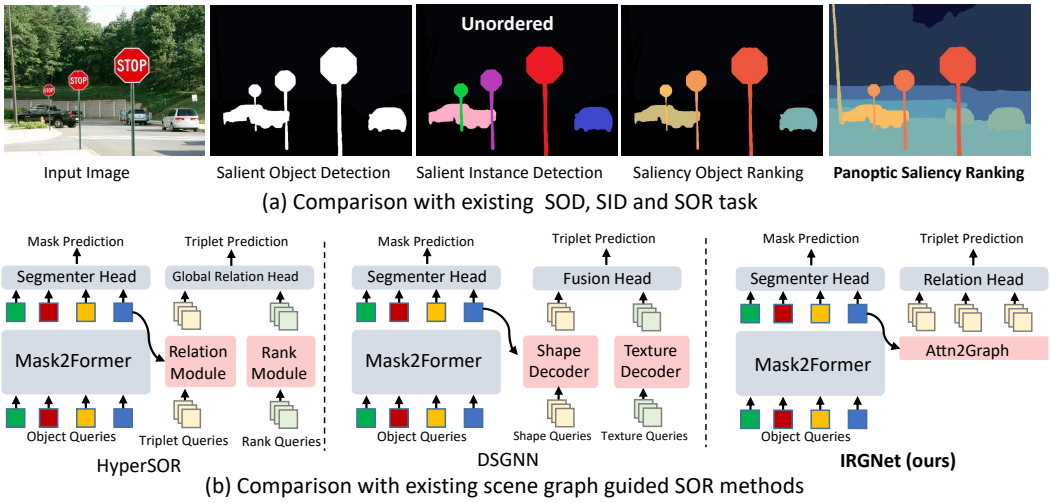


Figure 1: **Comparison at the task and method levels.** (a) The proposed PSR task ranks all objects in the scene rather than focusing solely on salient ones. (b) In contrast to prior approaches (e.g., HyperSOR [18] and DSGNN [26]) that rely on heavy relation extractors (369.4M and 488.4M parameters, respectively), our method directly derives relation graphs from self-attention, requiring only an additional 27.5M parameters.

for panoptic saliency ranking. It comprises 17,861 images with 146,460 annotated instances, spanning 13 superclasses and 376 fine-grained subcategories. In addition, we provide relation graph annotation for each image to facilitate explicit modeling of semantic and spatial relationships among objects, thereby enabling more accurate saliency ranking.

PSR requires not only accurate segmentation of each object instance but also the prediction of their saliency rankings, thereby necessitating a deeper and more structured understanding of inter-object relationships. Existing SOR methods can be broadly categorized into two types: relation-graph guided methods and direct saliency ranking methods. In particular, relation-graph guided approaches (e.g., HyperSOR [18] and DSGNN [26]) have demonstrated the superiority of explicitly modeling inter-object relationships for saliency ranking. As shown in Fig. 1(b), these methods typically adopt a two-branch architecture, where one branch performs object segmentation and the other predicts a relation graph, followed by a fusion network to produce the final saliency ranking. Despite their effectiveness, these approaches suffer from a common limitation: *they rely on sophisticated relation networks to predict inter-object relationships, leading to slow inference speed and higher computational overhead.*

To overcome this limitation, we propose an efficient Intrinsic Relation Graph Network (**IRGNet**), which reinterprets Transformer self-attention as an implicit relational matrix and seamlessly transforms it into an explicit relation graph, thereby obviating the need for additional relation modeling and significantly improving inference speed. The key insight is that the query and key tokens in self-attention can be interpreted as object queries, while the attention weights naturally encode pairwise relationships between objects. However, directly converting a dense self-attention matrix into a meaningful relation graph is non-trivial, as they contain abundant noisy and redundant connections. To address this issue, we further introduce a *Sparsity Gating* Module (SGM), which selectively distills dense pairwise interactions into a sparse and structured relation graph. Extensive experiments on PSR18K and three widely used SOR benchmarks demonstrate that our method achieves state-of-the-art performance while requiring significantly fewer parameters and offering substantially faster inference. In summary, our contributions are three-fold:

- **A New Task.** We introduce *Panoptic Saliency Ranking (PSR)*, a new task that predicts the saliency ranking of all objects in a scene, enabling a more holistic and fine-grained understanding of visual saliency.
- **A New Dataset.** We present *PSR18K*, the first large-scale benchmark for PSR, comprising 17,861 images with 146,460 annotated instances across 13 superclasses and 376 fine-grained categories, along with annotated relation graphs.



Figure 2: **Annotation pipeline of PSR18K**, consisting of five stages: stimuli presentation, gaze data collection, gaze data processing, saliency rank verification, and relation graph generation.

- **An Efficient Model.** We propose *IRGNet*, an efficient framework that leverages the intrinsic relational structure of Transformer self-attention to construct explicit relation graphs, eliminating the need for heavy relation extractors and achieving SOTA performance and faster inference.

2 Related Works

Salient object ranking (SOR) [26, 23, 4] requires not only identifying each salient object instance but also estimating the relative saliency of each object. As the pioneering work, [8] first addressed this task by ranking salient objects based on the level of agreement among multiple observers regarding which objects are perceived as salient. Building upon this idea, Siris *et al.* [20] proposed a deep learning framework that jointly predicts saliency ranks and the corresponding object masks by integrating bottom-up and top-down attention cues. Tian *et al.* [24] further introduced a bi-directional object–context prioritization framework for saliency ranking, leveraging both spatial and object-level attention mechanisms. Sun *et al.* [23] propose a partitioned saliency ranking paradigm, which segments unordered salient instances into partitions and then ranks them based on the correlations among these partitions. As scene graphs provide an intuitive representation for modeling inter-object relationships, recent saliency ranking approaches have commonly integrated scene graphs into their frameworks [26, 18]. Qiao *et al.* [18] propose a novel relation-aware hypernetwork-based SOR framework, which leverages scene context to guide saliency ranking via learning the scene graph in an explicit manner. Wu *et al.* [26] propose a domain separation graph neural network that incorporates both shape and texture information of objects while suppressing task-irrelevant noise. **Differences:** 1) Unlike SOR, which focuses solely on ranking salient objects, PSR aims to predict the saliency ranking of *all objects* in the scene. 2) Unlike previous approaches that rely on complex relation extractor, our method can directly generate relation graphs by leveraging the self-attention, leading to a simpler and more efficient architecture.

3 Proposed PSR18K

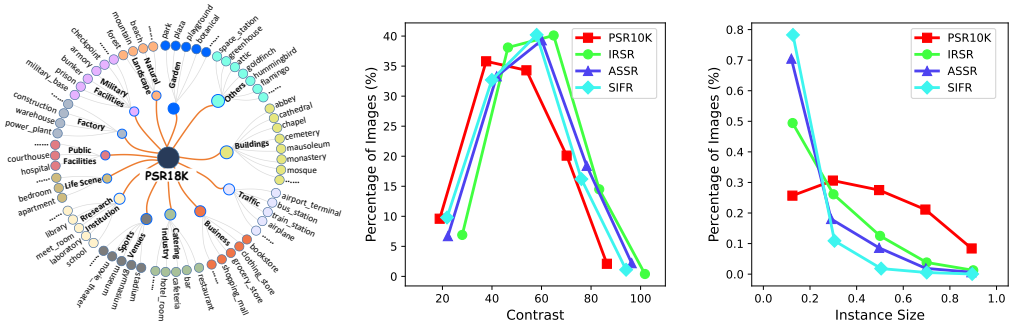
In this section, we introduce the proposed PSR18K dataset. We first describe the construction of the PSR18K dataset and then present its key statistics, such as object categories and instance sizes.

