

# Language-Guided Salient Object Ranking (Supplementary Material)

Anonymous CVPR submission

Paper ID 3993

In Supp. A, we first elaborate on the implementation details of other methods to establish a basis for comparison. We then provide a detailed implementation of our method, including loss function, text parsing, and text encoder. Next, in Supp. B, we compare the computational costs of the different methods. In Supp. C, we discuss potential conflicts that may exist between objects’ positional context and relations, and compare the proposed method with other graph-based SOR approaches. In Supp. D, we ablate the effects of the number of graph convolution layers, the number of rank decoder layers, sentence features from different sources, and so on. Finally, in Supp. E, we present additional visual results.

## A. Implementation Details

**Implementations of Other Methods.** For the salient instance/object detection methods [8, 19], we use the average saliency intensity to calculate the saliency ranks, following the method described in [13]. For instance segmentation methods [3, 9, 26], we modify the output layers to correspond with rank numbers from saliency ranking datasets and reinterpret category labels as rank labels.

For the large vision-and-language model-based method GiT [25], we adopt their instance segmentation setting, which adapts the window-based ViT [7, 14] to predict instance masks as a 1D sequence in raster order using an auto-regressive decoding paradigm. The number of output classes is adjusted according to the ranks of the two standard ranking datasets. A response example for instance segmentation in a sub-region of the input image is formatted as:  $\{(\text{image})\langle \text{local feature} \rangle \text{“instance segmentation”} : \langle c \rangle \langle x_1 \rangle \langle y_1 \rangle \dots \langle x_n \rangle \langle y_n \rangle\}$ , where  $\langle c \rangle$  is the class (rank) label, and  $\langle x_1 \rangle \langle y_1 \rangle \dots \langle x_n \rangle \langle y_n \rangle$  represent the  $n$  offsets of the predicted mask in the corresponding sub-region.

For the language-guided segmentation method (*i.e.*, X-Decoder [30]), we input the same detailed text descriptions as ours, generated by [4] and capped at 256 tokens. After extracting visual and language features using Focal-T [29] for images and BERT [6] for texts, the X-Decoder concatenates the language features with visual query embeddings and feeds them into transformer decoder layers for cross-modal feature fusion. The output layer of the X-Decoder is mod-

ified to meet the specific requirements of the two saliency ranking benchmarks.

For a fair comparison, backbone of instance segmentation methods [3, 9, 25, 26] and language-guided segmentation method [30] are also pretrained on MS-COCO [15].

**More Implementation Details.** We employ both binary cross-entropy loss and dice loss [16] for supervising salient instance mask predictions. They are denoted as  $\mathcal{L}_{mask} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}$ , where the coefficients  $\lambda_{ce}$  and  $\lambda_{dice}$  are both set to 5.0. For the rank predictions, we utilize ranking loss [18], denoted as  $\mathcal{L}_{rank}$ . The overall training loss is therefore:  $\mathcal{L} = \mathcal{L}_{mask} + \lambda_{rank}\mathcal{L}_{rank}$ , where  $\lambda_{rank}$  is empirically set as 2.0. Our framework is based on Detectron2 [27] and requires approximately one day to converge on eight A100 GPUs.

**Text Parsing.** To get the object and relation phrases, we use spaCy [1] to extract all noun and relation phrases from input descriptions. Subsequently, we filter out the pronouns (*e.g.*, she, he, it), atmosphere words (*e.g.*, a soft or cozy atmosphere), and non-instance phrases (*e.g.*, game, environment, street, city). If noun or relation phrases repeat, we retain only the first occurrence of each phrase.

**Text Encoder.** For the input description, we first tokenize it into 256 tokens. We truncate the tokenized text to 256 tokens if it exceeds this length, or pad it to meet this size if it is shorter. Additionally, we insert two special tokens: a [CLS] token at the start and a [SEP] token at the end of the description. We then feed this tokenized description into BERT [6] (“bert-base-uncased”) to obtain the corresponding output word embeddings. The output from the [CLS] token, designed to capture the context of the entire input description, is used as the sentence features for the input textual description. In the training process, we freeze the text encoder to stabilize the optimization process.

## B. Computational Analysis

**Computational Analysis of Different Methods.** In Table A1, we compare the FLOPs, parameters, inference speed, and performances of different methods on ASSR test set [21]. Although X-Decoder [30], which also processes input language descriptions using a text encoder (*i.e.*, BERT),

Table A1. Comparisons of FLOPs, Parameters, and Runtime of different methods. ‘-’ denotes the methods that do not use a text encoder or the results are not available. FLOPs are calculated for an input size of  $480 \times 640$ . Inference speed is tested on a single 4090 GPU.

Methods	Reference	Visual Encoder	Text Encoder	FLOPs ↓	Parameters ↓	Runtime ↓	SA-SOR ↑	SOR ↑	MAE ↓
X-Decoder [30]	[CVPR’23]	Focal-T	BERT	87.76G	148.41M	65ms	0.609	0.851	0.075
PSR [22]	[ACMMM’23]	ResNet-50	-	87.60G	50.88M	40ms	0.651	0.849	0.079
Ours	-	ResNet-50	BERT	139.44G	155.47M	68ms	<b>0.733</b>	<b>0.882</b>	<b>0.065</b>
GiT [25]	[ECCV’24]	GiT-B	-	-	158.74M	1398ms	0.541	0.854	0.101
OCOR [23]	[CVPR’22]	Swin-L	-	332.68G	401.67M	115ms	0.594	0.875	0.101
DSGNN [28]	[CVPR’24]	Swin-L	-	234.11G	250.21M	325ms	0.765	0.860	0.051
Ours	-	Swin-L	BERT	368.58G	327.03M	123ms	<b>0.787</b>	<b>0.895</b>	<b>0.049</b>

Table A2. Computation costs of different LVLMs.

LVLMs	LLaVa [17]	OPERA [12]	GPT-4V [2]	InstructBLIP [4]
Parameters ↓	3.81G	7.91G	1.7T	7.91G
Runtime ↓	1.89s	12.21s	≈7s	0.33s

Table A3. Computation costs of each component in our method.

Settings	TGVM	TAVR	FLOPs ↓	Parameters ↓	SA-SOR ↑
I			77.52G	153.50M	0.687
IV	✓		102.72G	154.29M	0.713
VIII	✓	✓	139.44G	155.41M	<b>0.733</b>

achieves slightly faster performance, our ResNet-50-based method significantly surpasses it across three metrics (SA-SOR: 20.36%; SOR: 3.64%; MAE: 23.07%). Our model’s slightly slower inference speed can be attributed to the additional components for language-aware graph reasoning and rank decoder branches. Although these components increase computational overhead, they significantly enhance our model’s ability to determine the saliency ranking of objects based on their interactions within the multimodal graph. Moreover, although PSR [22] achieves faster speeds, its performance lags behind our method by large margins (SA-SOR: 12.59%, SOR: 3.89%, MAE: 21.53%).

Compared to GiT [25], which predicts masks and ranks as a 1D sequence in a time-consuming autoregressive manner, the proposed (Swin-L-based) method leverages LVLM-generated descriptions to achieve significantly better performance (SA-SOR: 44.18%; SOR: 4.92%; MAE: 54.46%) while being approximately  $12\times$  faster. Notably, despite processing text input, our approach achieves  $1.6\times$  faster inference speeds than DSGNN [28] while delivering better performance, with a 4.19% improvement in the SOR metric.

**Computation Analysis of Different LVLMs.** As our method uses the LVLM-generated descriptions to assist the SOR task, we report the computational costs of different LVLMs in Table A2. We use InstructBLIP [9] as the default LVLM, as it is very efficient, and our method is robust to descriptions generated by different LVLMs (Table 4 in our paper).

**Computation Analysis of Different Components.** We further report the computational costs of different components of our model based on the ResNet-50 backbone in Table A3. Although introducing the TGVM and TAVR modules increases the computational overheads, they are essential for enhancing the performance of our model.

