Language-Guided Salient Object Ranking

Fang Liu Yuhao Liu Ke Xu Shuquan Ye Gerhard Petrus Hancke Rynson W.H. Lau
Department of Computer Science, City University of Hong Kong
{fawnliu2333, yuhaoLiu7456, kkangwing, shuquanye}@gmail.com
{gp.hancke, Rynson.Lau}@cityu.edu.hk

Abstract

Salient Object Ranking (SOR) aims to study human attention shifts across different objects in the scene. It is a challenging task, as it requires comprehension of the relations among the salient objects in the scene. However, existing works often overlook such relations or model them implicitly. In this work, we observe that when Large Vision-Language Models (LVLMs) describe a scene, they usually focus on the most salient object first, and then discuss the relations as they move on to the next (less salient) one. Based on this observation, we propose a novel Language-Guided Salient Object Ranking approach (named LG-SOR), which utilizes the internal knowledge within the LVLM-generated language descriptions, i.e., semantic relation cues and the implicit entity order cues, to facilitate saliency ranking. Specifically, we first propose a novel Text-Guided Visual Modulation (TGVM) module to incorporate semantic information in the description for saliency ranking. TGVM controls the flow of linguistic information to the visual features, suppresses noisy background image features, and enables the propagation of useful textual features. We then propose a novel Text-Aware Visual Reasoning (TAVR) module to enhance model reasoning in object ranking, by explicitly learning a multimodal graph based on the entity and relation cues derived from the description. Extensive experiments demonstrate superior performances of our model on two SOR benchmarks.

1. Introduction

The Salient Object Ranking (SOR) task is recently proposed to study how humans shift their attention across different objects in a scene. By mimicking how humans sequentially perceive the scene, SOR models can facilitate many downstream tasks, including autonomous driving [12, 47], important people detection [46, 59], and scene understanding [15].

Siris *et al.* [48] first propose to model the relations among salient objects and global scene context for saliency rank prediction across the objects. Tian *et al.* [53] propose to jointly model object-based and spatial attention for inferring the saliency rank. Sun *et al.* [51] propose to first partition salient objects into different groups and then model their relations based on a dense pyramid transformer. Guan and Lau [14] explore human pose cues to learn high-level interactions between humans and surrounding objects for ranking prediction. Despite their success, these methods often produce inconsistent saliency ranks with respect to those from humans. For example, existing methods may predict a higher saliency rank for the relatively larger motorcycle rather than the rider (Fig. 1(A)), or neglect some small salient objects but include those noisy background objects (Fig. 1(B)). The main reason for this limitation is that unlike humans who can easily identify/associate high-level semantic relations among objects (*e.g.*, "*riding*" and "*playing table tennis*" for the two examples in Fig. 1) when they determine the viewing order, existing methods do not consider such cues explicitly.

We observe that when Large Vision-Language Models (LVLMs) [5, 19, 35] describe a scene, they tend to start with the most salient object in the image and interpret rich semantic information (*i.e.*, attributes and relationships). For example, the caption generated by an LVLM [5] for Fig. 1(A) starts from the most salient object (*i.e.*, *"A man"*), and describes not only the spatial relation (*i.e.*, *"next to"*) but also the semantic relation (*i.e.*, *"ride"*) to the other object (*i.e.*, *"motorcycle"*). In Fig. 1(B), we can also see that the LVLM [5] describes the semantic relation (*i.e.*, *"playing table tennis"*) and mimics humans' attention shifts to describe objects sequentially, *i.e.*, from *"one holding a paddle"* to *"the other standing in front of him"* and *"other people observing the game"*. This observation inspires us to incorporate the semantic relations and implicit orders from the descriptions generated by LVLMs for salient object ranking.

Based on the above observation, we propose in this paper a novel language-guided salient object ranking method (called LG-SOR), which takes an image and a language description (generated by an LVLM) as input and learns to condition the ranking process on the extracted visual stimuli and semantic guidance. LG-SOR includes two novel modules. First, we propose a novel Text-Guided Visual Modulation (TGVM) module to incorporate the semantic information

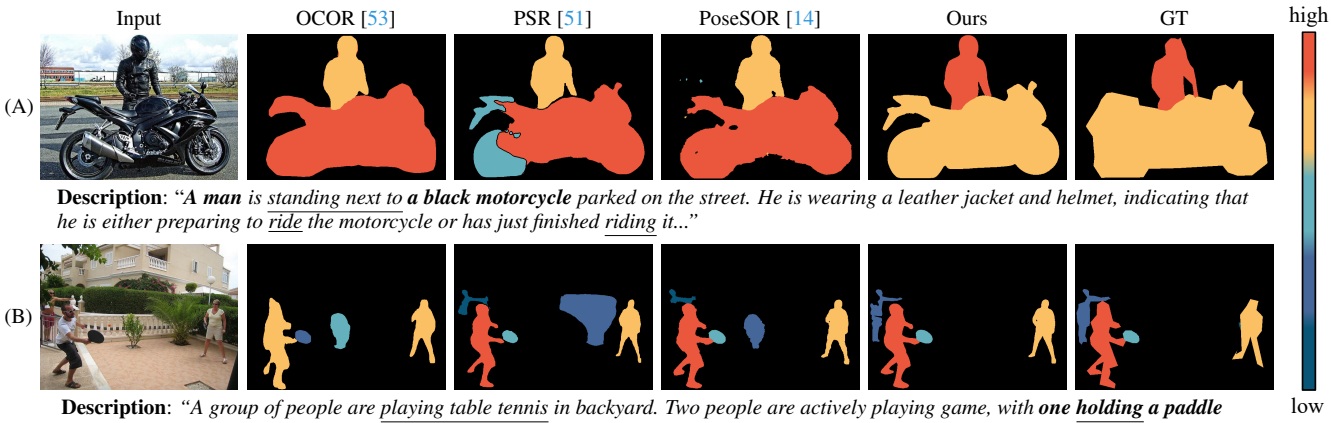

Figure 1. Existing methods [14, 51, 53] may predict unrealistic saliency ranks when they fail to understand the semantic relations among different salient objects. We observe that LVLMs (*e.g.*, [5]) tend to focus on salient objects in an image, and their generated descriptions contain semantic relations of these objects. (In each description, texts related to the salient objects are marked in **bold** while those related to object relations are underlined.) We therefore propose a new method, LG-SOR, to explicitly exploit such cues in the language descriptions from LVLMs to enhance saliency ranking.

in the text description with the visual features, which helps filter noisy background information and allows useful information to propagate further. Second, we propose a novel Text-Aware Visual Reasoning (TAVR) module to harness the implicit order within the entity cues and enhance ranking order reasoning within relation cues from the language description. It explicitly parses the description into entity cues (*e.g.*, "*a man, a black motorcycle*" in Fig. 1(A)) and relation cues (*e.g.*, "*standing next to, ride*"), and learns to construct a multimodal graph. As shown in Fig. 1, our approach can understand the semantic relations between different objects according to the input text descriptions, predicting correct rank orders and their masks.