3.1 Eye-tracking-based Annotation

The proposed PSR18K dataset is built upon the ADE20K dataset [30]. Following prior work [4], we adopt an eye-tracking-based pipeline for saliency annotation, as illustrated in Fig. 2. **(i) Viewing Protocol.** Gaze data are collected using an Eyeso remote eye tracker operating at 150 Hz. Stimuli are displayed on a 24-inch monitor, with each image resized proportionally to fit the screen resolution. In contrast to previous studies [4], participants are instructed to attend to all objects in the scene rather than focusing solely on salient regions. No explicit time constraint is imposed, allowing natural viewing behavior and sufficient exploration of the scene. Twelve participants with normal or corrected-to-normal vision take part in the study. **(ii) Gaze Data Collection.** To ensure data quality, the eye tracker was recalibrated every 100 images, during which participants were encouraged to take short breaks to reduce visual fatigue. The entire data collection process lasted 12 months to ensure the quality of the data obtained from the participants. Participants' attention was assessed indirectly through standard eye-tracking statistics, including fixation duration distributions and spatial consistency across observers. Abnormally short fixations and unstable gaze patterns were filtered out

Table 1: **PSR18K vs. existing SOR datasets** in terms of the number of images, the number of instances, and the number of instances per image. Large: (instance size $\geq 30\%$), Medium: ($5\% \leq$ instance size $\leq 30\%$), Small: (instance size $\leq 5\%$).

Datasets	Publication Year	Total Images	Total Instances	Rank Everything	Numbers of Labeled Instance Per Image (%)								
					2	3	4	5	6	7	8	9	10+
ASSR [21]	$CVPR_{20}$	11500	49445	✗	9.8	14.0	12.5	63.6	-	-	-	-	-
IRSR [14]	$TPAMI_{21}$	8988	30176	✗	34.1	29.9	17.5	9.4	5.0	2.5	1.7	-	-
SIFR [4]	$CVPR_{24}$	8389	52173	✗	-	16.0	18.7	17.0	13.1	10.3	7.4	5.1	12.5
PSR18K	Submission	17861	146460	✓	-	-	2.3	8.3	13.5	21.5	22.1	14.4	17.9



(a) Category Distribution (b) Color Contrast (c) Instance Size

Figure 3: **Statistics of PSR18K dataset.** Compared to SOR datasets, the PSR18K dataset shows more diverse object categories, lower color contrast, and balanced distribution of instance sizes.

following established practice. **(iii) Gaze Data Processing.** Raw gaze trajectories were segmented into fixations and saccades using a velocity-based algorithm [19], and saccades were excluded from subsequent annotation. Following standard practice [4], fixations with durations shorter than 200 ms were discarded. To alleviate the central bias introduced by screen-centered stimulus presentation, the first fixation in each trial was removed. After preprocessing, fixation points from all participants were aggregated on a per-image basis to form human ground-truth saliency maps, where fixation density reflects perceived saliency ranks. **(iv) Saliency Rank Verification.** Due to the inherent randomness of gaze behaviors and inter-subject variability, raw eye-tracking data alone cannot always produce accurate and consistent saliency rankings for every image. To ensure the reliability of the rankings, we introduce a verification step in which three independent annotators review the saliency orderings derived from aggregated gaze data and correct only those cases with clearly erroneous. As shown in Fig. 2, although the car in the lower-right region is perceptually more salient than the surrounding ground, the aggregated fixation density mistakenly favors the ground region; such rare inconsistencies are rectified during verification. **(v) Relation Graph Generation.** To reduce annotation cost while ensuring high-quality labels, we adopt a human–AI collaborative annotation pipeline. Initial relation graphs are generated by GPT-5 and subsequently refined and verified by 12 human annotators to ensure accurate relation graph.

3.2 Statistics of PSR18K

Table 1 provides a quantitative comparison between PSR18K and widely used SOR benchmarks, including ASSR [21], IRSR [14], and SIFR [4]. In addition, Fig. 3 presents detailed dataset statistics, including the distributions of object categories, color contrast, and instance sizes.

(i) Comparison with SOR Benchmarks. As shown in Table 1, PSR18K contains 146,460 annotated instances, nearly tripling the scale of the largest existing SOR dataset (ASSR). Moreover, PSR18K exhibits a substantially higher instance number, with an average of 8.2 instances per image. Notably, 17.9 of images contain more than 10 instances, significantly exceeding that of existing SOR benchmarks. **(ii) Category Distribution.** As illustrated in Fig. 3(a), PSR18K spans 13 superclasses and 376 fine-grained subcategories, exhibiting nearly a fourfold increase in categorical diversity compared with ASSR, IRSR, and SIFR, which are all derived from 80 categories MS-COCO [11]. **(iii) Color Contrast.** Unlike previous approaches [4] that calculate color contrast across the entire image, we

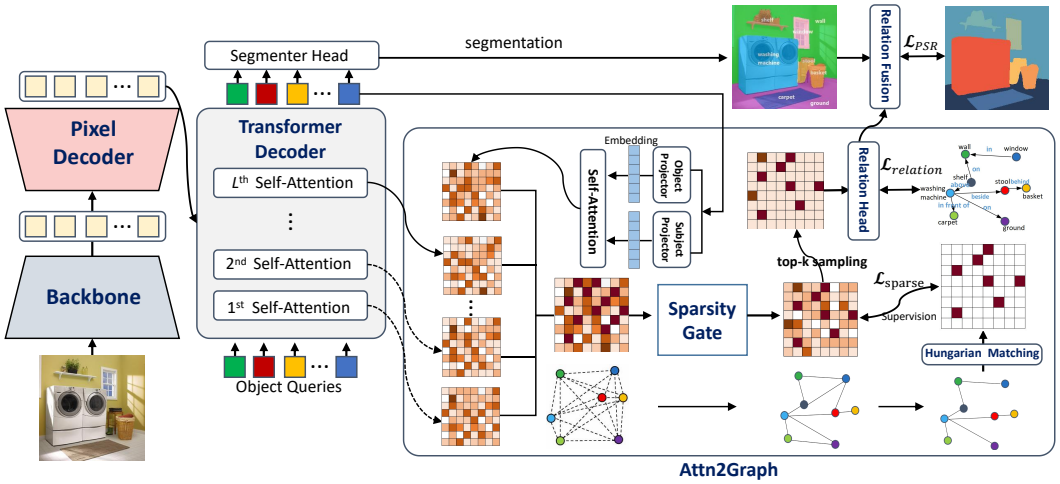


Figure 4: **Overall Architecture of IRGNet**. The framework comprises three key components: (a) Mask2Former serves as the segmentation backbone to produce segmentation maps; (b) the proposed relation extractor, Attn2Graph, transforms the inherent self-attention weights of the Transformer into an explicit relation graph; and (c) the relation fusion module integrates the constructed relation graph into the segmentation branch to yield the final saliency ranking.

propose computing color contrast between adjacent ranking objects, which better reflects the difficulty of ranking. The lower the color contrast of adjacent ranking objects, the greater the difficulty of ranking. As illustrated in Fig. 3(b), PSR18K exhibits consistently lower color contrast compared to existing SOR datasets. **(iv) Instance Size.** In contrast to existing SOR datasets that annotate only salient objects, PSR18K provides annotations for all objects within each image. Consequently, it exhibits a more balanced and diverse distribution of instance sizes.

4 Method

In this section, we first present an overview of the proposed IRGNet architecture. We then introduce the proposed relation extractor, Attn2Graph, which converts the intrinsic self-attention weights of Transformers into explicit relation graphs. Finally, we present a sparsity loss that facilitates the model to learn sparse yet high pairwise relations.

4.1 Model Architecture

As illustrated in Fig. 4, the proposed IRGNet comprises three components: (1) a segmentation network for panoptic segmentation, (2) a lightweight Attn2Graph module for relation extraction, and (3) a relation fusion module for final saliency ranking. *(i) Segmentation Network.* Following prior works [18, 26], we adopt Mask2Former [3] as the segmentation backbone. Specifically, the backbone first extracts low-resolution feature maps from the input image. The pixel decoder then progressively upsamples these features to produce high-resolution representations while maintaining computational efficiency. Finally, the Transformer decoder takes object queries together with the high-resolution features as input to generate the segmentation map. *(ii) Attn2Graph.* The proposed Attn2Graph can convert self-attention weights of the Transformer decoder into explicit relation graphs, eliminating the need for sophisticated relation extractors. To mitigate the gap between the dense self-attention matrix and the sparse nature of real-world relation graphs, we further introduce a Sparsity Gating Module (SGM), which enforces structured sparsification of attention-derived relations while preserving the most informative interactions. *(iii) Relation Fusion Module.* For the relation fusion module, we adopt the rank prediction graph (RPG) [18]. Different from the original design, we remove multi-path graph attention blocks since the relation graph is explicitly constructed by Attn2Graph. Instead, we retain only the hypernetwork guidance block, which utilizes the learned relation graph to guide relation-aware feature fusion for final saliency ranking.