## C. More Discussions

**How to handle the conflicts between intrinsic positional context and objects’ relations in our model?** Our model processes them with the proposed TAVR module, which can enhance the reasoning of saliency ranks by explicitly modeling the extracted entity and relation cues from the input language description. Specifically, we initialize the graph node features by the visual features (which contain the intrinsic positional context of object instances) and entity cues in the description. We then use the relation cues to construct the edge connections of different nodes. After building this graph, we perform a graph reasoning process by graph convolution, which can empower the model to reason the saliency degree of an object, based on its relations with other objects learned from the graph structure. In this way, when the objects’ intrinsic positional context and their relations in the description conflict, the rank orders of the object can be enhanced or redefined according to its relations with other objects.

**Comparisons with Graph-based SOR Methods.** IRSR [18] introduces multiple graphs to capture local contrast, global contrast, person priors, and interactions between instances for ranking prediction. DSGNN [28] extracts shape and texture features for all objects, using them to build separate shape and texture graphs, with edges based on image features. QAGNet [5] leverages Graph Attention Network [24] to directly obtain edge features and aggregate node features according to multi-scale visual object features. HyperSOR [20] utilizes scene graph annotations to learn the semantic context of the scene.

In short, these graph-based SOR methods [5, 18, 28] model semantic relations via graph construction, where the semantics are *implicitly* learned from image features or scene

Table A4. Ablation study of activation function in gating mechanism. \* denotes the default setting.

Activation	SA-SOR $\uparrow$	SOR $\uparrow$	MAE $\downarrow$
Sigmoid	0.730	0.880	0.066
Tanh	<b>0.733</b>	<b>0.882</b>	<b>0.065</b>

Table A5. Ablation study of the number of graph convolution layers used in TAVR.

Layers	SA-SOR $\uparrow$	SOR $\uparrow$	MAE $\downarrow$
1	0.714	0.876	0.066
2	<b>0.733</b>	0.882	<b>0.065</b>
3	0.731	<b>0.883</b>	0.065

Table A6. Ablation study of number of rank decoder layers.

Layers	SA-SOR $\uparrow$	SOR $\uparrow$	MAE $\downarrow$
1	0.716	0.865	0.068
3	<b>0.733</b>	<b>0.882</b>	<b>0.065</b>
6	0.728	0.880	0.066

graph annotations. In contrast, our model *explicitly* extracts the semantic relations (e.g., parsed entity and relation cues) from the input descriptions.

## D. Additional Ablations

**Ablation Study of Different Activation Function.** We ablate different activation functions of the gating mechanism in Table A4. The results show that the tanh activation function can achieve the best performance, and we also treat it as a default setting.

**Number of Graph Convolution Layers.** In Table A5, we study the effect of varying the number of graph convolution layers. The results show that increasing the number of layers (e.g., from 1 to 2) enhances the model’s reasoning ability, but adding too many layers (e.g., 3) does not boost all metrics (with SOR being higher but SA-SOR and MAE worse). This may be attributed to the lengthy descriptions generated by LVLMs, where extra layers may over-complicate the model’s inference, introducing noise that negatively impacts the performance. Thus, we choose two graph convolution layers for our model.

**Number of Rank Decoder Layers.** Table A6 illustrates the impact of varying the number of rank decoder layers. The results show that using three layers yields the best performance across three metrics, surpassing both the single-layer and six-layer setups. Thus, we have selected three layers as the optimal configuration for our rank decoder.

**Different Sentence Features from Text Encoder.** In Table A7, we explore a variant where sentence features  $L_s$  are derived by averaging features from individual word tokens instead of using the [CLS] token, denoted as [CLS]  $\rightarrow$

Table A7. Ablation study of different sentence features.

Sentence Features	SA-SOR $\uparrow$	SOR $\uparrow$	MAE $\downarrow$
[CLS]	<b>0.733</b>	<b>0.882</b>	<b>0.065</b>
Words	0.729	0.880	0.066

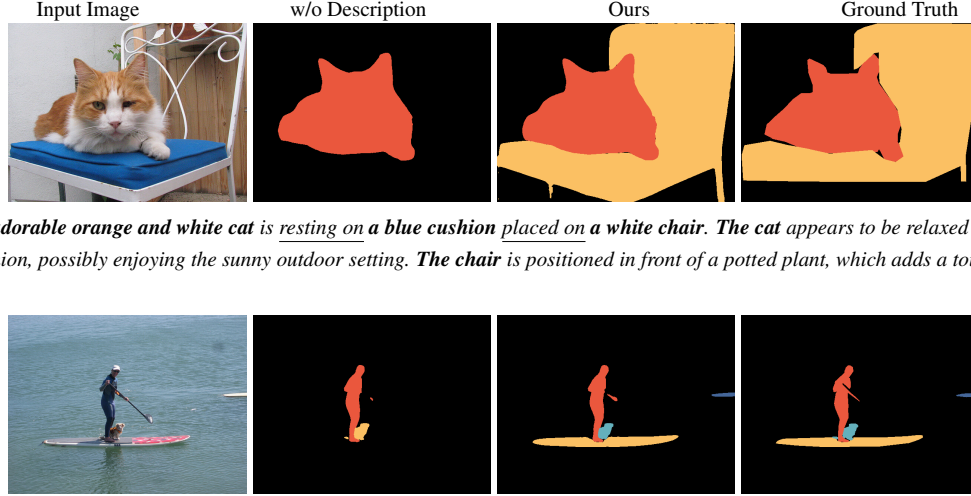
Table A8. Ablation study of different descriptions.

Captions	SA-SOR $\uparrow$	SOR $\uparrow$	MAE $\downarrow$
Wrong Descriptions	0.668	0.847	0.075
Ours	<b>0.733</b>	<b>0.882</b>	<b>0.065</b>

Words. The results indicate that representing sentence features with pooled word embeddings adversely affects performance. This validates that the [CLS] token more effectively captures the global context of the entire input textual description.

**Ablation Study of Wrong Description.** To further test our model, we deliberately feed completely wrong descriptions (using descriptions generated for other images) with the target images to our model. For example, we randomly sample one description and treat it as a completely wrong description for all target images. The results (denoted as “Wrong Descriptions”) are shown in Table A8, which indicates that using completely wrong descriptions indeed harms the model’s performance. This is expected, as the erroneous text describes another image, which is unrelated to the corresponding target images. However, since visual features are still used in our approach, our model can utilize visual features to aid in ranking and mitigate the negative effects of incorrect input descriptions, thus achieving comparable performance to other methods.

**Visual Comparisons of Introducing Language Description.** Fig. A1 illustrates the effectiveness of incorporating language descriptions into saliency ranking. The second column shows the model’s results without descriptions, where both TGVM and TAVR modules are removed, and we directly input the multiscale visual features into the transformer decoder to generate the ranking predictions. The visual results demonstrate that our approach effectively leverages the rich semantic information (e.g., salient objects and their relations) of textual descriptions to enhance saliency ranking. For example, in the first row, the model without using the description erroneously highlights only the most salient object (*‘the cat’*), but our method ranks *‘the cat’* as the most salient and *‘a chair’* as the second simultaneously, aligning with their orders in the description. The second row presents a more complex scenario. Unlike the model without descriptions, which only identifies the most salient object, our approach accurately recognizes *‘a woman’* and *‘a surfboard’* as the most and second salient objects, respectively, before shifting focus to *‘a small dog’* and *‘other surfboards’*.



**Description:** “An adorable orange and white cat is resting on a blue cushion placed on a white chair. The cat appears to be relaxed and comfortable as it lounges on the cushion, possibly enjoying the sunny outdoor setting. The chair is positioned in front of a potted plant, which adds a touch of greenery to the scene ...”

**Description:** “A woman is standing on a surfboard in the middle of a body of water, holding a paddle. She is accompanied by a small dog sitting on the surfboard next to her. Both the woman and the dog appear to be enjoying their time on the water. There are several other surfboards visible in the background, suggesting that there may be other people out on the water ...”

Figure A1. The effectiveness of introducing language description for saliency ranking.

This capability stems from our method’s effectiveness in extracting valuable information from the multimodal graph, utilizing implicit orders and relationships (e.g., ‘standing on’ and ‘sitting on’) detailed in the textual description.