Our main contributions can be summarized as follows:

- We propose the first language-guided saliency ranking approach (LG-SOR), which leverages language descriptions to guide salient object ranking.
- We propose a novel TGVM module to exploit the semantic information in the description to learn global semantic context information, and a novel TAVR module to enhance the reasoning of saliency ranks by explicitly modeling the extracted entity and relation cues from the description.
- We conduct extensive experiments to analyze our approach and show that it outperforms state-of-the-art methods.

2. Related Work

Salient Object Ranking (SOR). Islam *et al.* [22] propose to combine the binary saliency maps from multiple observers to infer a saliency consensus map for each input image. These saliency consensus maps are then considered as ground truth for model training. Later, Siris *et al.* [48, 49] propose and formulate the salient object ranking task as to determine the order of focusing and then shifting the attention across different objects by an observer, following the human attention process [40]. Subsequently, a number of methods are proposed to improve the SOR performance via designing different learning techniques (*e.g.*, neural graphs [6, 36, 44, 62] and transformers [50, 51, 53]), and modeling different contextual cues (*e.g.*, object positions [10], object-based and spatial attention [53], relations between object groups [51], and human poses [14]). More recently, Guan and Lau [13] propose to model human foveal and peripheral visions, by predicting salient objects one by one in a sequential manner to form the saliency rank.

Unlike the above existing works, which either overlook or only implicitly model low-level spatial and high-level semantic relations among objects derived from image features, our method explicitly captures these relations and the implicit orders derived from the language description, which can be readily obtained from LVLMs.

Salient Object Detection (SOD) aims to detect the visually distinctive object(s) in an image. Recent SOD methods are deep learning based. Some of them focus on fusing or enhancing features of different levels via, *e.g.*, dynamic convolution [41], recurrent blocks [43], attention mechanisms [38, 45, 71], and transformer-based architectures [34, 42]. Other methods use auxiliary tasks/modalities, *e.g.*, image captioning [69], depth [72], light fields [29], and edge detection [70], to facilitate saliency detection. There is also a group of methods [9, 26, 52, 54, 55, 68] proposed to detect salient objects at the instance level.

Despite their success, these methods only detect salient objects in an image, and do not consider ranking the saliency order of the detected objects.

Language-Guided Segmentation (LGS) aims to segment the target object(s) in an image according to the input description. A key difference among existing LGS methods is that they use different strategies to fuse linguistic and visual features, *e.g.*, feature concatenation [18], graph-based blocks [17, 20, 21], attention mechanisms [61], and transformers-based feature fusion in the encoder [63], decoder [8, 33, 60, 65, 74] or both sides [64].

LGS is different from our language-guided SOR in two-fold. First, LGS descriptions describe the target objects, serving to differentiate them from other objects. They usually focus on the attributes, such as colors, positions, and appearances, and may sometimes contain low-level spatial (*e.g.*, next to and behind) relations to other objects. In contrast, our descriptions describe the entire scene. They tend to describe salient objects with their implicit order, and focus on the spatial and semantic relations among them. Second, while LGS methods prioritize cross-modal matching between discriminative textual and visual features of the target object for the segmentation, our LG-SOR approach focuses on extracting the semantic relations and the implicit orders from the descriptions for saliency ranking.

Large Vision-Language Models (LVLMs) [2, 75] are recently proposed to provide visual comprehension and reasoning by borrowing the strong reasoning capability of Large Language Models (LLMs) [4, 27, 56, 66]. LVLMs have achieved impressive success in several vision tasks, *e.g.*, image captioning [5, 35], question answering [30, 66, 67], reasoning segmentation [25, 73], object detection [57, 58], and few-shot learning [76].

In this work, we explore how to leverage LVLMs for saliency ranking, by extracting useful knowledge in the language descriptions generated by LVLMs to guide the SOR model to understand the relations of different objects.

3. Proposed Method

We observe that the way large vision-language models (LVLMs) [5, 19, 35] describe images coincides with human scene perception. LVLMs, like humans, prioritize and describe the most salient objects first, and then shift the focus to the less salient objects, while also incorporating the relationships between these objects. Inspired by this observation, we propose in this paper the Language-Guided Salient Object Ranking (LG-SOR) method.

Fig. 2 illustrates the LG-SOR framework. Given an input image $\mathbf{I}$, we first employ an LVLM (*e.g.*, [5]) to generate a detailed textual description $\mathbf{T} \in \mathbb{R}^{N_w}$ comprising N_w words, prompted by a predefined instruction such as "*Write a detailed description for the image*". We then extract multi-scale visual features $\{\mathbf{V}_i \in \mathbb{R}^{H_i \times W_i \times C_i}\}_{i=1}^{4}$ using a visual encoder and text features (including word features $\mathbf{L}_w \in \mathbb{R}^{N_w \times C_l}$ and sentence features $\mathbf{L}_s \in \mathbb{R}^{1 \times C_l}$) using a text encoder, where C_i, H_i, W_i denote the number of channels, height, and width of the visual features in the i-th scale and C_l denotes the number of channels of the text features. To selectively incorporate language cues into visual features, we propose the Text-Guided Visual Modulation (TGVM) module to enhance the visual features $\mathbf{V}_i$ with both word-level $\mathbf{L}_w$ and sentence-level $\mathbf{L}_s$ textual features through a modulation operation. This produces semantics-enhanced visual features $\mathbf{F}_m$, which are then decoded via a transformer decoder to produce semantics-enhanced salient object embeddings $\mathbf{O}_s$. We then propose the Text-Aware Visual Reasoning (TAVR) module to construct a multi-modal graph for robust object reasoning. Taking the entity cues $\mathbf{L}_e$ and relation cues $\mathbf{L}_r$ (derived from the language description), and the learned semantics-enhanced salient object embeddings $\mathbf{O}_s$ as input, the TAVR module predicts order-aware object embeddings $\mathbf{O}_g$. Finally, the first-scale visual features $\mathbf{V}_1$ and the semantics-enhanced salient object embeddings $\mathbf{O}_s$ are exploited for saliency mask prediction through the mask head [3]. The semantics-enhanced visual features $\mathbf{F}_m$ and order-aware salient object embeddings $\mathbf{O}_g$ are used to predict the saliency ranks via the rank decoder. The rank decoder processes the input $\mathbf{O}_g$ and $\mathbf{F}_m$ through N_d standard transformer decoder layers and a linear layer to produce the rank score for each object. Finally, we combine the rank scores with the salient instance masks (predicted by the mask head) to produce the output ranking result.