4.2 Self-Attention is an Intrinsic Relation Graph

Motivation. Existing SOR methods [26, 18] typically rely on dedicated relation networks to model inter-object relationships. Despite their effectiveness, such designs inevitably introduce substantial parameters and computational overhead. In contrast, we observe that Transformer-based models inherently capture pairwise relationships among objects through the self-attention mechanism. Specifically, by replacing query and key tokens with object representations, the resulting attention scores naturally encode pairwise relations between objects. Motivated by this observation, we propose a lightweight relation extractor, Attn2Graph, which converts self-attention into an explicit relation graph.

Attn2Graph. Formally, given a set of object representations $F = \{f_1, \ldots, f_N\}$, the self-attention operation in layer l produces the attention matrix

$$A^l = \text{Softmax}\Big(\frac{Q^l (K^l)^\top}{\sqrt{d}}\Big), \quad Q^l = F^l W_Q^l,\ K^l = F^l W_K^l, \tag{1}$$

where $A^l \in \mathbb{R}^{N\times N}$ encodes pairwise interactions among all object queries. Intuitively, A^l serves as an *implicit relation graph*, where each entry A_{ij}^l reflects the relation strength from object i to j.

To capture relation information across different levels, we aggregate attention matrices from all L decoder layers. Empirically, we find that higher-layer attention maps capture more meaningful object-level relations, while lower-layer attention primarily encodes local dependencies (see Fig. 8). To enhance relation modeling, we additionally compute a self-attention relation matrix from the final decoder layer:

$$A^{L+1} = \text{SA}\big(F^{L+1} W_s, F^{L+1} W_o\big), \tag{2}$$

where W_s and W_o project object queries into subject and object spaces, and $\text{SA}(\cdot,\cdot)$ denotes self-attention operation.

However, directly converting these dense matrices into a relation graph is non-trivial. We observe a fundamental mismatch: self-attention produces fully connected pairwise dense relations ($\mathcal{O}(N^2)$ relations for N object queries), whereas the groundtruth relation graphs are sparse. For instance, in the PSR18K dataset with object queries $N = 100$, the density of annotated relations is only 2.5×10^{-4}. This sparsity poses a significant challenge for a naive relation head, which must identify a handful of meaningful edges among an overwhelming number of irrelevant pairs. To bridge this gap, we introduce a Sparsity Gating Module (SGM) that filters the dense pairwise relation map into a considerably sparser one and selectively emphasizes high relational edges while suppressing noisy connections.

Sparsity Gating Module. Given the dense relation map $A \in \mathbb{R}^{N\times N}$ derived from self-attention, our goal is to learn a sparse subgraph that preserves only meaningful relations. Let $\{f_i\}_{i=1}^L$ denote object queries from each decoder layer. For each pair (i, j), we construct a relation embedding

$$h_{ij} = \text{concat}(f_i, f_j, A_{ij}), \tag{3}$$

which jointly encodes object semantics and their interaction strength. We then estimate the reliability of each relation via an energy-based scoring function:

$$E_{ij} = W_2\, \sigma(W_1 h_{ij}), \tag{4}$$

where W_1 and W_2 are learnable parameters and $\sigma(\cdot)$ denotes ReLU. Intuitively, lower energy indicates higher relational plausibility. The energy is converted into a soft gating probability $g_{ij} = \sigma(-E_{ij})$, yielding a dense gating matrix $G \in (0, 1)^{N\times N}$. To obtain a sparse yet differentiable structure, we adopt a Gumbel-Top-k strategy. Specifically, we inject Gumbel noise into the gate logits:

$$\tilde{g}_{ij} = \frac{\exp((\log g_{ij} + \gamma_{ij})/\tau)}{\sum_{j'} \exp((\log g_{ij'} + \gamma_{ij'})/\tau)}, \tag{5}$$

where $\gamma_{ij} \sim \text{Gumbel}(0, 1)$ and τ is a temperature parameter. For each node i, only the top-k relations are retained:

$$S_i = \text{TopK}(\tilde{g}_{i,:}), \quad \hat{G}_{ij} = \begin{cases} 1, & j \in S_i \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Method	#Para.(M)	FPS	Backbone	Relation-Guided	PSR18K		
					SA-SOR ↑	SOR ↑	MAE ↓
PSR [23](MM 2023)	**225.3**	**18.9**	Swin-L	✗	0.486	0.534	0.138
SeqRank [7](AAAI 2024)	644.2	2.4	Swin-L	✗	0.436	0.527	0.166
QAGNet-U [4](CVPR 2024)	218.8	4.9	Swin-L	✗	0.525	0.576	0.156
PanoRank [22](AAAI 2025)	345.2	8.6	Swin-L	✗	0.521	0.583	0.173
DSGNN [26](CVPR 2024)	695.4	1.7	Swin-L	✓	0.543	0.598	0.152
HyperSOR [18](TPAMI 2024)	576.4	2.1	Swin-L	✓	0.565	0.626	0.142
IRGNet (Ours)	<u>234.5</u>	<u>18.7</u>	Swin-L	✓	**0.602**	**0.653**	**0.121**

Table 2: **Quantitative comparisons on PSR18K.** The best result is in **bold** and the second is <u>underlined</u>. The symbols ↑ and ↓ indicate "larger is better" and "smaller is better".

Finally, we aggregate gated relation representations across layers to produce the explicit relation graph:

$$\hat{R} = \sigma\left(\text{MLP}\left(\sum_{l=1}^{L} \hat{G}^l A^l + \hat{G}^{L+1} A^{L+1}\right)\right), \tag{7}$$

where $\hat{R}$ is the predicted relation graph, $\hat{G}^l$ denotes the sparsity gating at layer l, and MLP denotes a three-layer perceptron with ReLU activations.

4.3 Training and Inference

To train IRGNet, we adopt a multi-task learning strategy. In addition to saliency ranking and relation extraction, we introduce a novel sparsity loss to facilitate relation extraction. The overall objective is formulated as

$$\mathcal{L} = \lambda_0 \mathcal{L}_{PSR} + \lambda_1 \mathcal{L}_{relation} + \lambda_2 \mathcal{L}_{sparse}, \tag{8}$$

where $\mathcal{L}_{PSR}$ is the panoptic saliency ranking loss, $\mathcal{L}_{relation}$ is the relation extraction loss, $\mathcal{L}_{sparse}$ is a sparsity loss, λ_0, λ_1 and λ_2 are balancing coefficients.

Panoptic Saliency Ranking Loss: Following prior work [18, 26], we adopt a combination of binary cross-entropy and Dice loss to optimize saliency prediction. *Relation Extraction Loss:* To explicitly guide relation reasoning, we impose supervision on the predicted relation graph:

$$\mathcal{L}_{relation} = \text{BCE}(\hat{G}, G), \tag{9}$$

where $\hat{G}$ and G denote the predicted and groundtruth relation graphs, respectively. *Sparsity Loss:* To align with the intrinsic sparsity of real-world relations, we proposed a density-aware weighted binary cross-entropy loss:

$$\mathcal{L}_{sparse} = -\left[\frac{1}{\rho}\, y \log(\sigma(x)) + (1-y)\log(1-\sigma(x))\right], \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function, $y \in \{0, 1\}$ is the groundtruth label, and ρ denotes the empirical graph density. This formulation compensates for the extreme imbalance between positive and negative edges by up-weighting positive samples, thereby preventing trivial dense predictions and encouraging the model to learn sparse yet high pairwise relations.