**Description Comparisons of LVLMs.** In Table A9, we compare textual descriptions generated by various LVLMs [2, 4, 12, 17]. The results show that LLaVa [17] and OPERA [12] sometimes produce hallucinations that are not present in the original image. For example, as shown in Table A9, while the image contains ‘red chairs’, LLaVa erroneously identifies one of these chairs as a refrigerator, leading to a description that mistakenly includes ‘a refrigerator’. Similarly, OPERA generates descriptions that incorrectly label ‘three adults’, whereas the image actually depicts two women and two children, with an erroneous categorization of three individuals as adults. On the other hand, compared to InstructBLIP [4], the descriptions from GPT-4V [2] are more accurate and contain richer semantic information. For example, while InstructBLIP simply states ‘a woman is standing next to the cake’, GPT-4V provides a more detailed description, such as ‘a senior woman with blonde hair’, demonstrating its superior ability to capture and convey richer semantic details.

## E. More Visual Results

We present the comparison of our method against the eleven top-performing methods listed in Table 1, as shown in Fig. A2 – A7. These methods include Mask2Former [3], GiT [25], X-Decoder [30], ASSR [21], IRSR [18], OCOR [23], PSR [22], SeqRank [10], QAG-Net [5], DSGNN [28], and PoseSOR [11].

Fig. A2 – A4 demonstrate that our model effectively ranks salient objects based on their implicit orders in the description. For instance, the first row of Fig. A2 shows our model prioritizing ‘a man’ and ‘a wooden bench’ over ‘a boat in the water’, following the implicit narrative sequence in the description. Similarly, in first row of Fig. A3, the model identifies ‘a man’ and ‘a magazine with a picture of a woman on the cover’ as the most and second-most salient objects, respectively, with ‘a refrigerator’ ranked third. Even when background elements such as ‘a sink’ and ‘a clock’ are present, our approach effectively extracts useful information from the language description and suppresses these distractions. These results further validate that our method can decipher the implicit orders of salient objects from textual descriptions while effectively minimizing background noise.



Furthermore, our method exhibits robustness in capturing saliency by effectively deciphering positional context and object relationships, even when the order in the language description does not precisely align with the ground truth, which is demonstrated in Fig. A5 – A7. For example, the third row of Fig. A5 initially describes ‘a woman’, ‘a baby’ then ‘a white stuffed teddy bear’. But, according to the ground truth, ‘a baby’ should be the most salient. Despite this, our approach successfully utilizes relationships (e.g., ‘holding’ and ‘being enjoyed while held’) and the entity phrases to construct a multimodal graph, enhancing the model’s reasoning for the salient objects’ ranking orders. In the last row of Fig. A6, our method effectively learns the positional context from the multimodal graph by integrating semantic textual information with visual features.

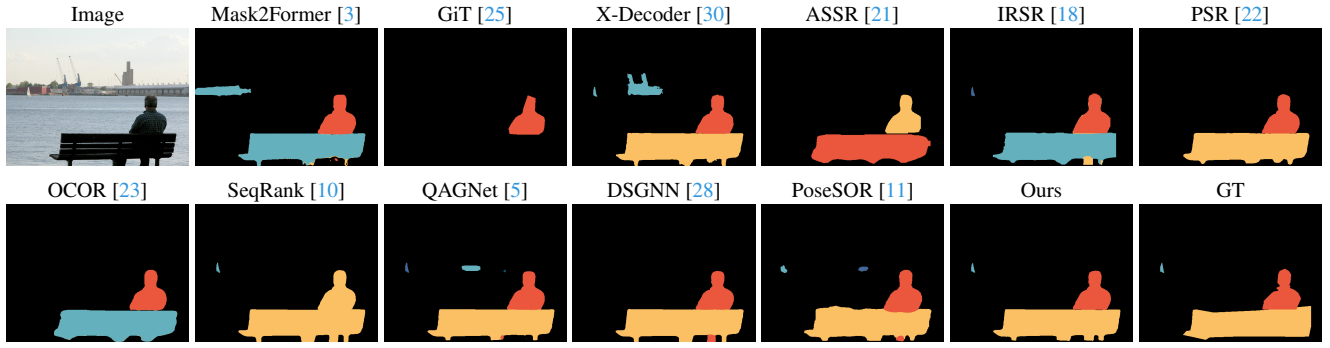


## References

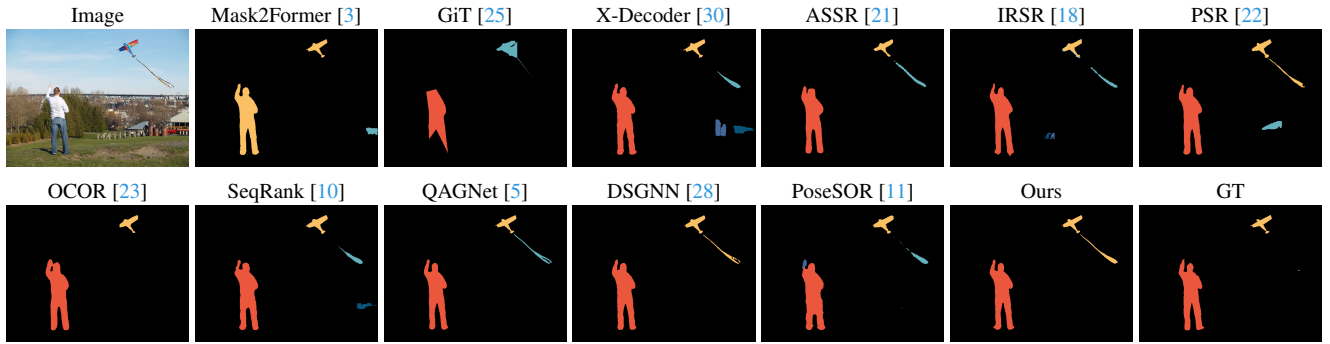
- [1] spacy. <https://spacy.io/>. 1
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2, 4, 6
- [3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 4, 7, 8, 9, 10, 11, 12
- [4] 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, 4, 6
- [5] 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, 4, 7, 8, 9, 10, 11, 12
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. 1
- [7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1
- [8] 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. 1
- [9] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 1
- [10] Huankang Guan and Rynson W.H. Lau. Seqrank: Sequential ranking of salient objects. In *AAAI*, 2024. 4, 7, 8, 9, 10, 11, 12
- [11] Huankang Guan and Rynson WH Lau. Posesor: Human pose can guide our attention. In *ECCV*, 2024. 4, 7, 8, 9, 10, 11, 12
- [12] 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. 2, 4, 6
- [13] 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. 1
- [14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 1
- [15] 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. 1
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 4, 6
- [18] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE TPAMI*, 2021. 1, 2, 4, 7, 8, 9, 10, 11, 12
- [19] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021. 1
- [20] 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
- [21] 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, 4, 7, 8, 9, 10, 11, 12
- [22] Chengxiao Sun, Yan Xu, Jialun Pei, Haopeng Fang, and He Tang. Partitioned saliency ranking with dense pyramid transformers. In *ACM MM*, 2023. 2, 4, 7, 8, 9, 10, 11, 12
- [23] 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. 2, 4, 7, 8, 9, 10, 11, 12
- [24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. 2
- [25] 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. 1, 2, 4, 7, 8, 9, 10, 11, 12
- [26] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 1
- [27] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 1
- [28] 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, 4, 7, 8, 9, 10, 11, 12
- [29] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *NeurIPS*, 2021. 1
- [30] 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. 1, 2, 4, 7, 8, 9, 10, 11, 12

Table A9. Comparisons of language descriptions generated by different LVLMs. We also display the parsed **nouns** and **relations**, highlighted in gray, using spaCy for each description. Only the left image was input into the LVLMs; the GT image on the right is only used for illustration.