3.1. Text-Guided Visual Modulation (TGVM)

While language descriptions can provide rich semantic information, they can also introduce noisy information (*e.g.*, descriptions of non-salient background objects) to the saliency ranking process. To address this problem, we propose a novel Text-Guided Visual Modulation (TGVM) module to learn semantics-enhanced salient object representations conditioned on both visual and text features. TGVM learns to selectively incorporate useful textual information while suppressing background noise. It first utilizes the modulation operation (as shown in Fig. 2) to produce semantics-enhanced visual features $\mathbf{F}_i$, conditioned on visual features ($\mathbf{V}_i$) and text features ($\mathbf{L}_w$, and $\mathbf{L}_s$), and then a transformer decoder is adapted to generate the output semantics-enhanced salient object embeddings $\mathbf{O}_s$.

Modulation Operation. To selectively integrate valuable textual information, we learn to scale and shift the attendance of the text features by considering contextual information. Specifically, we process visual features $\mathbf{V}_i$ and word features $\mathbf{L}_w$ by adding position embeddings and projecting them to a unified dimension. We then use cross-attention to obtain a text-attended visual representation $\mathbf{M}_i$. However, this representation may inherit noise from word-level features. Hence, we incorporate global semantics from sentence features $\mathbf{L}_s$ to adjust the attendance of potential noisy features. We employ a regression layer [23] to learn the scaling and shifting

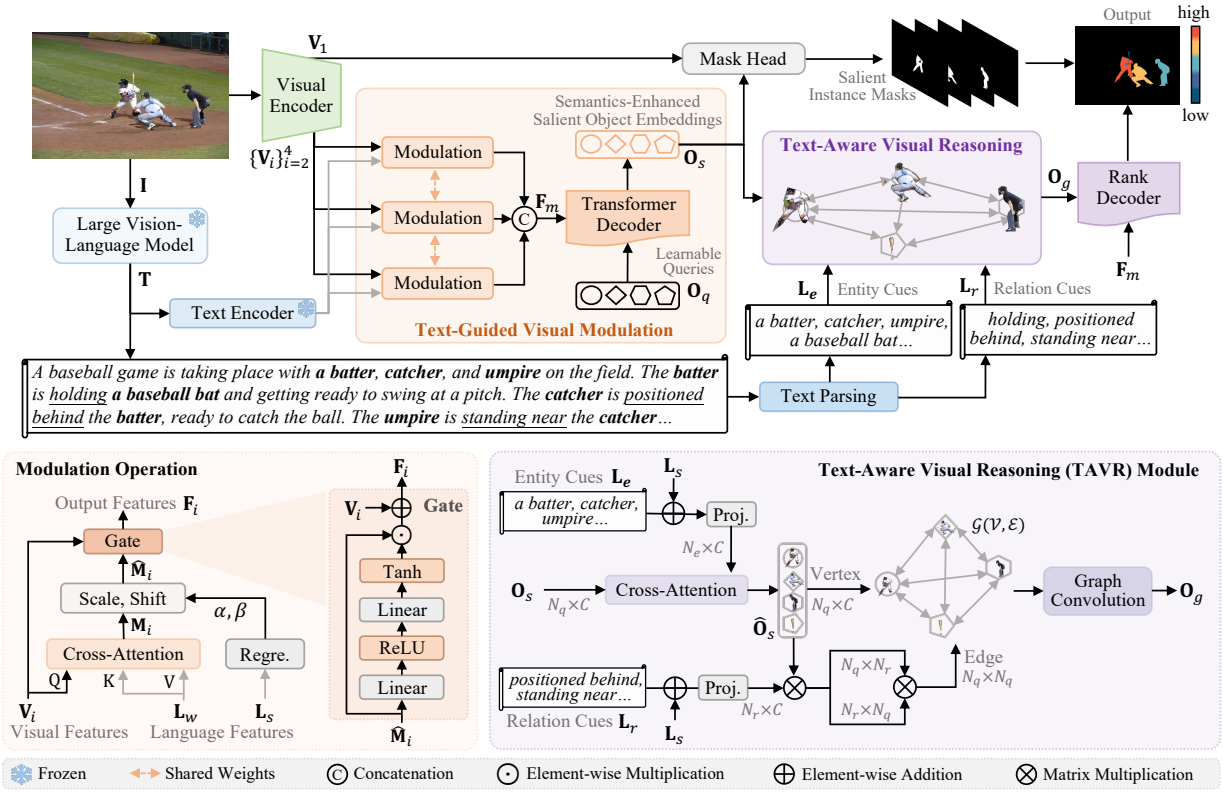

Figure 2. Overview of our LG-SOR approach. Given an input image $\mathbf{I}$ and its corresponding description $\mathbf{T}$ generated by the LVLM, we first apply a visual encoder and a text encoder to extract visual and text features. We then propose a novel Text-Guided Visual Modulation (TGVM) module to selectively incorporate textual features as guidance to learn semantics-enhanced salient object embeddings $\mathbf{O}_s$ for predicting the saliency instance masks. We further propose a novel Text-Aware Visual Reasoning (TAVR) module to leverage the parsed entity and relation cues from the description to learn order-aware object embeddings $\mathbf{O}_g$ based on $\mathbf{O}_s$ for predicting the saliency ranks.

parameters from $\mathbf{L}_s$, which are then used to modulate the text-attended features $\mathbf{M}_i$ to produce the enhanced visual features $\hat{\mathbf{M}}_i$, as:

$$\hat{\mathbf{M}}_i = \mathbf{M}_i \odot (1+\alpha) + \beta, \quad \text{with } \alpha, \beta = \phi_{Regre}(\mathbf{L}_s), \tag{1}$$

where $\alpha \in \mathbb{R}^{1\times C}$ and $\beta \in \mathbb{R}^{1\times C}$ are scaling and shifting parameters, respectively. ϕ_{Regre} represents a regression layer and C denotes the number of channels.