5 Experiment

5.1 Experimental Settings

Datasets and Evaluation Metrics. We conduct experiments on the proposed *PSR18K* dataset. The PSR18K dataset is split into training, validation, and test sets with 12,503, 1,786, and 3,572 samples, respectively. In addition, we evaluate the generalization ability of our IRGNet on three widely used SOR benchmarks, including ASSR [21], IRSR [14], and SIFR [4]. For the SOR benchmarks, IRGNet is first pre-trained on the PSR18K training set and then fine-tuned on the target SOR datasets, with the Attn2Graph module frozen. For evaluation, we adopt three standard metrics: Salient Object Ranking (SOR) [8], Segmentation-Aware SOR (SA-SOR) [13], and Mean Absolute Error (MAE).

Method	#Para.(M)	ASSR			IRSR			SIFR		
		SA-SOR ↑	SOR ↑	MAE ↓	SA-SOR ↑	SOR ↑	MAE ↓	SA-SOR ↑	SOR ↑	MAE ↓
Swin-L [15] Backbone										
OCOR-U [24](CVPR 2022)	401.7	0.641	0.884	0.079	0.518	0.815	0.100	0.439	0.744	0.049
OCOR-L [24](CVPR 2022)	401.7	0.647	0.894	0.086	0.506	0.818	0.105	0.443	0.746	0.053
PSR [23](MM 2023)	225.3	0.746	<u>0.915</u>	0.071	<u>0.664</u>	<u>0.890</u>	0.075	0.595	0.783	0.052
SeqRank [7](AAAI 2024)	644.2	0.685	0.870	0.072	0.576	0.822	0.062	0.525	0.779	0.049
DSGNN [26](CVPR 2024)	695.4	0.761	0.853	<u>0.047</u>	0.606	0.807	0.067	0.586	0.765	0.047
QAGNet-U [4](CVPR 2024)	218.8	0.779	0.859	0.049	0.647	0.824	0.077	<u>0.621</u>	0.798	0.042
QAGNet-L [4](CVPR 2024)	218.8	<u>0.787</u>	0.856	0.048	0.647	0.824	0.077	<u>0.621</u>	<u>0.799</u>	<u>0.041</u>
LG-SOR [12] (CVPR 2025)	-	<u>0.787</u>	0.895	0.049	0.634	0.835	<u>0.050</u>	-	-	-
PanoRank [22] (AAAI 2025)	345.2	0.770	0.887	0.139	0.638	0.824	0.124	0.615	0.775	0.086
IRGNet (Ours)	**234.5**	**0.812**	**0.921**	**0.043**	**0.694**	**0.921**	**0.048**	**0.648**	**0.825**	**0.039**

Table 3: **Quantitative comparisons on SOR datasets.** "-L" and "-U" represent limited and unlimited model version. The symbols ↑ and ↓ indicate "larger is better" and "smaller is better". The best result is in **bold** and the second is <u>underlined</u>.

No.	Attn2Graph	SGM	$\mathcal{L}_{sparse}$	Para. (M)	FPS	PSR18K		
						SA-SOR ↑	SOR ↑	MAE ↓
Baseline	✗	✗	✗	207	20.4	0.505	0.556	0.164
Model ①	✔	✗	✗	222.3	19.5	0.559	0.614	0.146
Model ②	✔	✔	✗	234.5	18.7	0.584	0.631	0.132
Model ③	✔	✔	✔	234.5	18.7	**0.602**	**0.653**	**0.121**

Table 4: **Ablation Study**. Effectiveness of each component of the IRGNet.

Implementation Details. Our model is optimized by AdamW with an initial learning rate of 10^{-4}, a weight decay of 10^{-4}, and a total batch size of 16. Based on the ablation studies presented in Appendix D, we set Top-K to 15 in our experiments, and we set loss weights λ_0, λ_1, and λ_2 to 2, 1, and 1, respectively. We set the number of object queries to 100, size of embedding dimensions, d = 256 inheriting the design of Mask2Former. The subject projector and object projector are both MLPs with three fully connected layers, with embedding dimension d = 256 and ReLU as the activation function. We conducted 5 experiments, and the results exhibit a stable variance of ±0.01 in the SA-SOR metric. All experiments were conducted on a NVIDIA 4090 GPU.

5.2 Comparison to State-of-the-art Methods

We compare IRGNet with 10 state-of-the-art (SOTA) SOR methods, including IRSR [13], SOR [5], OCOR [24], PSR [23], SeqRank [7], DSGNN [26], QAGNet [4], HyperSOR [18], PanoRank [22], and LG-SOR [12].

Quantitative Comparison on PSR18K. As shown in Table 2, compared with relation-guided approaches (i.e., DSGNN and HyperSOR), IRGNet has approximately 3× fewer parameters, while achieving an average improvement of around 5% in terms of SA-SOR and delivering nearly 9× faster inference. IRGNet achieves substantially better performance than non-relation-guided methods (e.g., QAGNet and PSR), with an average improvement of approximately 10% in SA-SOR, while maintaining a comparable parameter count.

Despite the superiority of IRGNet, existing SOR methods are inherently limited in leveraging the relational graph annotations provided in PSR18K. Therefore, to ensure a fair and meaningful comparison, we further evaluate IRGNet against explicit relation modeling methods in Appendix B. As shown in Table 5, IRGNet achieves comparable or superior performance to explicit relation modeling methods while maintaining higher inference speed.

Quantitative Comparison on SOR Datasets. As shown in Table 3, IRGNet consistently outperforms all competing saliency ranking methods across all evaluation metrics. While PSR [23] and QAGNet [4] employ fewer parameters, IRGNet achieves superior performance, yielding improvements of approximately 6% and 2% in SA-SOR on the ASSR dataset, respectively. Additional visual comparisons on the SOR task are provided in Appendix C.