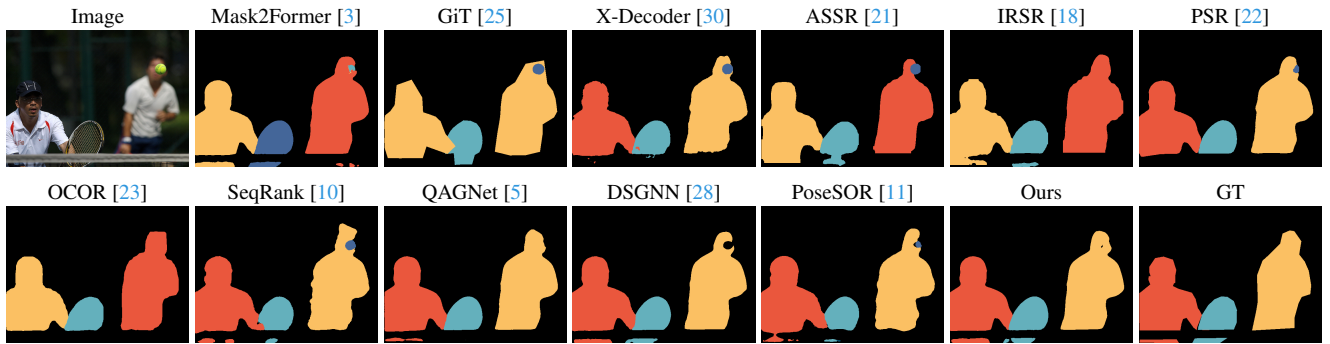
Visual example, Birthday Celebration:	
 	
User	Write a detailed description for the image. + [Left Image]
LLaVa [17]	<p>A family is gathered around a dining table, celebrating a special occasion. A woman is holding a white napkin, standing next to a young boy who is looking at a white cake with a pink ribbon and a single pink flower. Another woman is holding a baby in the background. The cake is on a glass cake stand on a dining table. The room has a red chair and a refrigerator, suggesting a homey kitchen setting. The atmosphere is one of celebration, likely a birthday or a similar event.</p> <p><b>nouns:</b> [a family, a dining table, a special occasion, a woman, a white napkin, a young boy, a white cake with a pink ribbon and a single pink flower, another woman, a baby, a glass cake stand, a red chair, a refrigerator]</p> <p><b>relations:</b> [gathered around, celebrating, holding, standing next to, looking at, on]</p>
OPERA [12]	<p>A group of people gathered around a small white cake on a table. A woman is standing in front of the cake, holding a knife, preparing to cut it. There are two children present, one on the left side of the table and the other on the right side. They seem to be excited about the upcoming cake-cutting ceremony. In addition to the children, there are three adults in the scene, two on the left side of the table and one on the right side. All of them are smiling and appear to be enjoying the celebration.</p> <p><b>nouns:</b> [a group of people, a small white cake, a table, a woman, a knife, two children, the left side of the table, the right side of the table, three adults]</p> <p><b>relations:</b> [gathered around, standing in front of, holding, preparing to cut]</p>
GPT-4V [2]	<p>A festive family gathering unfolds in a well-lit, homey kitchen. A senior woman with blonde hair, wearing a brown tank top and a festive headband, excitedly opens a gift, her expression one of surprise and delight. To her right, a young boy in a blue button-up shirt attentively watches her reaction. Nearby, a younger woman holding a toddler observes the scene, adding to the familial atmosphere. On the table, a pink and white birthday cake with a single candle suggests a celebration, possibly a birthday. The room is decorated with personal photos and vibrant red chairs, enhancing the cheerful, family-oriented setting.</p> <p><b>nouns:</b> [a festive family gathering, a senior woman with blonde hair, a brown tank top, a festive headband, a gift, a young boy in a blue button-up shirt, a younger woman, a toddler, a pink and white birthday cake with a single candle, personal photos, vibrant red chairs]</p> <p><b>relations:</b> [wearing, opens, watches, holding, observes, decorated with]</p>
InstructBLIP [4]	<p>A group of people gathered around a white cake on a table. A woman is standing next to the cake, preparing to cut it with a knife. There are two young boys in the scene, one on the left side of the table and the other on the right side. They appear to be excited about the upcoming birthday celebration. In addition to the people and the cake, there are several chairs scattered around the room. One chair can be seen on the left side of the image, while another is located on the right side. There is also a dining table in the scene, likely used for serving the cake and other refreshments during the celebration.</p> <p><b>nouns:</b> [a group of people, a white cake, a table, a woman, a knife, two young boys, chairs, a dining table]</p> <p><b>relations:</b> [gathered around, standing next to, preparing to cut, on the left side, on the right side, appear to be excited, scattered around, located on]</p>
User	Write a short description for the image. + [Left Image]
InstructBLIP [4]	<p>A group of people standing around a table with a cake on it.</p> <p><b>nouns:</b> [a group of people, a table, a cake]</p> <p><b>relations:</b> [standing around, with, on]</p>



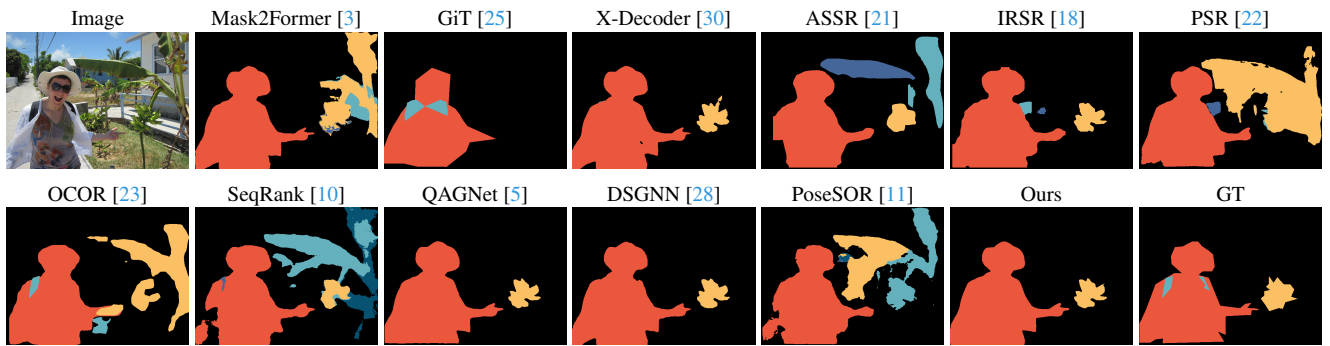
**Description:** “A man is sitting on a wooden bench near a body of water, possibly the ocean. He is wearing a plaid shirt and appears to be enjoying the view. The scene also includes a boat in the water and a crane in the background, ...”



**Description:** “A man standing on a grassy field, flying a colorful kite in the sky. He is wearing a white shirt and appears to be enjoying the outdoor activity. In the background, there is a bridge spanning across the landscape, ...”



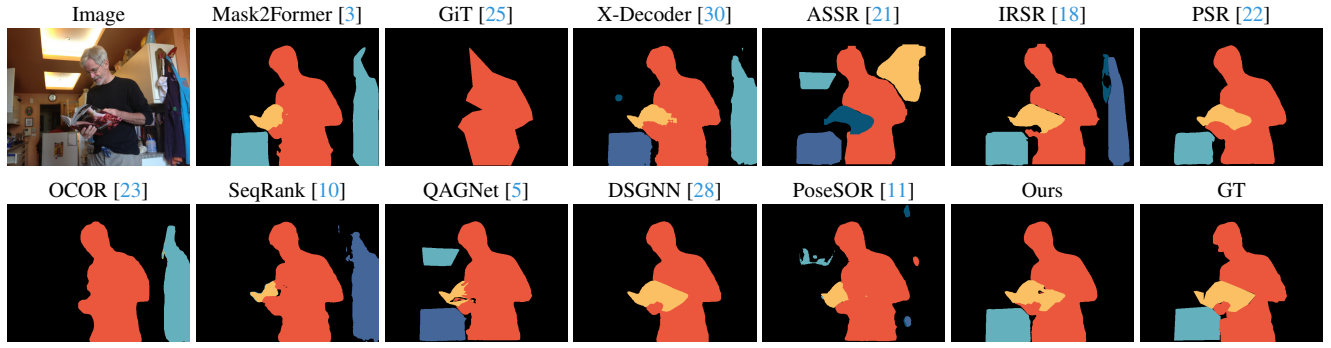
**Description:** “Two men are playing a game of tennis on a tennis court. One of them is holding a tennis racket, ready to hit the ball that is coming towards him. The other man is standing nearby, possibly waiting for his turn to hit the ball ...”



