NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

Paper Code Contact

# OneRef: Unified One-tower Expression Grounding and Segmentation with Mask Referring Modeling

Linhui Xiao[1,2,3], Xiaoshan Yang[1,2,3], Fang Peng[1,2,3], Yaowei Wang[2,4], Changsheng Xu[1,2,3*]
[1] MAIS, Institute of Automation, Chinese Academy of Sciences [2] Pengcheng Laboratory
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences [4] HIT (Shenzhen)
E-mail: xiaolinhui16@mails.ucas.ac.cn,
Github: https://github.com/linhuixiao/OneRef

Dec 10th –Dec 15th , 2024

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Contents

- Motivation
- Methodology
- Experiments
- Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Contents

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Motivation

Visual Grounding (VG) aims to ground a region referred by an expression query text in a specific image.

The generalized VG / referring tasks include Referring Expression Comprehension (REC), Phrase Grounding (PG), and Referring Expression/Image Segmentation (RES/RIS). In REC/PG, the grounding region is represented by a rectangular boundary box, while in RES/RIS, it is represented by an irregular fine-grained segmented mask of the referred object .

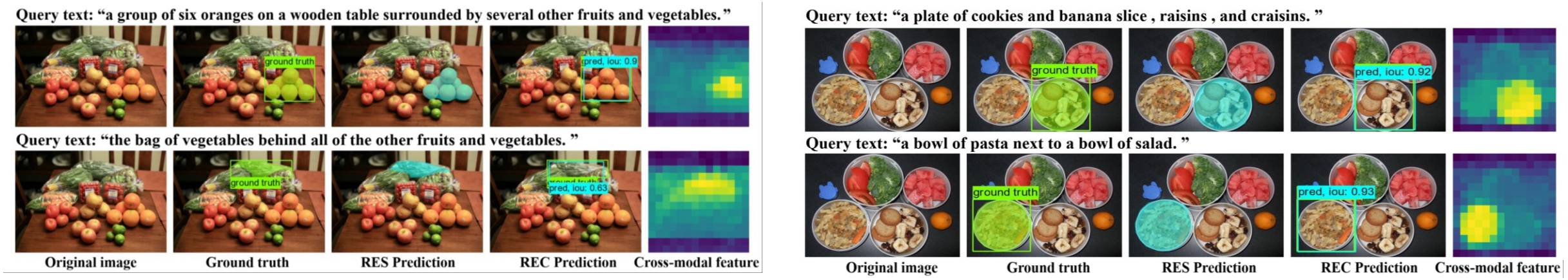