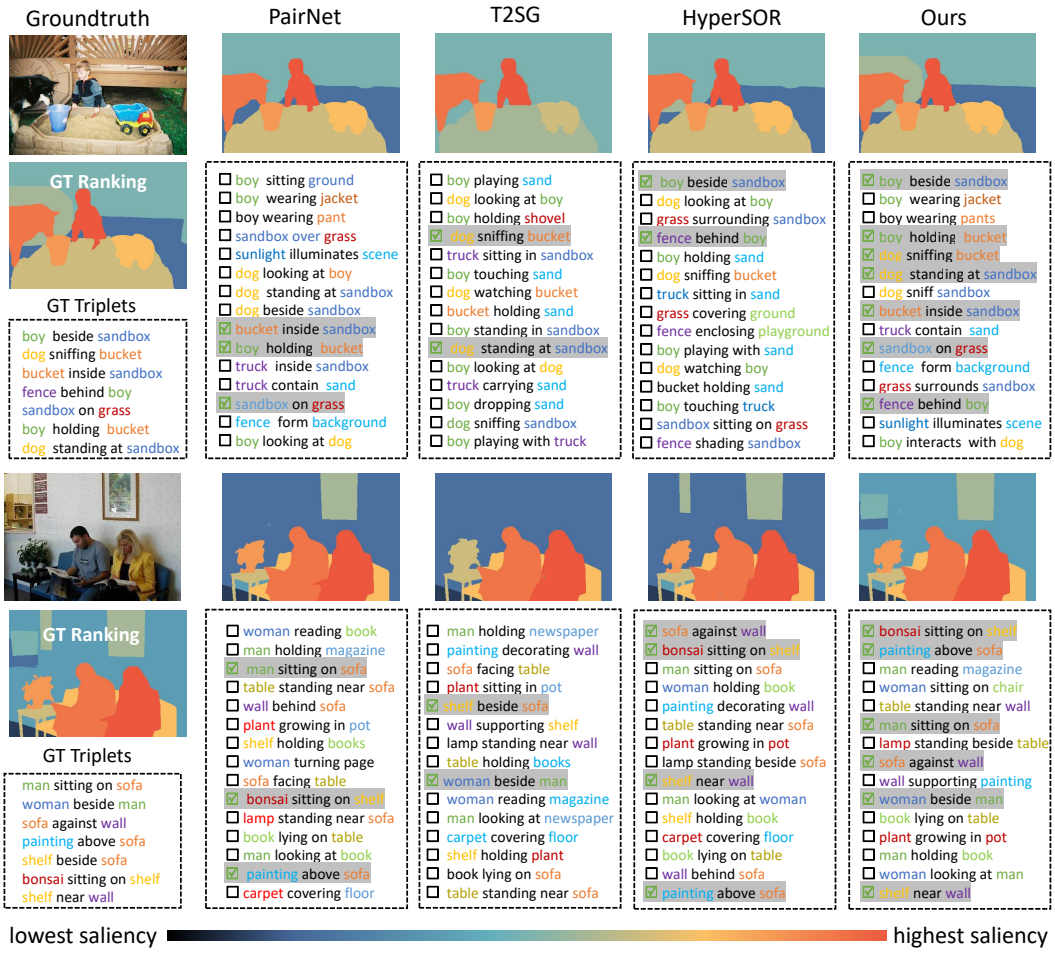


Figure 5: **Visualization of saliency ranking and the top 15 predicted triplets**. Triplets that match the groundtruth are marked by gray background.

Visualization Comparison. Fig. 5 presents saliency ranking results along with the predicted scene graphs, visualized as triplet lists. We compare IRGNet with graph-guided saliency ranking methods (e.g., HyperSOR) and explicit relation modeling approaches (e.g., T2SG [16] and PairNet [25]). Compared to these methods, IRGNet yields more accurate instance segmentation, more reliable saliency rankings, and more comprehensive triplet relation predictions.

5.3 Ablation Study

As shown in Table 4, we directly adopt Mask2Former as the baseline model and progressively incorporate the proposed modules to validate their effectiveness. More detailed ablation studies are provided in the Appendix D.

Effectiveness of Attn2Graph. We first incorporate Attn2Graph into the baseline model. As shown in Table 4, Model ① significantly improves the baseline by 5.4% and 5.8% in terms of SA-SOR and SOR on the PSR18K dataset, respectively, demonstrating the effectiveness of Attn2Graph.

Effectiveness of SGM. We further incorporate the SGM into the Model ①. As we can see, Model ② improves Model ① by approximately 2.5% and 1.7% in terms of SA-SOR and SOR metrics, which demonstrates the effectiveness of the SGM. See Appendix for visual analysis of SGM (Fig. 6).

Effectiveness of $\mathcal{L}_{sparse}$. Finally, we incorporate the proposed $\mathcal{L}_{sparse}$ into Model ②, forming the complete IRGNet. As shown in Table 4, Model ③ further improves upon Model ② by approximately

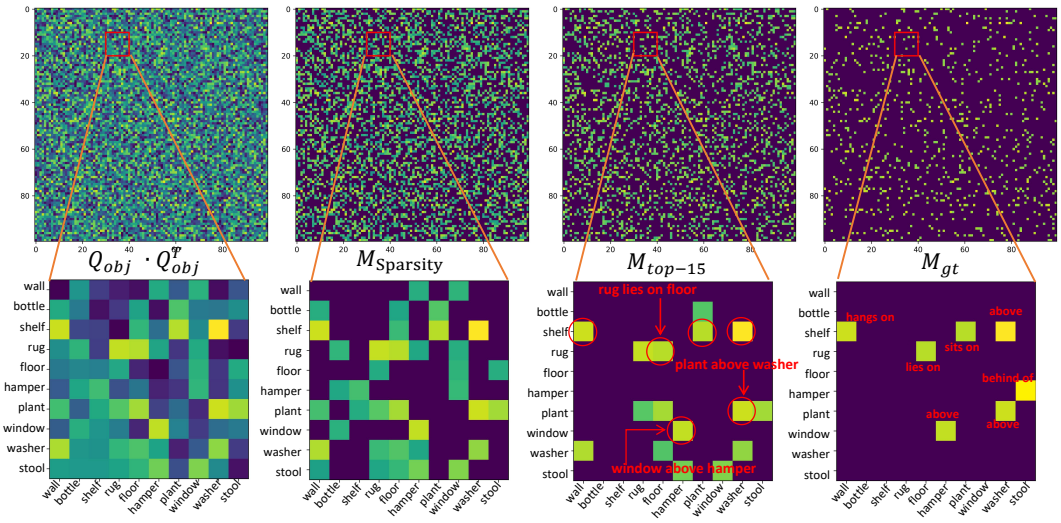


Figure 6: **Correlation Between Self-attention and Relation Graph.** From left to right: the self-similarity matrix computed by object queries ($Q_{obj}Q_{obj}^T$), the sparsified relation matrix $M_{sparsity}$ after the Sparsity Gating Module, the further refined matrix M_{top-k} obtained via the Gumbel-Top-k strategy, and the groundtruth relation matrix M_{gt}.

1.6% and 2.2% in terms of SA-SOR and SOR, demonstrating the effectiveness of the $\mathcal{L}_{sparse}$. In Appendix G, we discussed the limitations and future work.

5.4 Correlation Between Self-attention and Relation Graph

Figure 6 presents a visual comparison between the learned self-attention maps and the resulting relation graph, demonstrating how the proposed Attn2Graph and SGM effectively transform dense, unstructured attention weights into a sparse, interpretable graph. Specifically, the visualization illustrates that the model successfully identifies salient inter-object relationships—such as spatial proximity, semantic affinity, and functional interaction—while suppressing irrelevant or noisy connections. The high correspondence between the sparsified attention mask $M_{top\text{-}k}$ and the ground-truth relation matrix M_{gt} further validates that the SGM learns to approximate the underlying relational structure, rather than merely performing arbitrary edge pruning. This alignment confirms that the self-attention mechanism indeed encodes meaningful pairwise associations, which can be explicitly extracted and refined for saliency ranking.

6 Conclusion

In this paper, we introduced a new task termed Panoptic Saliency Ranking (PSR), which aims to predict the saliency ranking of all objects in a scene. To support this task, we constructed the PSR18K dataset, the first large-scale benchmark with panoptic annotations, saliency rankings, and relation graphs. Furthermore, we proposed IRGNet, an efficient intrinsic relation graph network that converts Transformer self-attention into explicit relation graphs. Extensive experiments on PSR18K and existing SOR benchmarks demonstrate that the proposed method achieves state-of-the-art performance with significantly fewer parameters and faster inference speed.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62501495 and 62501493, the Fundamental Research Funds for the Central Universities under Grant 2682026ZT007 and 2682025CX052, and the Sichuan Science and Technology Program under Grant 2025ZNSFSC1427. The authors declare no competing interests.

References

[1] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 507–515, 2016. 1

[2] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9962–9971, 2020. 1

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 5

[4] Bowen Deng, Siyang Song, Andrew P French, Denis Schluppeck, and Michael P Pound. Advancing saliency ranking with human fixations: Dataset models and benchmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28348–28357, 2024. 3, 4, 7, 8

[5] Hao Fang, Daoxin Zhang, Yi Zhang, Minghao Chen, Jiawei Li, Yao Hu, Deng Cai, and Xiaofei He. Salient object ranking with position-preserved attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16331–16341, 2021. 8

[6] Jiawei Fu, Tiantian Zhang, Kai Chen, and Qi Dou. Hybrid reciprocal transformer with triplet feature alignment for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 13

[7] Huankang Guan and Rynson WH Lau. Seqrank: Sequential ranking of salient objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1941–1949, 2024. 7, 8

[8] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7142–7150, 2018. 3, 7

[9] Jiankai Li, Yunhong Wang, Xiefan Guo, Ruijie Yang, and Weixin Li. Leveraging predicate and triplet learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28369–28379, 2024. 13

[10] Lin Li, Chuhan Zhang, Dong Zhang, Chong Sun, Chen Li, and Long Chen. Interaction-centric knowledge infusion and transfer for open-vocabulary scene graph generation. In *NeurIPS*, 2025. 13