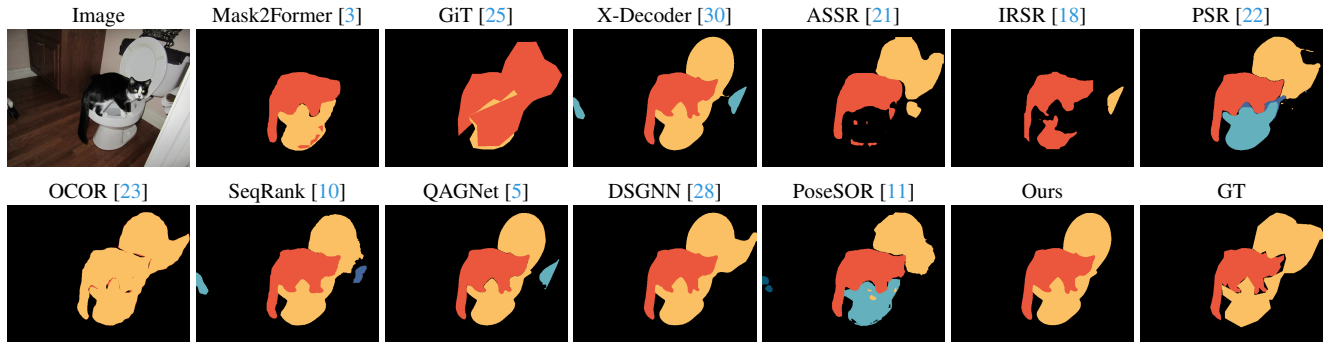
**Description:** “A woman wearing a hat and sunglasses is standing on a sidewalk next to a banana tree. She appears to be pointing at something in the distance, possibly enjoying the scenery around her. The banana tree is located towards ...”

high  low

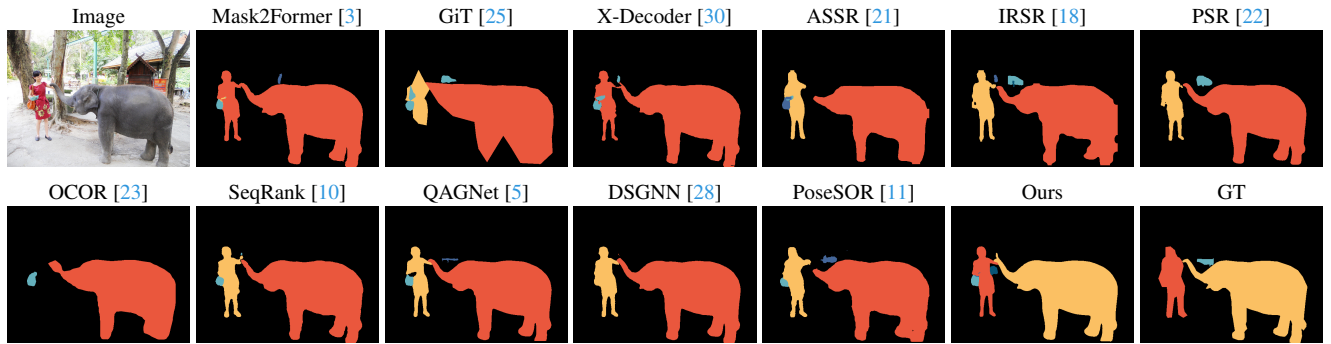
Figure A2. Qualitative comparison of our method with eleven best-performing methods in Table 1.



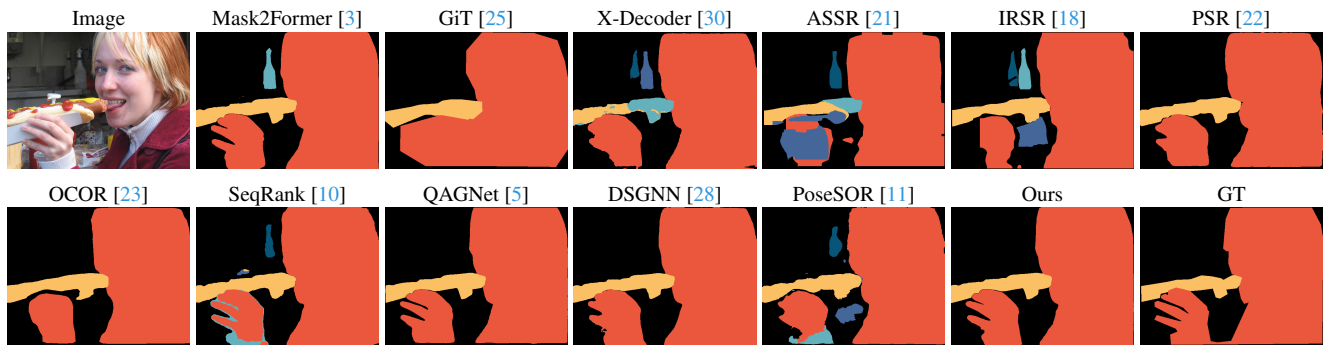
**Description:** “A man is standing in a kitchen, holding a magazine with a picture of a woman on the cover. He is wearing a black shirt and appears to be reading the magazine. The kitchen has a refrigerator, a sink, and a clock on the wall ...”



**Description:** “A black and white cat is standing on top of a white toilet in a bathroom. The cat appears to be enjoying its time on the toilet, as it seems to be licking its paw ...”



**Description:** “A woman is standing next to a small elephant in a dirt area. The elephant is reaching out its trunk towards the woman, who appears to be petting it. The elephant’s trunk is touching the woman’s hand, and she is smiling ...”

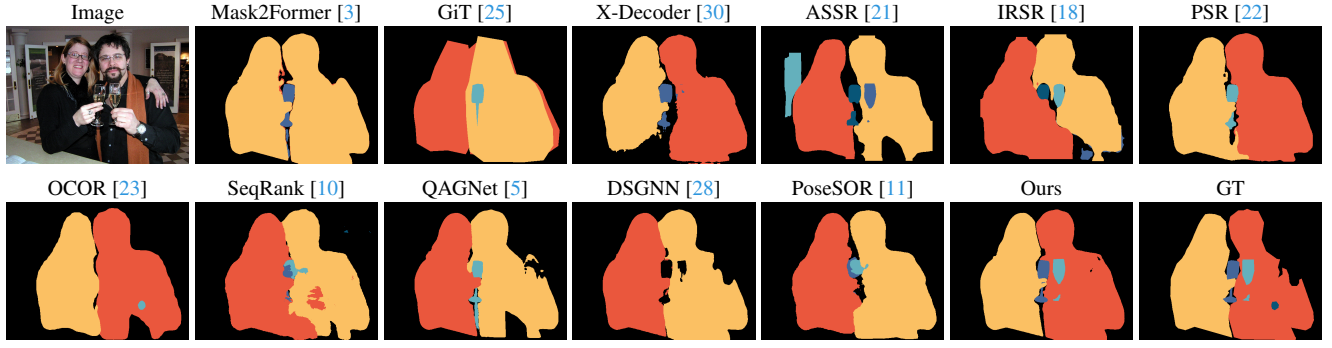


**Description:** “A young woman wearing a red jacket is biting into a long hot dog covered in mustard and ketchup. The hot dog appears to be quite large, taking up a significant portion of the woman’s mouth. ...”

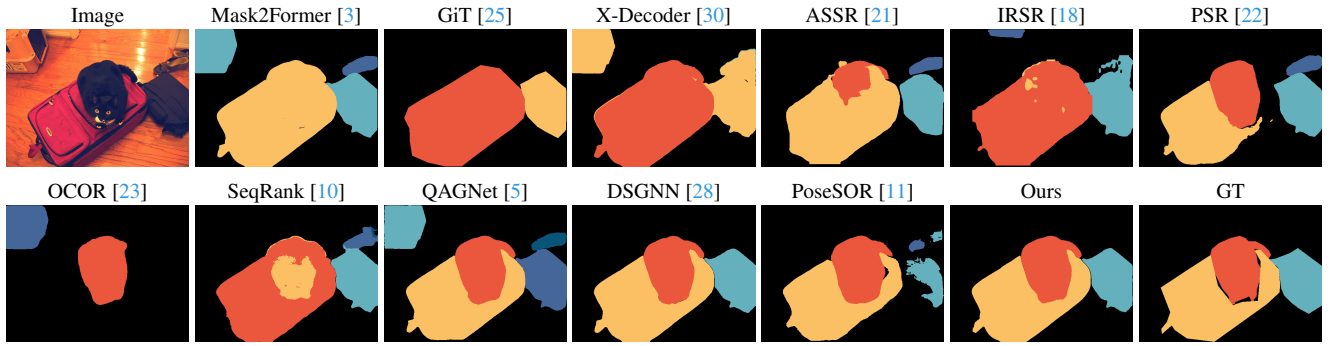
high  low

Figure A3. Qualitative comparison of our method with eleven best-performing methods in Table 1.