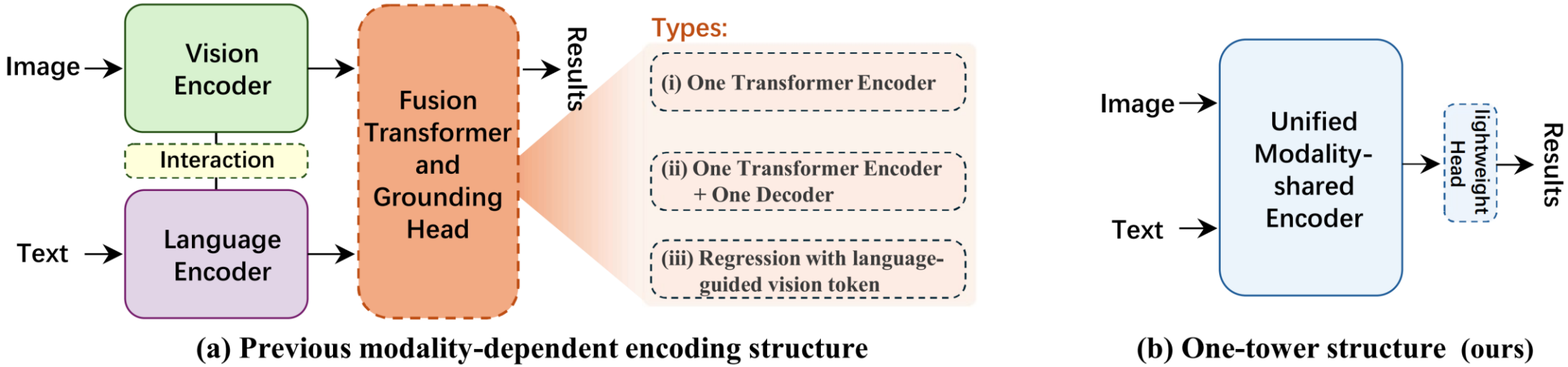

**Fig.1 Referring Expression Comprehension (REC) and Segmentation (RES) tasks.**

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Motivation

Constrained by the separate encoding of vision and language, existing grounding and referring segmentation works heavily rely on bulky Transformer-based fusion en-/decoders and a variety of early-stage interaction technologies.

These previous structures restrict the alignment between visual and language modalities, while also introducing significant computational overhead.




(a) Previous modality-dependent encoding structure

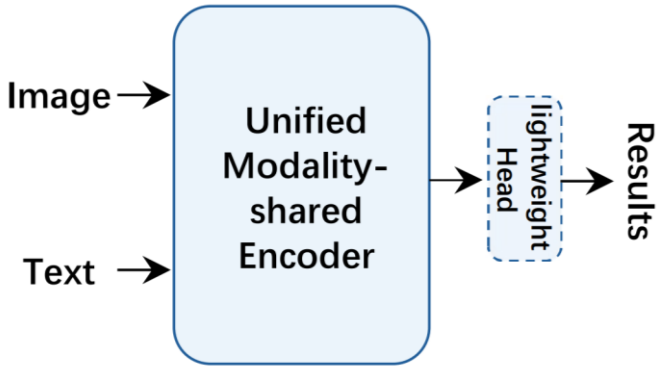
(b) One-tower structure (ours)

**Fig.3 Previous REC/RES model and Our OneRef model.**

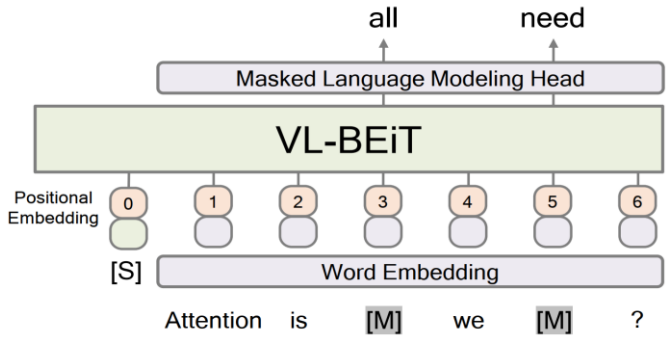
NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Motivation

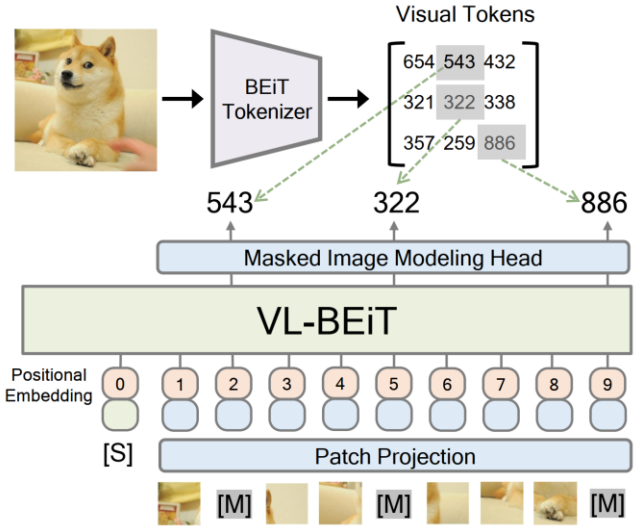
However, the existing one-tower backbone networks (e.g., BEiT-3, VL-BEiT) utilize masked visual language modeling (MVLM) as a self-supervised pre-training paradigm. Nevertheless, the vanilla MVLM employs alternating mask language modeling (MLM) and mask image modeling (MIM), making it difficult to capture the nuanced referential relationship between image-text pairs in the referring task.