A gating block is additionally utilized to learn to suppress the influence of noisy textual information. Specifically, we feed the enhanced visual features $\hat{\mathbf{M}}_i$ into a gating block to generate the language-aware visual features $\mathbf{F}_i \in \mathbb{R}^{H_i\times W_i\times C}$, as:

$$\mathbf{F}_i = \phi_{Gate}(\hat{\mathbf{M}}_i) \odot \hat{\mathbf{M}}_i \oplus \mathbf{V}_i, \tag{2}$$

where ϕ_{Gate} represents a two-layer MLP that includes a linear layer with a ReLU activation followed by a linear layer with a Tanh activation. With the guidance of both local (word-level) and global (sentence-level) textual features, the TGVM module selectively integrates linguistic information into the visual features to enhance useful semantic contexts while filtering out noisy background information.

After obtaining language-aware visual features $\mathbf{F}_i$ at different scales, we flatten their spatial dimensions and concatenate them together to form the multi-scale language-aware visual features $\mathbf{F}_m \in \mathbb{R}^{N_m\times C}$ (where $N_m = \sum_{i=2}^{4} H_i \times W_i$). We then use a transformer decoder [3] to process $\mathbf{F}_m$ and learnable queries $\mathbf{O}_q \in \mathbb{R}^{N_q\times C}$, where N_q is the number of queries, to produce the semantics-enhanced salient object embeddings $\mathbf{O}_s \in \mathbb{R}^{N_q\times C}$ as the output of TGVM.

3.2. Text-Aware Visual Reasoning (TAVR)

With the semantics-enhanced salient object embeddings $\mathbf{O}_s$, we now aim to determine the ranks of these salient objects via joint modeling of visual stimuli and the implicit order and relations within the textual description. To better exploit the implicit object orders and to prevent the extraction of vague relation features from sentence embeddings, we first parse the description to identify words and phrases that indicate objects and relations. These are exploited by a novel Text-Aware Visual Reasoning (TAVR) module to learn order-aware salient object embeddings $\mathbf{O}_g$.

Text Parsing. We employ spaCy [1] to parse the LVLM-generated description $\mathbf{T}$, allowing us to identify the words and phrases that represent objects and relationships. These extracted elements are then fed into the text encoder to generate the entity cues $\mathbf{L}_e \in \mathbb{R}^{N_e \times C_l}$ and relation cues $\mathbf{L}_r \in \mathbb{R}^{N_r \times C_l}$. N_e and N_r denote the numbers of entities and relations.

Visual Reasoning. The TAVR module aims to construct a multimodal visual-linguistic graph for reasoning the salient object ranks, based on the extracted entity cues $\mathbf{L}_e$ and relation cues $\mathbf{L}_r$. As shown in Fig. 2, the TAVR module takes the text-enhanced salient object embeddings $\mathbf{O}_s$ and text features, including entity cues $\mathbf{L}_e$, relation cues $\mathbf{L}_r$, and sentence features $\mathbf{L}_s$, as inputs to produce the order-aware salient object embeddings $\mathbf{O}_g$.

Specifically, we first add the sentence features $\mathbf{L}_s$ to both the entity cues $\mathbf{L}_e$ and relation cues $\mathbf{L}_r$ to enhance contextual understanding. Next, we ensure their dimensions align with the desired feature space C using a projection layer. To capture the interplay between the text-enhanced salient object embeddings $\mathbf{O}_s$ and the entity features $\mathbf{L}_e$, we employ a cross-attention mechanism. Here, $\mathbf{O}_s$ are used as a query and $\mathbf{L}_e$ as both key and value. This strategy effectively encodes the implicit orders within the input description, resulting in entity-aligned salient object embeddings $\hat{\mathbf{O}}_s \in \mathbb{R}^{N_q \times C}$.

We then explicitly construct a fully-connected multimodal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \{\vartheta\}_{i=1}^{N_q}$ denotes N_q vertexes, and $\mathcal{E}$ contains $N_q \times N_q$ edges. We use the entity-aligned embeddings $\hat{\mathbf{O}}_s = \{o_i\}_{i=1}^{N_q} \in \mathbb{R}^{N_q \times C}$ to denote the vertex features and $\mathbf{A} \in \mathbb{R}^{N_q \times N_q}$ as the edge adjacency matrix. Unlike previous methods [6, 36, 62] that infer the implicit relationships only from image features, we also incorporate explicit relation cues $\mathbf{L}_r$ derived from the language description to compute the adjacency matrix $\mathbf{A}$, as:

$$\mathbf{A} = \mathbf{R}_1 \otimes \mathbf{R}_2 = \mathcal{S}(\mathbf{R}) \otimes \mathcal{S}(\mathbf{R}^\top), \quad \text{with } \mathbf{R} = \hat{\mathbf{O}}_s \otimes \mathbf{L}_r^\top, \tag{3}$$

where $\mathcal{S}$ is the softmax operation. $\mathbf{R} \in \mathbb{R}^{N_q \times N_r}$ is the affinity matrix between $\hat{\mathbf{O}}_s$ and $\mathbf{L}_r$. We apply the $\mathrm{softmax}$ function along both the first and second dimensions of $\mathbf{R}$ to derive $\mathbf{R}_1 \in \mathbb{R}^{N_q \times N_r}$ and $\mathbf{R}_2 \in \mathbb{R}^{N_r \times N_q}$, respectively. The adjacency matrix $\mathbf{A}$ is then computed through matrix multiplication of $\mathbf{R}_1$ and $\mathbf{R}_2$. We apply graph convolution [24, 28] to the established visual-linguistic multimodal graph $\mathcal{G}$ to produce the order-aware salient object embeddings $\mathbf{O}_g \in \mathbb{R}^{N_q \times C}$ as the output of TAVR:

$$\mathbf{O}_g = \sigma\left(\mathbf{A}(\mathbf{I}_s + \hat{\mathbf{O}}_s) \otimes \mathbf{W}\right), \tag{4}$$

where σ is the ReLU function, $\mathbf{I}_s$ is an identity mapping, and $\mathbf{W} \in \mathbb{R}^{C \times C}$ is a learnable parameter matrix that facilitates the adaptation/refinement of node features. By aggregating information from neighboring nodes in the graph, graph convolution empowers the model to reason the saliency degree of an object. This reasoning process considers not only the object's intrinsic visual features and semantics captured by $\hat{\mathbf{O}}_s$, but also its relations with other objects as encoded in the graph structure.