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[12] Fang Liu, Yuhao Liu, Ke Xu, Shuquan Ye, Gerhard Petrus Hancke, and Rynson WH Lau. Language-guided salient object ranking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29803–29813, 2025. 8

[13] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8321–8337, 2021. 7, 8

[14] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8321–8337, 2022. 4, 7

[15] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 8

[16] Changsheng Lv, Mengshi Qi, Liang Liu, and Huadong Ma. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17197–17206, June 2025. 9, 13

[17] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: The dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018. 1

[18] Minglang Qiao, Mai Xu, Lai Jiang, Peng Lei, Shijie Wen, Yunjin Chen, and Leonid Sigal. Hypersor: Context-aware graph hypernetwork for salient object ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):5873–5889, 2024. 2, 3, 5, 6, 7, 8

[19] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78, 2000. 4

[20] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Inferring attention shift ranks of objects for image saliency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12133–12143, 2020. 3

[21] Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, and Rynson W.H. Lau. Inferring attention shift ranks of objects for image saliency. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12130–12140, 2020. 4, 7

[22] Jia Song, Chenglizhao Chen, Xu Yu, and Shanchen Pang. Fine-grained perception in panoramic scenes: A novel task, dataset, and method for object importance ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6950–6958, 2025. 7, 8

[23] Chengxiao Sun, Yan Xu, Jialun Pei, Haopeng Fang, and He Tang. Partitioned saliency ranking with dense pyramid transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1874–1883, 2023. 3, 7, 8

[24] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5882–5891, 2022. 3, 8

[25] Jinghao Wang, Zhengyu Wen, Xiangtai Li, Zujin Guo, Jingkang Yang, and Ziwei Liu. Pair then relation: Pair-net for panoptic scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10452–10465, 2024. 9, 13

[26] Zijian Wu, Jun Lu, Jing Han, Lianfa Bai, Yi Zhang, Zhuang Zhao, and Siyang Song. Domain separation graph neural networks for saliency object ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3964–3974, 2024. 2, 3, 5, 6, 7, 8

[27] Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, and Yaqi Xie. Hiker-sgg: Hierarchical knowledge enhanced robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28233–28243, 2024. 13

[28] Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12118–12128, 2026. 15

[29] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European conference on computer vision*, pages 211–229. Springer, 2020. 1

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 3

A Summary of the Appendix

To facilitate a deeper understanding of the main paper, we present supplementary material with additional details, organized as follows:

- Appendix B further compare our IRGNet with explicit relation modeling methods.
- Appendix C provided addtional visual results on SOR task.
- Appendix D provided addtional ablation study on PSR18K dataset.
- Appendix E provided typical application scenarios of PSR.
- Appendix G discussed the limitations and future work.

B Additional Quantitative Results on PSR Task

Comparison with explicit relation modeling methods. Although our proposed IRGNet significantly outperforms existing SOR methods on the PSR18K dataset, these SOR approaches are unable to leverage the relational graph annotations provided in PSR18K. Therefore, in Table 5, we further evaluate IRGNet against SOTA Comparison with explicit relation modeling methods (ACC [10], T2SG [16], HQSG [6], PairNet [25], HiKER-SGG[27], DRM [9], which can leverage the relational graph annotations to improve model performance.

As shown in Table 5, while ACC [10] achieves a slightly higher accuracy (SA-SOR: 0.611 vs. 0.602 for our method), our approach demonstrates a more favorable speed–accuracy trade-off. Unlike explicit relation modeling methods (e.g., ACC) that rely on additional relation extractors, IRGNet leverages the intrinsic relational structure already encoded in self-attention, resulting in a more principled and efficient formulation. Thus, under the same backbone configuration (Swin-L), IRGNet significantly reduces computational overhead, operating at 18.7 FPS compared to ACC's 5.7 FPS, yielding nearly three times higher inference throughput. Moreover, excluding ACC [10], IRGNet achieves SOTA performance across all evaluation metrics on the PSR18K dataset.

Method	**#Para.(M)**	**FPS**	**Backbone**	**Relation-Guided**	**PSR18K**		
					SA-SOR ↑	SOR ↑	MAE ↓
HiKER-SGG[27](CVPR 2024)	425.8	5.6	Swin-L	✔	0.563	0.581	0.141
DRM [9](CVPR 2024)	673.7	2.2	Swin-L	✔	0.574	0.604	0.134
PairNet [25](TPAMI 2024)	<u>357.4</u>	<u>6.1</u>	Swin-L	✔	0.564	0.603	0.135
HQSG [6](CVPR 2025)	642.5	3.6	Swin-L	✔	0.583	0.631	0.131
T2SG [16](CVPR 2025)	476.1	4.3	Swin-L	✔	0.597	0.643	0.125
ACC [10](NeurIPS 2025)	586.3	5.7	Swin-L	✔	**0.611**	<u>0.648</u>	**0.119**
IRGNet (Ours)	**234.5**	**18.7**	Swin-L	✔	<u>0.602</u>	**0.653**	<u>0.121</u>

Table 5: **Comparison with Panoptic Scene Graph Methods.** The best result is in **bold** and the second is <u>underlined</u>. The symbols ↑ and ↓ indicate "larger is better" and "smaller is better".

C Additional Visual Results on SOR Task

Visual Comparison. In Fig. 7, we provide a comprehensive visual comparison across several challenging scenarios, including scenes with multiple interacting objects, significant variations in object scales, low contrast between foreground and background, and cluttered or complex foreground compositions. These scenarios are particularly difficult as they require the model to not only precisely delineate object boundaries but also to correctly infer relative saliency under ambiguous visual cues. As illustrated in the Fig. 7, existing methods often suffer from issues such as blurred boundaries, missing small yet salient objects, or incorrect ranking orders when objects exhibit subtle saliency differences. In contrast, our IRGNet consistently generates more accurate saliency maps with sharper object boundaries and clearer structural details. Moreover, it demonstrates a stronger capability in preserving fine-grained regions and capturing less prominent but semantically important instances. Benefiting from the intrinsic relation modeling, IRGNet is able to better distinguish inter-object relationships, leading to more reliable and coherent saliency ranking results across diverse and complex scenes.

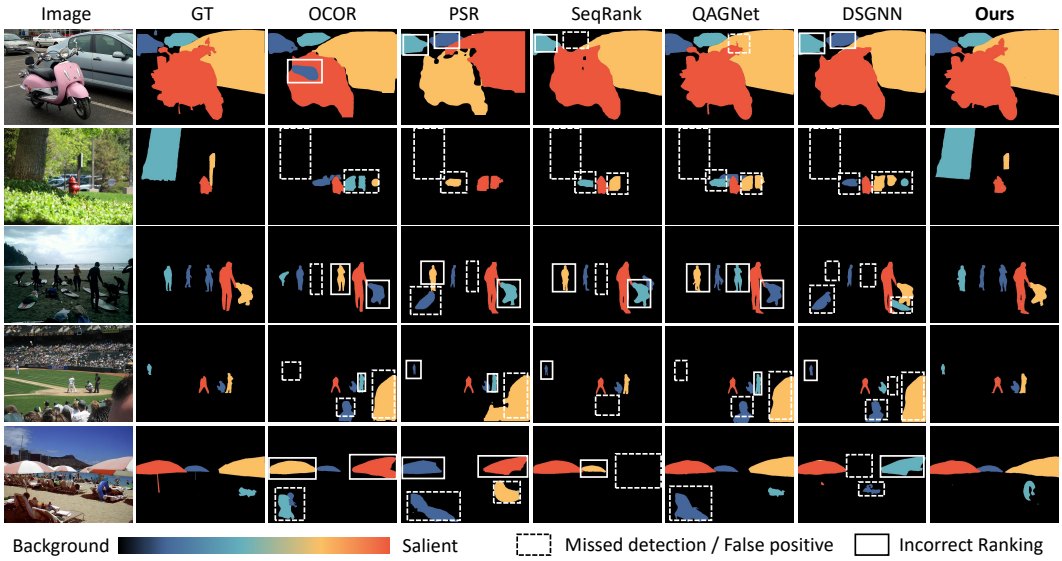


Figure 7: **Visual Comparison on SOR datasets**. Our IRGNet produces more accurate saliency maps with sharp boundaries and correct ranking orders..