(a) One-tower model




(b) Vanilla MLM




(a) Vanilla MVM

**Fig.4 One-tower model and the vanilla MVLM.**

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Contents

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Methodology

We propose ***OneRef***, a minimalist referring framework built on the modality-shared one-tower transformer that unifies the visual and linguistic feature spaces. To modeling the referential relationship, we introduce a novel MVLM paradigm called **Mask Referring Modeling (*MRefM*)**, which encompasses both **referring-aware mask image modeling (Referring MIM)** and **referring-aware mask language modeling (Referring MLM)**.

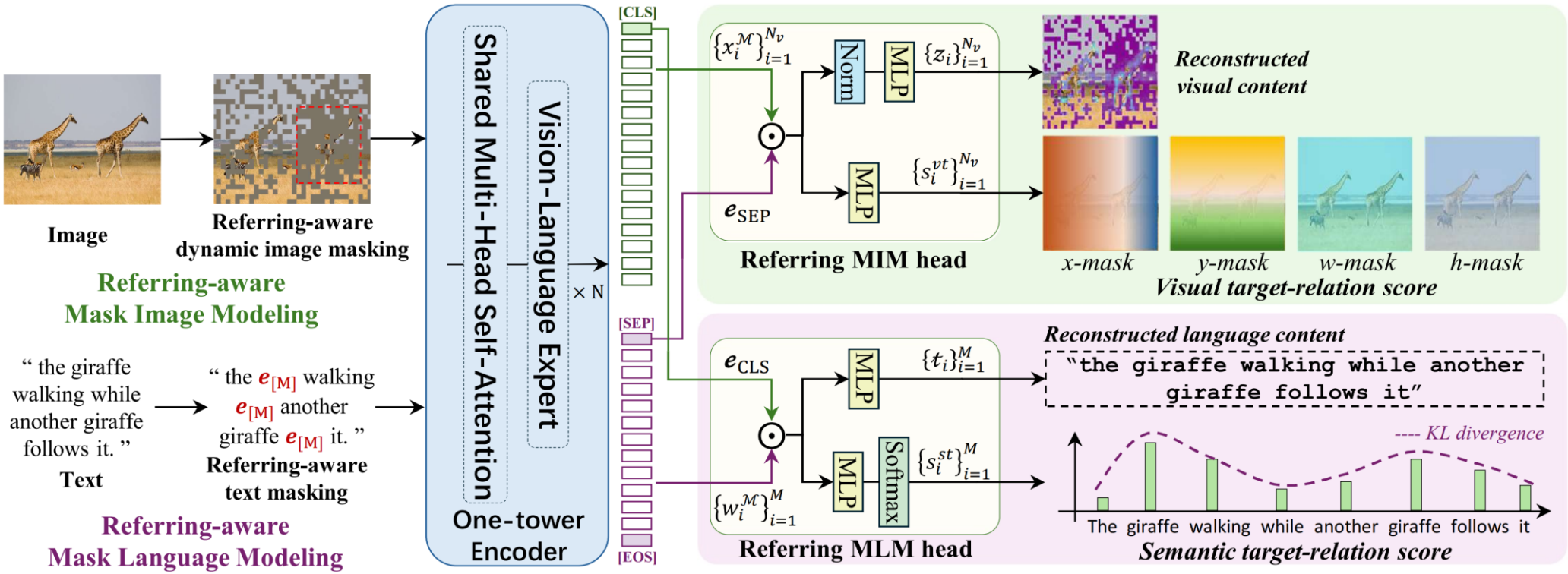



**Fig.5 Multimodal Mask Referring Modeling (MRefM) paradigm.**

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Methodology

Both **referring-aware mask image modeling (Referring MIM)** and **referring-aware mask language modeling (Referring MLM)** modules not only reconstruct modality-related content but also cross-modal referring content.