4. Experiments

4.1. Experimental Setups

Implementation Details. We employ ResNet-50 [16] and Swin-L [39], pretrained on the MS-COCO training set [31] as our image encoder, following previous SOR methods [6, 10, 36, 44, 51, 62]. We utilize BERT [7] as the text encoder to extract text features. The input descriptions are capped at a maximum length of 256 characters for all experiments, and the input images are resized to 1024×1024 following [6]. We employ binary cross-entropy loss and dice loss [32] for mask prediction and ranking loss [36] for ranking prediction. We train our LG-SOR 40,000 iterations with a batch size of 16 on eight A100 GPUs (80GB). The learning rate is initially set to $1e^{-5}$ and then reduced by 10 after 30,000 iterations. We use the AdamW optimizer with a 0.05 weight decay for model optimization. The number of LVLM output words (N_w), entities (N_e), relations (N_r), learnable queries (N_q), and transformer decoder layers in the rank decoder (N_d) are set to 256, 34, 24, 200, and 3, respectively.

Datasets, Methods, and Metrics. We conduct experiments on two standard SOR datasets, the ASSR [48] and IRSR [36]. The ASSR dataset contains 7,464 training images, 1,436 validation images, and 2,418 test images, and each image has five salient object ranks. The IRSR dataset contains 6,059 images for training and 2,929 images for testing, and each image has one to eight salient object ranks.

We compare our method with **seventeen** state-of-the-art methods, including eleven existing SOR methods (*i.e.*, RSDNet [22], ASSR [48], IRSR [36], PPA [10], PSR [51], OCOR [53], SeqRank [13], HyperSOR [44], QAGNet [6], DSGNN [62], PoseSOR [14]), one salient instance detection method (S4Net [9]), one salient object detection method (VST [37]), two instance segmentation methods (QueryInst [11], Mask2Former [3]), one large-vision language model-based method (GiT [57]), and one language-guided segmentation method (X-Decoder [77], whose input description is generated by [5]). For a fair comparison, we retrain all existing SOR methods on both ASSR and IRSR datasets based on their released codes (if available)[1], following previous SOR methods [13, 14, 44, 51]. For the competing methods from other related tasks, we modify and re-train them based on their released codes. Refer to Supp. A for more implementation details.

We use three metrics for performance evaluation, *i.e.*, Salient Object Ranking (SOR) [48, 49], Segmentation-

[1]For HyperSOR [44], we directly copy and report the results from their paper as their code is not available.

Table 1. Quantitative comparison. '-' denotes that the result is not available. SID: Salient Instance Detection. SOD: Salient Object Detection. IS: Instance Segmentation. LGS: Language-Guided Segmentation. LVLM: Large Vision Language Model. SOR: Salient Object Ranking. The best performance is marked in **bold**.

Method	Reference	Original Task	Backbone	ASSR Test Set [48]			IRSR Test Set [36]		
				SA-SOR↑	SOR↑	MAE↓	SA-SOR↑	SOR↑	MAE↓
S4Net [9]	[CVPR'19]	SID	ResNet-50	0.469	0.662	0.149	0.307	0.648	0.129
VST [37]	[ICCV'21]	SOD	T2T-ViT-T	0.434	0.649	0.104	0.254	0.584	0.094
QueryInst [11]	[ICCV'21]	IS	ResNet-101	0.614	0.843	0.098	0.502	0.807	0.082
Mask2Former [3]	[CVPR'22]	IS	ResNet-101	0.625	0.857	0.078	0.518	0.814	0.079
GiT [57]	[ECCV'24]	LVLM	GiT-B	0.541	0.854	0.101	0.371	0.817	0.122
X-Decoder [77]	[CVPR'23]	LGS	Focal-T	0.609	0.851	0.075	0.545	0.811	0.086
RSDNet [22]	[CVPR'18]	SOR	ResNet-101	0.499	0.717	0.158	0.471	0.729	0.112
ASSR [48]	[CVPR'20]	SOR	ResNet-101	0.637	0.815	0.105	0.350	0.702	0.109
IRSR [36]	[TPAMI'21]	SOR	ResNet-50	0.643	0.841	0.108	0.545	0.808	0.083
PPA [10]	[ICCV'21]	SOR	VoVNet-39	0.647	0.859	0.082	0.501	0.783	0.081
PSR [51]	[ACMMM'23]	SOR	ResNet-50	0.651	0.849	0.079	0.528	0.819	0.083
HyperSOR [44]	[TPAMI'24]	SOR	ResNet-101	0.653	0.830	0.101	-	-	-
Ours	-	SOR	ResNet-50	0.733	0.882	0.065	0.578	0.817	0.060
OCOR [53]	[CVPR'22]	SOR	Swin-L	0.594	0.875	0.101	0.482	0.813	0.079
SeqRank [13]	[AAAI'24]	SOR	Swin-L	0.661	0.865	0.081	0.553	0.821	0.075
QAGNet [6]	[CVPR'24]	SOR	Swin-L	0.772	0.867	0.052	0.618	0.825	**0.049**
DSGNN [62]	[CVPR'24]	SOR	Swin-L	0.765	0.860	0.051	0.609	0.812	0.058
PoseSOR [14]	[ECCV'24]	SOR	Swin-L	0.664	0.854	0.077	0.547	0.816	0.070
Ours	-	SOR	Swin-L	**0.787**	**0.895**	**0.049**	**0.634**	**0.835**	0.050

Aware SOR (SA-SOR) [36], and Mean Absolute Error (MAE). SOR computes the Spearman's rank-order correlation between the predicted and actual saliency ranking orders, emphasizing the relative saliency among objects rather than assigning specific ranks to individual objects. SA-SOR computes the Pearson correlation between the predicted and true saliency ranks, penalizing the misidentification of non-salient objects and incorrect rankings. MAE quantifies the average per-pixel difference between predicted and Ground-Truth saliency maps.