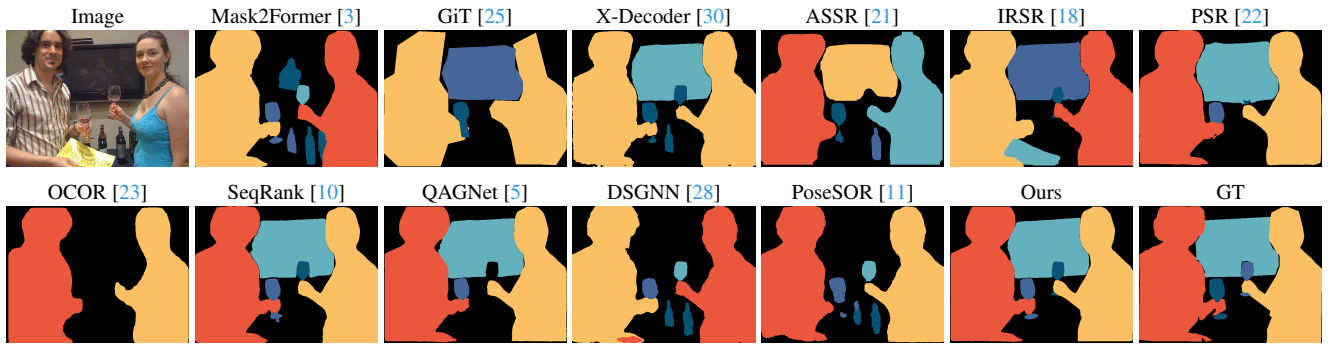




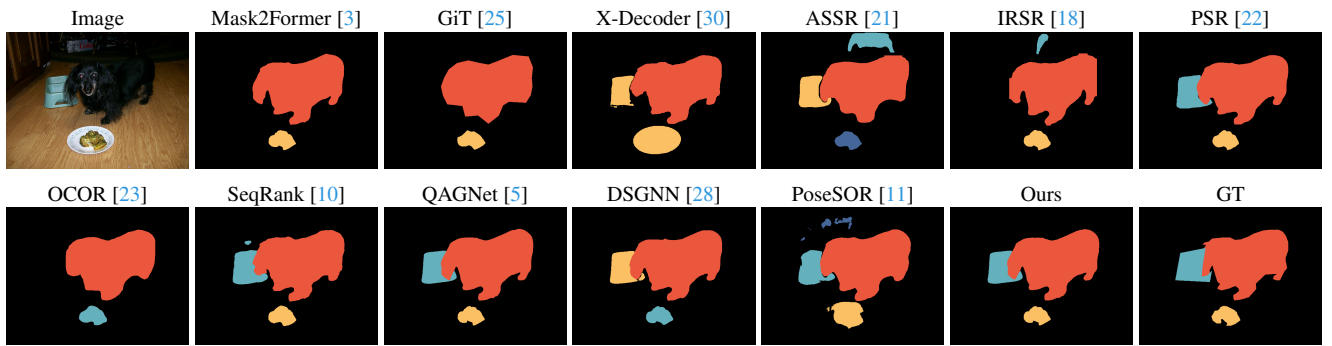
**Description:** “A man and a woman are standing together, holding wine glasses and posing for a picture. The man is wearing glasses and a scarf, while the woman is wearing a black shirt. They are both smiling ...”



**Description:** “A black cat is sitting on top of a red suitcase, which is placed on the floor. The cat appears to be looking at the camera. There is a backpack next to the suitcase. The scene is set in a room with a wooden floor. ...”



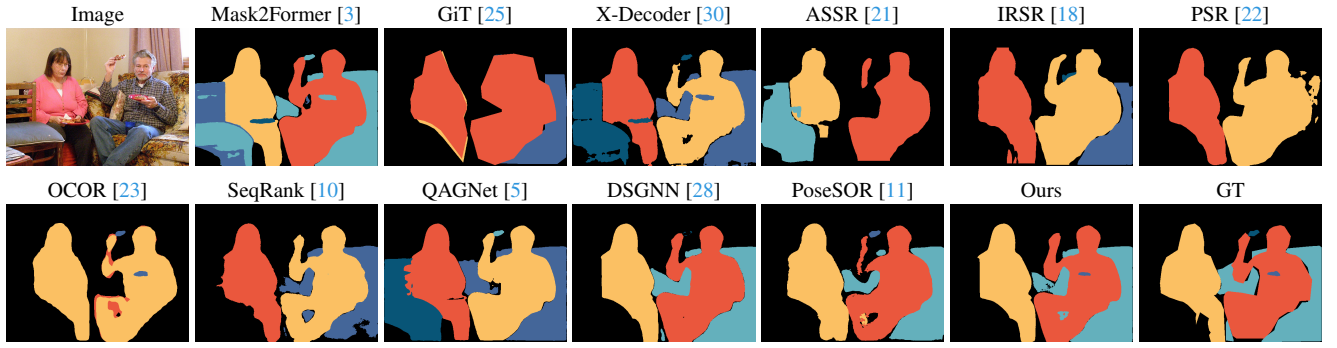
**Description:** “A man and a woman are standing together, holding wine glasses, and posing for a picture. They are surrounded by several wine bottles. The man is wearing a striped shirt, and the woman is wearing a blue dress ...”



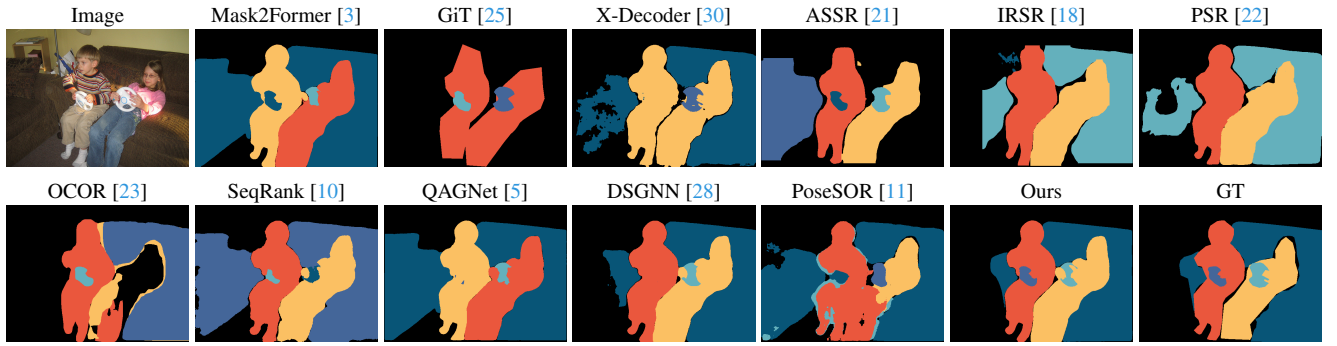
**Description:** “A small black dog is standing on a wooden floor next to a white plate containing broccoli. The dog seems to be curious about the broccoli and is staring intently at it. There are a bowl on the left side of the image ...”

high  low

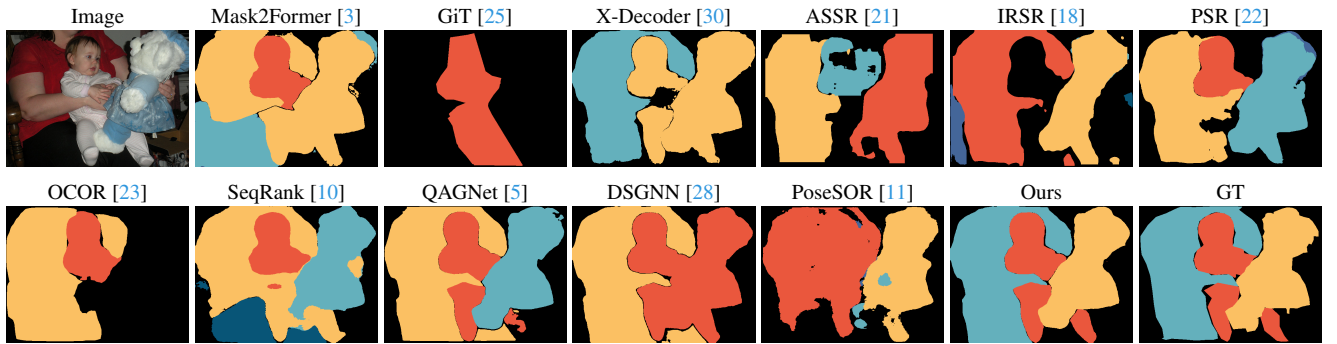
Figure A4. Qualitative comparison of our method with eleven best-performing methods in Table 1.