$(\lambda_0, \lambda_1, \lambda_2)$	PSR18K		
	SA-SOR ↑	SOR ↑	MAE ↓
(1, 1, 1)	0.597	0.648	0.127
(2, 1, 1)	**0.602**	**0.653**	**0.121**
(1, 2, 1)	0.586	0.644	0.123
(1, 2, 2)	0.591	0.646	0.123
(2, 2, 4)	0.595	0.635	0.125

(a) Different loss weights.

Top-K	PSR18K		
	SA-SOR ↑	SOR ↑	MAE ↓
5	0.576	0.631	0.132
10	0.585	0.642	0.128
15	**0.602**	**0.653**	**0.121**
20	0.598	0.650	0.121
25	0.596	0.636	0.122

(b) Different Top-K values.

Table 7: Ablation studies on loss weights and Top-K selection on PSR18K.

D Additional Results of Ablation Study

Effect of Different Loss Weights. In Table 7(a), we vary the weights of different components in the loss function. The results show that different weight settings have no significant impact on the final performance, indicating that our model is robust to variations in loss weighting.

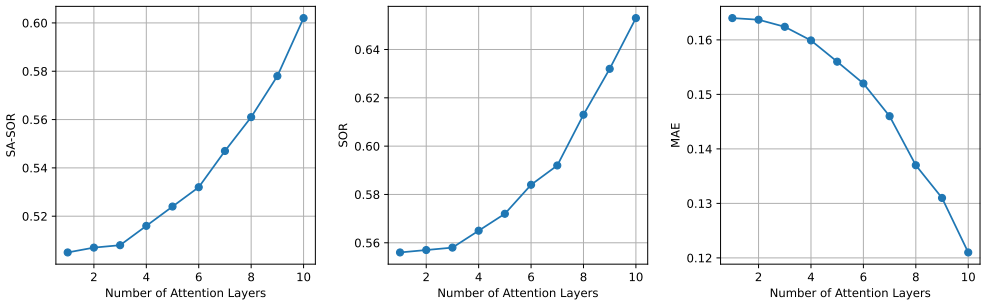
Effect of Different Top-K. In Table 7(b), varying the Top-K value has a noticeable impact on performance. When Top-K is set to 15, the model achieves the best results across all metrics, including SA-SOR, SOR, and MAE. Increasing or decreasing Top-K from this value leads to consistent performance degradation. Specifically, smaller Top-K values (e.g., 5) may limit the model's ability to capture sufficient relational information, while larger values (e.g., 20 and 25) introduce redundant or less informative relations, which can negatively affect ranking accuracy. These results indicate that a moderate Top-K strikes a balance between capturing informative relations and suppressing noise. Therefore, we set Top-K to 15 in our final model.

Different Number of Object Queries. In Table 6, we vary the number of object queries from 50 to 100 and 200. The results show that the performance remains largely stable across different settings, indicating that our model is robust to the number of object queries. Unless otherwise specified, we set the number of object queries to 100 in all experiments.

Object Queries	PSR18K		
	SA-SOR ↑	SOR ↑	MAE ↓
50	0.592	0.645	0.125
100	**0.602**	**0.653**	**0.121**
200	0.596	0.647	0.123

Table 6: Different numbers of relation queries.

Effect of Different Attention Layer. As shown in Fig. 8, we observe that performance gains become more pronounced when deeper attention layers



(a) SA-SOR vs Attention Layers. (b) SOR vs Attention Layers. (c) MAE vs Attention Layers.

Figure 8: **Effect of Different Attention Layer.** Deeper layers capture more semantic and global relations while earlier layers tend to focus on local patterns and short-range dependencies.

Method	ASSR				
	BLEU ↑	METEOR ↑	ROUGE-L ↑	CIDE ↑	SPICE ↑
Baseline	0.0576	0.2966	0.2454	0.0883	0.1791
PS-guided	0.0654	0.3012	0.2515	0.1012	0.1973
IS-guided	0.0798	0.3247	0.2601	0.1361	0.2346
SOR-guided	0.0795	0.3285	0.2589	0.1395	0.2298
PSR-guided	0.0847	0. 3514	0.2736	0.1651	0.2658

Figure 9: Image Captioning Guided by PS, IS, PSR and SOR. PS:Panoptic Segmentation, IS: Instance Segmentation

are incorporated, suggesting that deeper layers capture more semantic and global relations. In contrast, earlier layers tend to focus on local patterns and short-range dependencies, which are insufficient for capturing global relational structures. As the depth increases, the attention mechanism progressively encodes more abstract and object-level interactions, enabling more reliable relational reasoning.

E Applications

Image Compression. We provid quantitative results for the image compression application. Specifically, we adopt AEIC [28] as the compression baseline and compare with PS, IS SOR-guided and PSR-guided variants. As shown in Table 8, PSR-guided compression consistently improves over both the baseline and PS, IS and SOR approaches. This is because PSR explicitly provides the perceptual importance of different regions, enabling the compression model to allocate higher bit rates to visually important regions while aggressively compressing less important areas. In contrast, PS and IS only describe the semantic categories and instance structures, but cannot indicate which regions deserve more visual fidelity. In Fig. 10, we compare the visual quality of images compressed to half of their original size under the guidance of SOR and PSR. We observe that SOR-guided compression often leads to noticeable artifacts, including color distortion and blurring in large background regions. In contrast, PSR-guided compression preserves fine-grained details and maintains higher color fidelity across the entire image.

Image Captioning. We adopt Qwen2.5-VL-7B as the baseline model. As shown in the table 9, PSR also achieves better performance than SOR, PS and IS. The main reason is that PSR provides explicit saliency ranking information, guiding the model to focus more on visually important objects and generate descriptions centered around perceptually meaningful regions. Since Qwen2.5-VL-7B is already a powerful multimodal large language model with strong semantic understanding capabilities, the additional semantic information provided by PS and IS offers limited complementary benefits.

Method	ASSR		
	Bpp ↓	LPIPS ↓	DISTS ↓
Baseline(AEIC)	0.132	0.275	0.128
PS-guided	0.129	0.276	0.125
IS-guided	0.122	0.261	0.114
SOR-guided	0.112	0.246	0.115
PSR-guided	0.104	0.235	0.107

Table 8: Image Compression Guided by PS, IS, PSR and SOR. PS:Panoptic Segmentation, IS: Instance Segmentation