4.2. Main Results

Quantitative Comparison. Table 1 presents quantitative results. Our approach achieves state-of-the-art results across all metrics on both benchmarks. Notably, our SA-SOR and SOR scores surpass the latest methods PoseSOR [14] and DSGNN [62] by 17.47% and 4.19%, respectively, on the ASSR benchmark. This is because our method can effectively mine rich semantic information from the language description for the ranking process. In contrast, PoseSOR and DSGNN either neglect or only implicitly learn from just the vision features. In addition, compared to the multi-modal method (X-Decoder [77]), which also utilizes the same language description as input, our method shows impressive gains, improving the SA-SOR scores by 28.08% on ASSR dataset and by 16.70% on IRSR dataset. This is because the X-Decoder is unable to effectively extract and reason about the rich semantic information within the textual descriptions. In contrast, our approach leverages language-guided visual modulation and reasoning to exploit valuable information (*e.g.*, implicit orders and relations) in descriptions, while effectively suppressing background noise. Compared to GiT [57], which directly employs LVLM to predict instance masks and ranks without considering the modality discrepancy, our method exploits LVLM-generated descriptions as ancillary guidance to enhance the extraction of useful semantic information for salient object ranking, thus bringing in a significant performance improvement (*e.g.*, a 44.18% boost in the SA-SOR metric on the ASSR dataset). Refer to Supp. B for a computational analysis.

Qualitative Comparison. Fig. 3 shows the qualitative comparison of our method with nine best-performing methods chosen from Table 1. We can see that the SOR maps generated by our approach consistently outperform all compared methods across different types of scenes. Specifically, the examples in the first three rows show that our approach can effectively rank salient objects based on the implicit order conveyed by the description. For example, in the first row, our method effectively uses the perceived semantic information, such as the attribute (*"red"*) and implicit order (first *"girl"* and then *"elephant"*), to correctly identify *"the young girl"* as the most salient object, before shifting the focus to the less salient but relatively large *"elephant"*. In contrast, most other approaches either erroneously assign equal saliency to both the woman and the elephant [57, 77] or incorrectly highlight the elephant as the most salient object [13, 14, 51, 53].

Further, our method demonstrates robustness in capturing saliency, even when the order in the description does not exactly match with the ground truth, as validated in the fourth to sixth rows of Fig. 3, where objects' visual intrinsic positional context and relations in the description are effectively

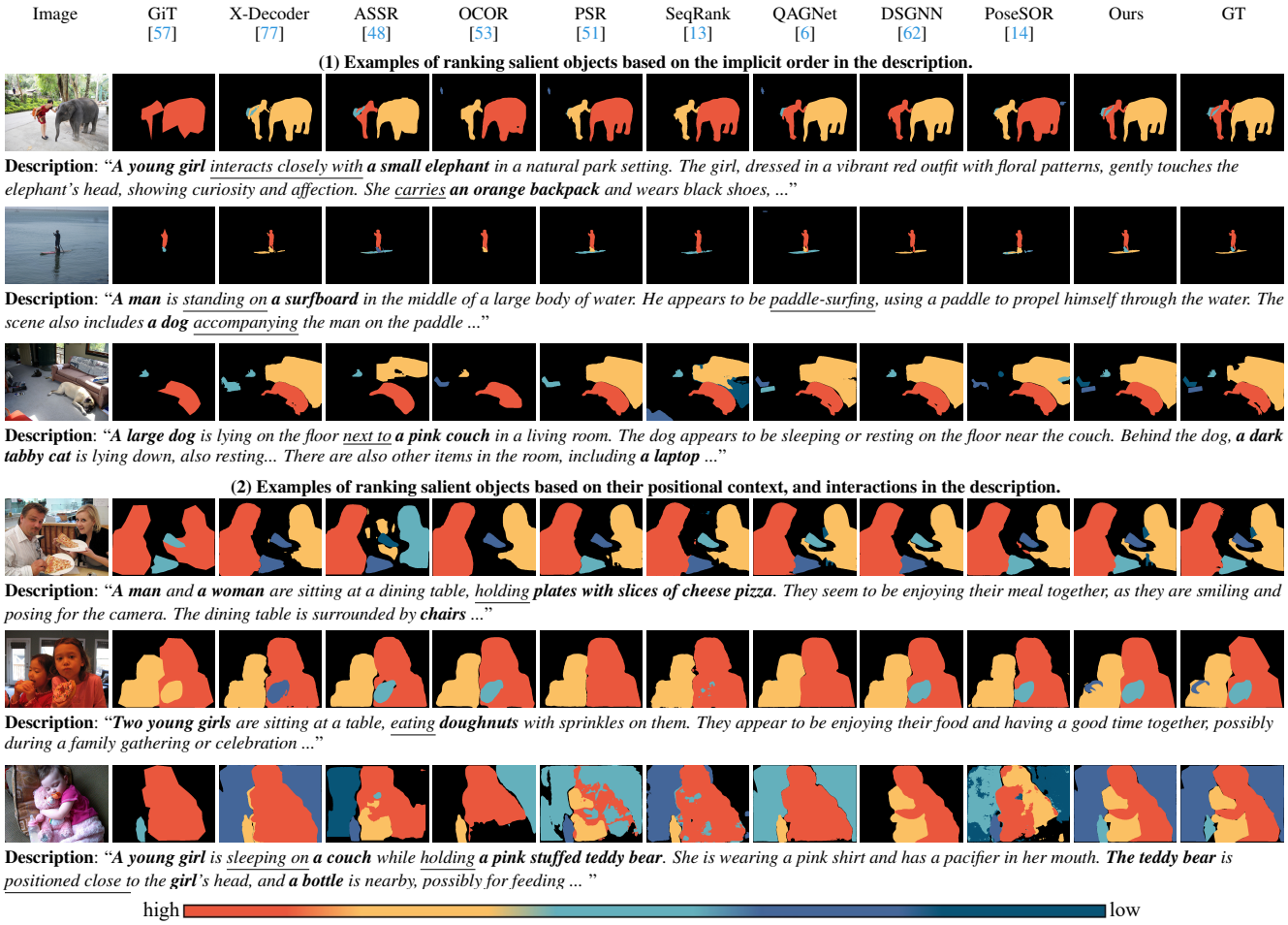

Figure 3. Qualitative comparison of our method with nine best-performing methods in Table 1.

leveraged to determine saliency. For example, in the sixth row, although "*couch*" is mentioned before "*pink teddy bear*" in the description, our method can still rank it as the least salient. This is attributed to the graph reasoning process, which exploits the objects' intrinsic positional context (*e.g.*, the couch is in the background, while the pink teddy bear is more prominently positioned in the foreground) in graph nodes and relational cues with other objects (*e.g.*, "*sleeping on*") in edges. Refer to Supp. C and Supp. E for more discussion and visual comparisons, respectively.