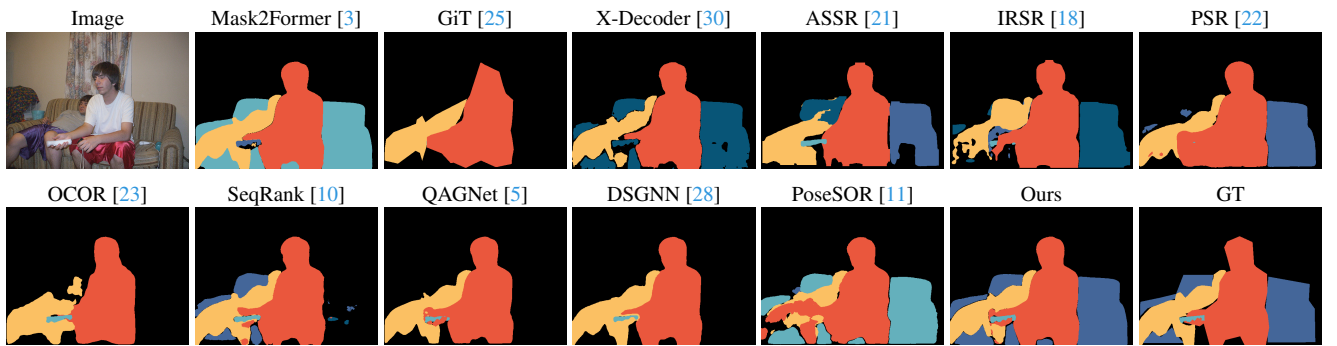
**Description:** “A man and a woman are sitting on a couch, enjoying pizza together. The man is holding a slice of pizza in his hand, while the woman is also holding a slice. There are two pizza slices visible in the scene ...”



**Description:** “A young boy and girl are sitting on a couch, playing a video game together. They are holding Wii controllers in their hands, fully engaged in the game. The room has a cozy atmosphere, ...”



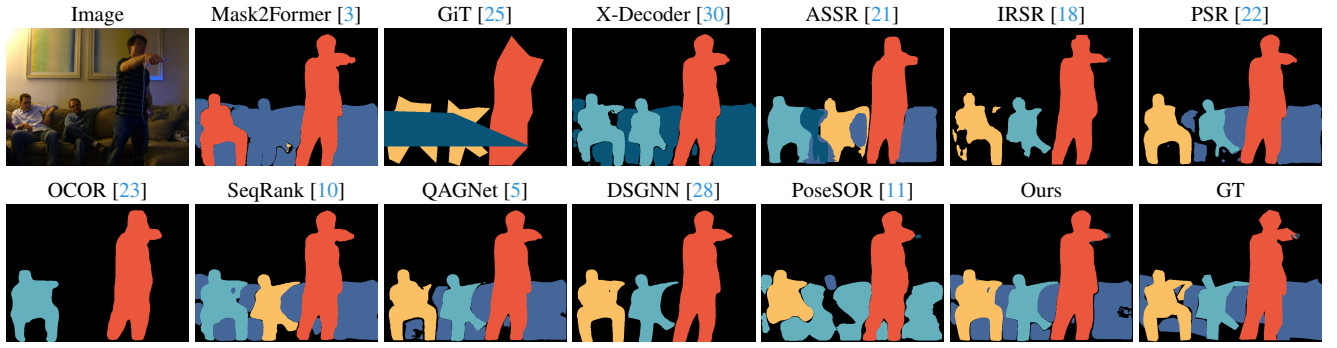
**Description:** “A woman is sitting in a chair and holding a baby close to her chest. The baby is wearing a pink outfit and appears to be enjoying being held by the woman. The woman is also holding a white stuffed teddy bear in her other hand. ...”



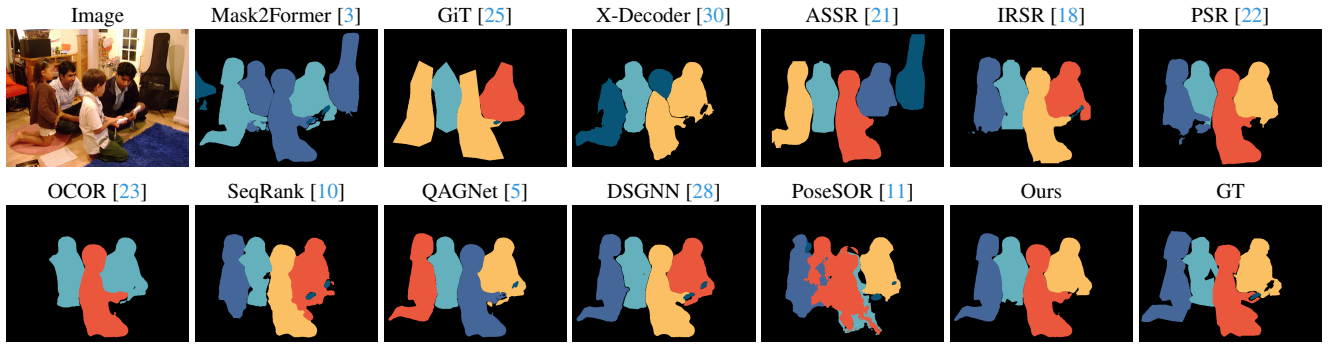
**Description:** “Two young men are sitting on a couch, playing a video game together. One of the men is holding a Wii remote, while the other is holding a Nintendo Wii controller. The couch is positioned against a wall, ...”

high  low

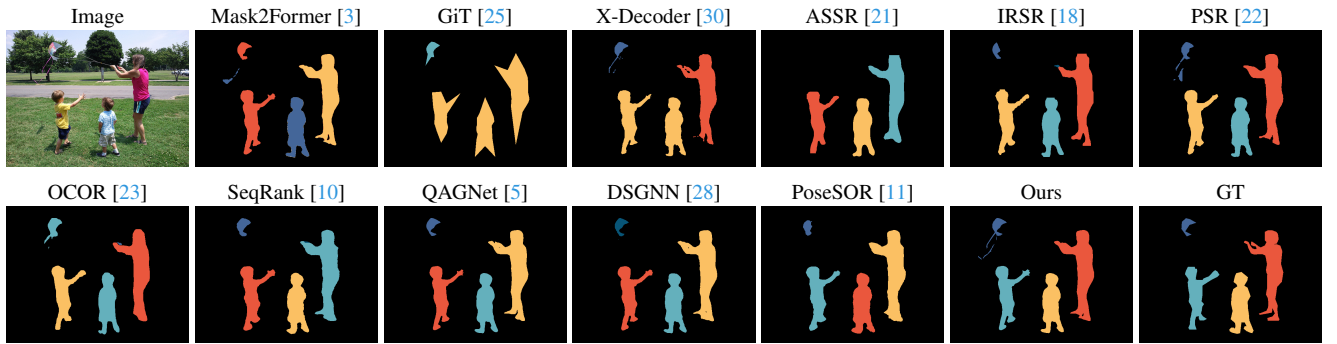
Figure A5. Qualitative comparison of our method with eleven best-performing methods in Table 1.