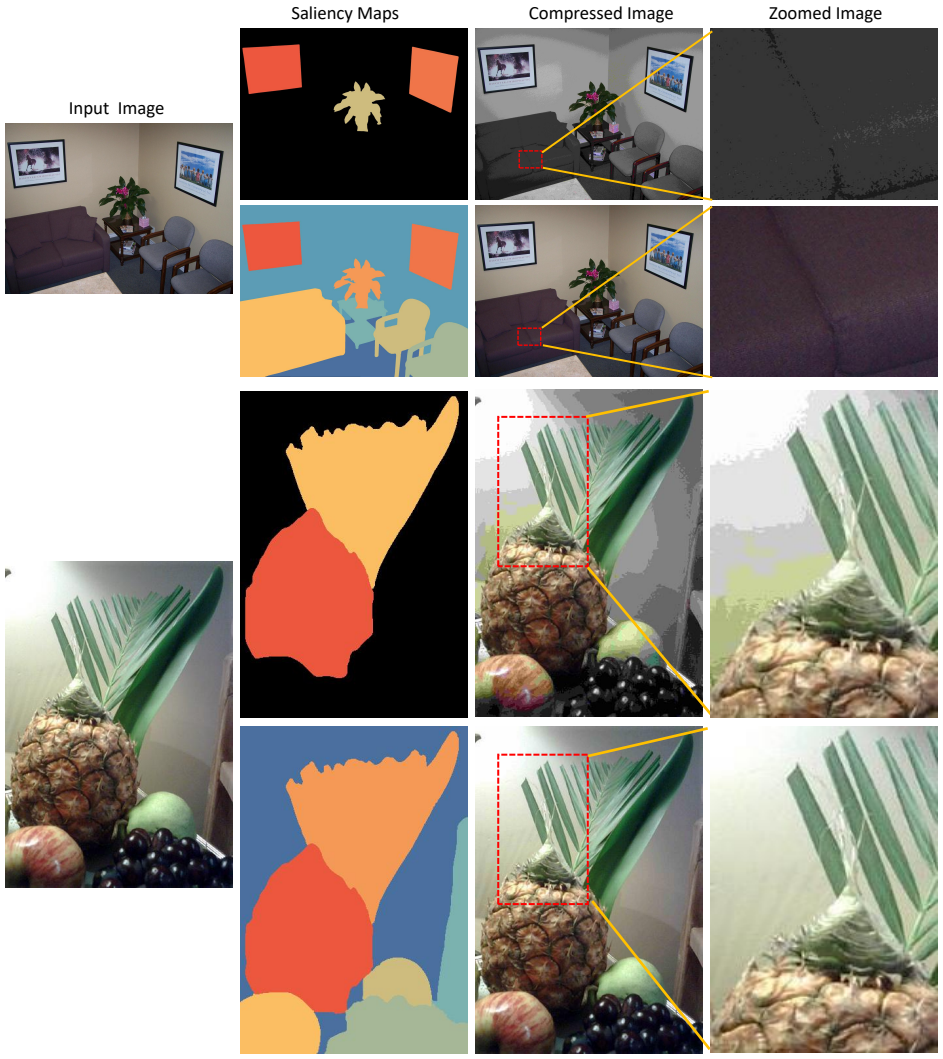


Figure 10: **Comparison of PSR and SOR for image compression**. Compared to SOR, PSR-guided compression better preserves fine-grained details and maintains higher color fidelity across the entire image.

F More Details

Details of the relation graph annotation. In this section, we provide the details about the relation graph annotation pipeline. Specifically, (1) Prompt. GPT-5 was used only to generate preliminary relationship proposals. The prompt used to instruct GPT-5 is: "Please describe the relationships between the positions and affiliations of all of objects in the imag". (2) Guidelines. After obtaining the preliminary relationship graphs, human annotators performed a two-stage verification process according to predefined annotation guidelines: (i) Completeness verification: checking whether all object instances in the image are included in the generated relationship graph. (ii) Correctness verification: examining whether the generated relationships accurately reflect their relations. (3) Statistics. According to our statistics, 78.1% of generated relationship graphs satisfy the completeness criterion, while 82.4% of the generated graphs satisfy the correctness criterion .Since the relationships between objects are relatively clear, we did not observe substantial disagreement among annotators. Ambiguous cases were resolved through discussion and consensus. (4) Hallucination. We also emphasize that GPT-5 only serves as an annotation acceleration tool, while the final annotations are produced and validated by human annotators. Therefore, the potential bias or hallucination introduced by GPT-5 does not directly affect the final annotation quality.

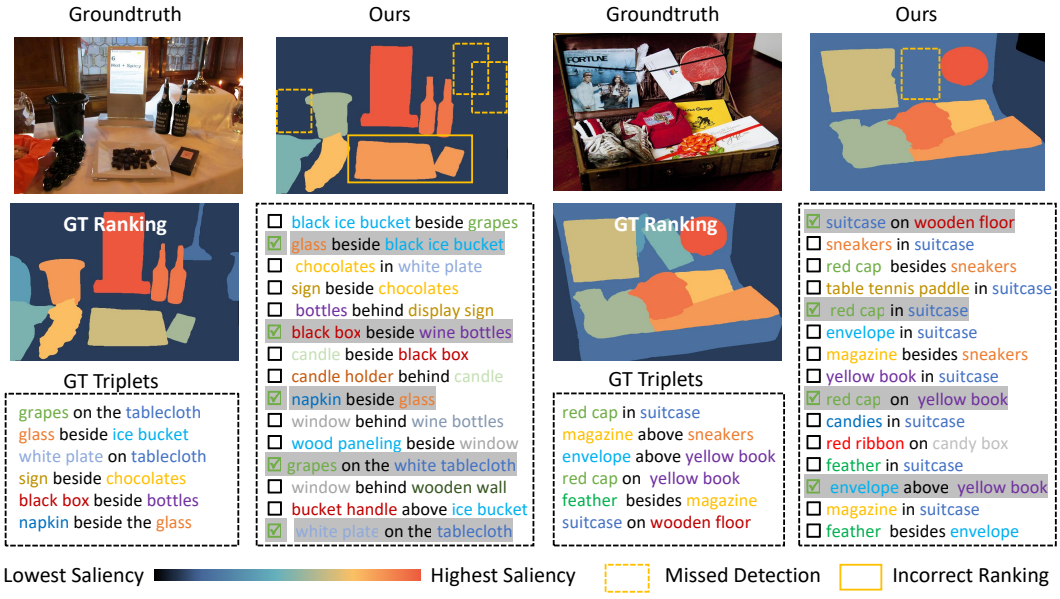


Figure 11: **Failure Cases**. Triplets that match the groundtruth are marked by gray background.

Details of training/testing protocols. For the PSR task, to ensure a fair comparison, we follow the official implementations and training protocols of all compared methods. Specifically, all methods are trained and evaluated on the PSR18K dataset. The PSR18K dataset is split into 12,503 training samples, 1,786 validation samples, and 3,572 test samples. Note that DSGNN, HyperSOR, and our IRGNet utilize relation graphs as additional supervisory signals according to their original formulations, and we report their results under the corresponding settings. For the SOR task, we follow previous works and evaluate all methods using the standard splits of ASSR, IRSR, and SIFR datasets. The ASSR dataset contains 7,646 training images, 1,436 validation images, and 2,418 test images, where each image is annotated with up to five salient instances ranked by saliency levels. The IRSR dataset contains 6,059 training images and 2,929 test images, with up to eight ranked salient instances per image. The SIFR dataset contains 8,389 images, including 6,701 training images and 1,688 testing images.

G Limitations and Future Work

Limitations. Despite achieving state-of-the-art performance, IRGNet still faces challenges in complex visual scenarios where inter-object saliency differences are subtle. As illustrated in Fig. 11, under such conditions, the model is prone to both missed segmentations and incorrect relational ordering, even when the overall structure of the predicted relation graph remains largely reasonable. These limitations can be attributed to two main factors: (1) the Mask2Former backbone may be insufficient for precise delineation of fine-grained regions in complex visual scenarios, and (2) although IRGNet effectively captures pairwise relations, it lacks more advanced structured reasoning mechanisms to resolve perceptual ambiguities and enforce global consistency.

Future Work. Several promising directions warrant further investigation. First, inspired by the strong performance of the Segment Anything Model (SAM) family in fine-grained segmentation, an important extension is to replace the Mask2Former backbone with a SAM-based variant to improve segmentation quality in visually ambiguous regions. Second, to enhance relational reasoning, future work could explore integrating pretrained large language models (LLMs) or multimodal foundation models, which may provide structured commonsense knowledge and contextual priors to better resolve ambiguities and enforce global consistency. Finally, current evaluation protocols rely on standard SOR metrics, including SA-SOR, SOR, and MAE, which are primarily designed for ranking salient objects. These metrics are not fully aligned with the holistic and fine-grained nature of PSR. Developing dedicated evaluation metrics tailored to panoptic saliency ranking remains an important open problem.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper provides the full set of assumptions and a complete experimental proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data used in this paper have been available on $GitHub$. If once the paper is accepted, we will provide detailed links.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The full details are provided with the main context and supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports the statistical significance of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper describes safeguards that have been put in place for responsible release of data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Meet the requirements.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Meet the requirements.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not Applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not Applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Editing (e.g., grammar, spelling, word choice)

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.