4.3. Ablation Study

Component Analysis. We conduct the ablation study based on the ASSR dataset [48] to verify the effectiveness of the proposed modules in Table 2. We build the baseline **(I)** by removing TGVM and TAVR modules, where learnable queries are fed into the rank decoder and mask head for ranking predictions. We then demonstrate the effectiveness of the proposed TGVM module by gradually adding the cross-attention block for word features $\mathbf{L}_w$ **(II)**, regression block for sentence features $\mathbf{L}_s$ **(III)**, and gate block **(IV)**. The results show that both the word-level and sentence-level text features are beneficial. They can facilitate the model in learning semantic information from the language description. In addition, a clear performance improvement can be seen after adding the gate block, validating its effectiveness.

We further study three alternative designs for learning complex relations in the language description: **(V)** we use semantics-enhanced salient object embeddings $\mathbf{O}_s$ obtained from TGVM as the vertex features, and sentence features $\mathbf{L}_s$ as edge features to construct the graph; **(VI)** we use entity-aligned salient object embeddings $\hat{\mathbf{O}}_s$ as vertex features, and global sentence features $\mathbf{L}_s$ to construct the edge adjacent matrix; and **(VII)** we use $\mathbf{O}_s$ as graph vertex features, and relation cues $\mathbf{L}_r$ as edge features. Setting **VIII** uses entity-aligned salient object embeddings $\hat{\mathbf{O}}_s$ as vertex features, and relation cues $\mathbf{L}_r$ as edge features. The results show that adding the parsed entity or relation cues from the multi-modal graph helps improve the performance. This confirms that these cues enhance the model's ability to comprehend complex relationships between different instances. As a result, our model can predict the importance of each instance

Table 2. Ablation analysis of different modules in LG-SOR. $\mathbf{L}_w$ and $\mathbf{L}_s$ denote word and sentence features. $\mathbf{L}_e$ and $\mathbf{L}_r$ are entity and relation cues.

Settings	TGVM			TAVR			SA-SOR ↑	SOR ↑	MAE ↓
	$\mathbf{L}_w$	$\mathbf{L}_s$	Gate	Graph	$\mathbf{L}_e$	$\mathbf{L}_r$			
I (Baseline)							0.687	0.852	0.073
II	✓						0.704	0.863	0.071
III	✓	✓					0.706	0.866	0.071
IV	✓	✓	✓				0.713	0.870	0.069
V	✓	✓	✓	✓			0.721	0.874	0.067
VI	✓	✓	✓	✓	✓		0.726	0.877	0.066
VII	✓	✓	✓	✓		✓	0.727	0.879	0.065
VIII	✓	✓	✓	✓	✓	✓	**0.733**	**0.882**	**0.065**

not only based on its inherent saliency but also through its relations and interactions with other instances in the scene.

Order of the Parsed Entities. We ablate the impact of entity ordering in Table 3. For the parsed entities, we first shuffle their orders and obtain the entities in a random order. Compared to the original entities with the normal order, which contains the implicit orders of the salient objects, the random order causes lower SA-SOR and SOR scores. The results also validate that the implicit ranking order in the description can help boost the saliency ranking performance.

Ablation Study of Different LVLMs. Considering that different LVLMs may generate captions with different levels of detail and accuracy, we first investigate the effects of various detailed descriptions produced by distinct LVLMs [2, 5, 19, 35] in rows 1 to 4 of Table 4. Although the results slightly differ across sources, the enriched semantic information from input text generally facilitates the model in reasoning about the importance of different objects. This performance variability further demonstrates the robustness of our model, highlighting its ability to extract useful information from text inputs with diverse details. In addition, the process of generating textual descriptions through these well-established and accessible LVLMs is automatic and scalable, ensuring the usability and scalability of our model. Although GPT-4V [2] achieves higher SA-SOR and SOR scores due to more accurate semantic details, its significance is restricted by its limited availability and high cost, making it less accessible to the general public. Thus, we take InstructBLIP [5] as our final choice.

Further, we prompt InstructBLIP to generate a short description using the instruction *"Write a short description for the image"*. The short descriptions are then used as input to train the model, allowing us to compare its performance with the model trained on long text descriptions, as shown in Table 4. The results reveal an obvious performance drop with short descriptions, which is primarily attributed to their lack of richer semantic content and contextual information, such as object attributes and interactions, in the descriptions. Refer to Supp. D for more comparisons.

Table 3. Ablation study of the order of entities.

Orders	SA-SOR ↑	SOR ↑	MAE ↓
Random	0.726	0.876	0.066
Normal	**0.733**	**0.882**	**0.065**

Table 4. Ablation study on caption sources. † denotes that the generated descriptions are short.

Sources	SA-SOR ↑	SOR ↑	MAE ↓
LLaVa [35]	0.723	0.881	0.065
OPERA [19]	0.725	0.882	0.066
GPT-4V [2]	**0.741**	**0.886**	**0.064**
InstructBLIP [5]	0.733	0.882	0.065
InstructBLIP [5]†	0.727	0.880	0.065

Description: "***A group of young men*** *are* *riding* ***skateboards*** *on a sidewalk. Some of them are wearing helmets, and ...* "

Figure 4. A failure case. Our method may fail to precisely rank salient objects when objects share a similar positional context, and the language description lacks semantic relationships.

5. Conclusion

In this paper, we have proposed the Language-Guided Salient Object Ranking (LG-SOR) approach, which harnesses the knowledge within the LVLM-generated descriptions to enhance object ranking by integrating semantic and order cues from the descriptions. Our approach contains two novel modules: Text-Guided Visual Modulation (TGVM) and Text-Aware Visual Reasoning (TAVR). The TGVM module effectively integrates semantic information from textual descriptions with visual features, filtering out background noise while propagating useful textual cues. Meanwhile, the TAVR module exploits the implicit order and improves the ranking reasoning by using parsed entity and relation cues in the language descriptions to construct a multimodal graph. Extensive experiments have demonstrated the superior performance of LG-SOR on two SOR benchmarks, validating its effectiveness.