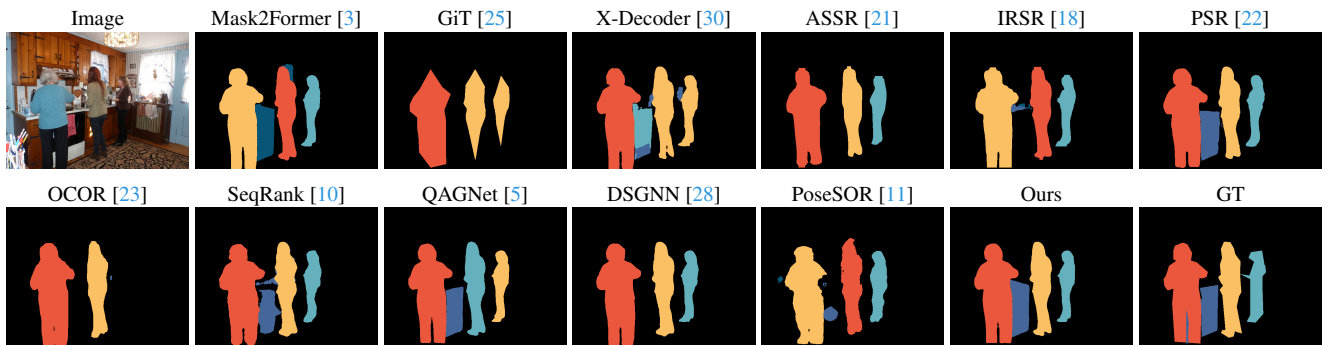
**Description:** “A group of people gathered in a living room, playing a video game together. A man is holding a Wii remote, and playing a video game. Other people are sitting on a couch, watching the man play. The living room is furnished ...”



**Description:** “A group of people gathered in a living room, playing a video game together. Two adults and two children are sitting on the floor, all holding Nintendo Wii controllers. They appear to be enjoying the game ...”



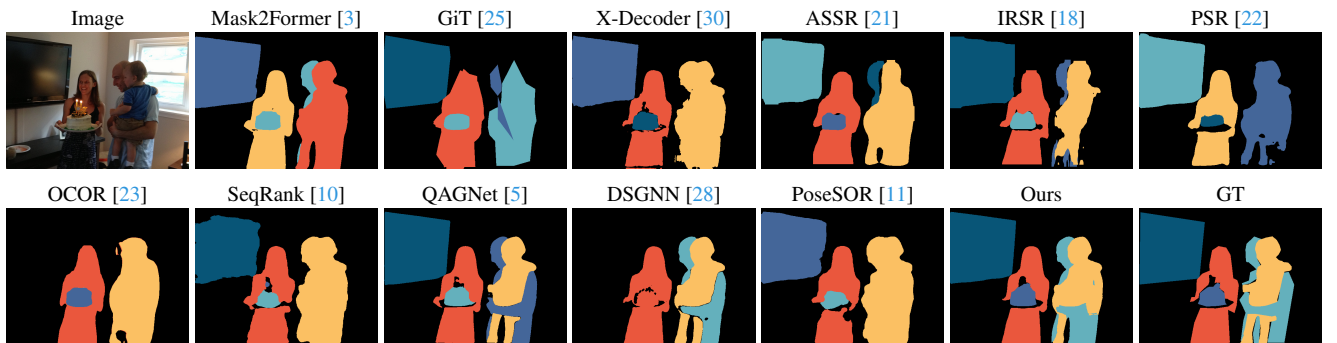
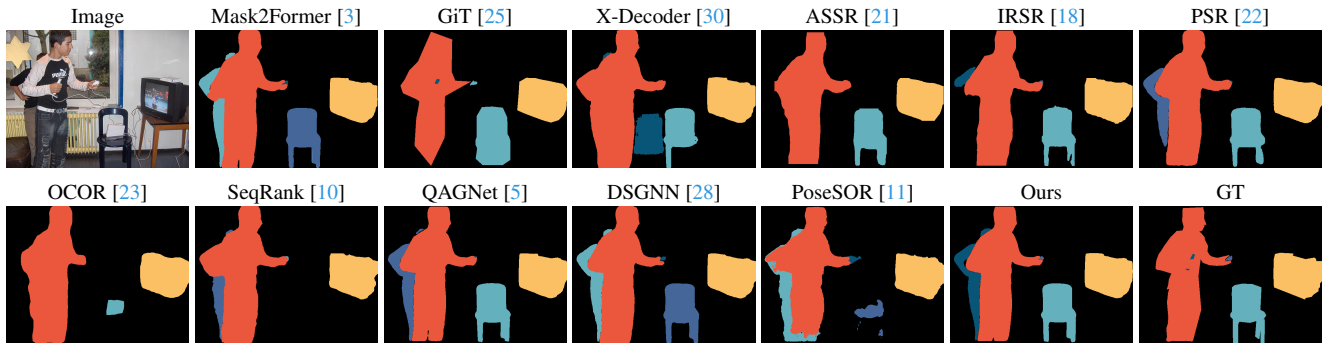
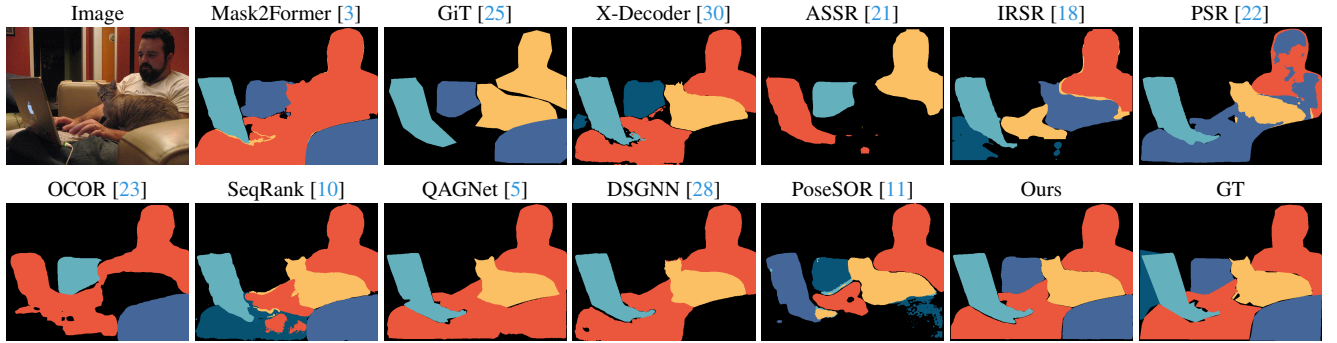
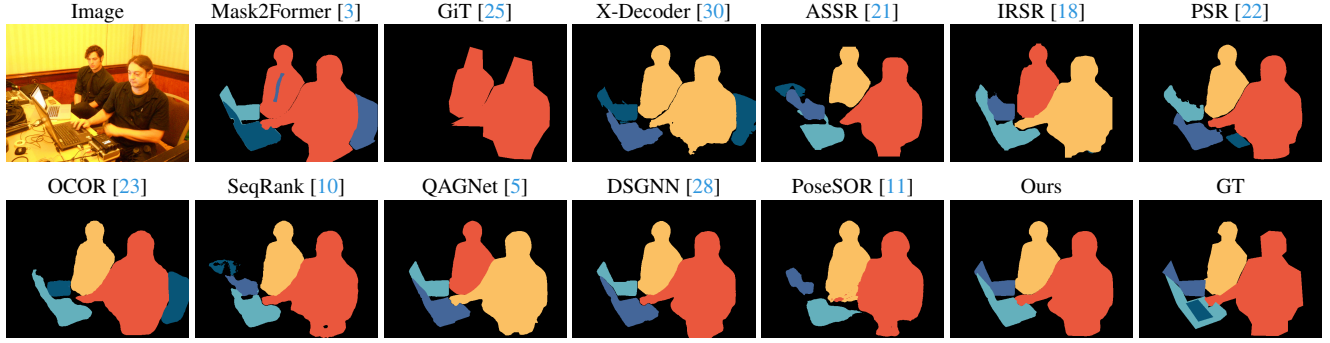
**Description:** “A woman and two young boys are playing with a kite in a park. The woman is holding the kite while the boys are standing next to her, looking up at the kite as it flies in the air. It appears to be a fun and enjoyable moment ...”



**Description:** “A group of three women are standing in a kitchen, preparing food and drinks. Two of the women are on the left side of the kitchen, while the third woman is on the right side. They appear to be engrossed in their tasks, ...”

high  low

Figure A6. Qualitative comparison of our method with eleven best-performing methods in Table 1.



high  low

Figure A7. Qualitative comparison of our method with eleven best-performing methods in Table 1.