Our approach does have limitations. For example, when salient objects share very similar semantic features (*e.g.*, positions and attributes) and the language description lacks sufficient semantic relationships, our model struggles to precisely infer their ranks, as shown in Fig. 4. As a future work, we would like to explore the incorporation of other modalities to assist the model in perceiving view and distance in a scene, thereby enabling more accurate ranking order prediction.

References

[1] spacy. https://spacy.io/. 5

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 3, 8

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 4, 5, 6

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *JMLR*, 2023. 3

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 5, 8

[6] Bowen Deng, Siyang Song, Andrew P French, Denis Schluppeck, and Michael P Pound. Advancing saliency ranking with human fixations: Dataset models and benchmarks. In *CVPR*, 2024. 2, 5, 6, 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. 5

[8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 3

[9] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019. 2, 5, 6

[10] 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 *ICCV*, 2021. 2, 5, 6

[11] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 5, 6

[12] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *WACV*, 2024. 1

[13] Huankang Guan and Rynson W.H. Lau. Seqrank: Sequential ranking of salient objects. In *AAAI*, 2024. 2, 5, 6, 7

[14] Huankang Guan and Rynson WH Lau. Posesor: Human pose can guide our attention. In *ECCV*, 2024. 1, 2, 5, 6, 7

[15] Yingchun Guo, Meng Zhang, Xiaoke Hao, and Gang Yan. Irnet-rs: image retargeting network via relative saliency. *Neural Computing and Applications*, 2024. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[17] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, pages 13332–13341, 2024. 3

[18] Ronghang Hu, Marcus Rohrbach, Trevor Darrell, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 3

[19] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024. 1, 3, 8

[20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020. 3

[21] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020. 3

[22] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, 2018. 2, 5, 6

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 5

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 3

[26] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[28] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, 2018. 5

[29] Zijian Liang, Pengjie Wang, Ke Xu, Pingping Zhang, and Rynson WH Lau. Weakly-supervised salient object detection on light fields. *IEEE TIP*, 2022. 2

[30] Jiaying Lin, Shuquan Ye, and Rynson W.H. Lau. Do multimodal large language models see like humans? In *arXiv*, 2024. 3

[31] 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 *ECCV*, 2014. 5

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[33] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *ICCV*, pages 22124–22134, 2023. 3

[34] Fang Liu, Yuhao Liu, Jiaying Lin, Ke Xu, and Rynson WH Lau. Multi-view dynamic reflection prior for video glass surface detection. In *AAAI*, pages 3594–3602, 2024. 2

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 3, 8

[36] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE TPAMI*, 2021. 2, 5, 6

[37] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021. 5, 6

[38] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *ICCV*, pages 7555–7564, 2021. 2

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[40] Ulric Neisser. *Cognitive psychology: Classic edition*. Psychology press, 2014. 2

[41] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, 2020. 2

[42] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Transcmd: Cross-modal decoder equipped with transformer for rgb-d salient object detection. *arXiv:2112.02363*, 2021. 2

[43] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, 2019. 2

[44] 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 TPAMI*, 2024. 2, 5, 6

[45] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, pages 13676–13685, 2020. 2

[46] Yu-Kun Qiu, Fa-Ting Hong, Wei-Hong Li, and Wei-Shi Zheng. Learning relation models to detect important people in still images. *IEEE TMM*, 2022. 1

[47] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *WACV*, 2024. 1

[48] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Inferring attention shift ranks of objects for image saliency. In *CVPR*, 2020. 1, 2, 5, 6, 7

[49] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Inferring attention shifts for salient instance ranking. *IJCV*, 2023. 2, 5

[50] Mengke Song, Linfeng Li, Dunquan Wu, Wenfeng Song, and Chenglizhao Chen. Rethinking object saliency ranking: A novel whole-flow processing paradigm. *IEEE TIP*, 2023. 2

[51] Chengxiao Sun, Yan Xu, Jialun Pei, Haopeng Fang, and He Tang. Partitioned saliency ranking with dense pyramid transformers. In *ACM MM*, 2023. 1, 2, 5, 6, 7

[52] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection. In *BMVC*, 2020. 2

[53] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *CVPR*, 2022. 1, 2, 5, 6, 7

[54] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 2022. 2

[55] Xin Tian, Ke Xu, and Rynson WH Lau. Unsupervised salient instance detection. In *CVPR*, 2024. 2

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 3

[57] Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. Git: Towards generalist vision transformer through universal language interface. In *ECCV*, 2024. 3, 5, 6, 7

[58] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2023. 3

[59] Xiao Wang, Zheng Wang, Toshihiko Yamasaki, and Wenjun Zeng. Very important person localization in unconstrained conditions: A new benchmark. In *AAAI*, 2021. 1

[60] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 3

[61] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, 2020. 3

[62] 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 *CVPR*, 2024. 2, 5, 6, 7

[63] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *ICCV*, 2023. 3

[64] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 3

[65] Zaiquan Yang, Yuhao Liu, Jiaying Lin, Gerhard Hancke, and Rynson WH Lau. Boosting weakly-supervised referring image segmentation via progressive comprehension. In *NeurIPS*, 2024. 3

[66] Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. Improving commonsense in vision-language models via knowledge graph riddles. In *CVPR*, pages 2634–2645, 2023. 3

[67] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3D Question Answering . *IEEE TVCG*, 30(03):1772–1786, 2024. 3

[68] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, 2016. 2

[69] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR*, 2019. 2

[70] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, 2019. 2

[71] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, 2020. 2

[72] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Joint learning of salient object detection, depth estimation and contour extraction. *IEEE TIP*, 2022. 2

[73] Youjun Zhao, Jiaying Lin, Shuquan Ye, Qianshi Pang, and Rynson WH Lau. Openscan: A benchmark for generalized open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2408.11030*, 2024. 3

[74] Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. In *NeurIPS*, 2023. 3

[75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 3

[76] Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. Llafs: When large-language models meet few-shot segmentation. In *CVPR*, 2024. 3

[77] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. 5, 6, 7