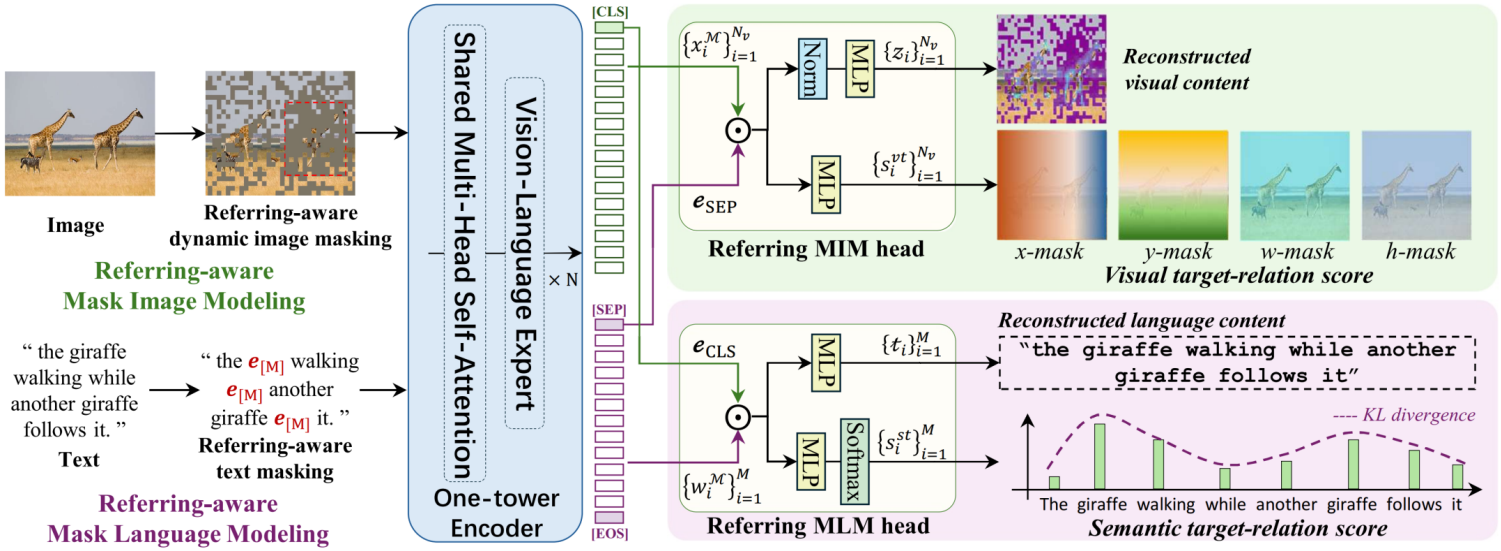



**Fig.5 Multimodal Mask Referring Modeling (MRefM) paradigm.**

**Vanilla MVLM:**

$$\mathcal{L}_{\text{MIM}} = -\sum_{\boldsymbol{x}\in\mathcal{I}}\sum_{i\in\mathcal{M}_v}\log p(\boldsymbol{z}_i|\boldsymbol{x}_i^{\mathcal{M}}) \qquad \mathcal{L}_{\text{MLM}} = -\sum_{\boldsymbol{x}\in\mathcal{T}}\sum_{i\in\mathcal{M}_w}\log p(\boldsymbol{t}_i|\boldsymbol{w}_i^{\mathcal{M}})$$

**Mask Referring Modeling (*MRefM*):**

$$\mathcal{L}_{\text{Referring MIM}} = -\sum_{\boldsymbol{x}\in\mathcal{I}}\sum_{i\in\mathcal{M}_v}\log p(\boldsymbol{z}_i|(\boldsymbol{x}_i^{\mathcal{M}}\odot\boldsymbol{e}_{\texttt{SEP}})) - \sum_{\boldsymbol{x}\in\mathcal{I}}\sum_{i\in[1,N_v]}\log p(\boldsymbol{s}_i^{vt}|(\boldsymbol{x}_i^{\mathcal{M}}\odot\boldsymbol{e}_{\texttt{SEP}}))$$

$$\mathcal{L}_{\text{Referring MLM}} = -\sum_{\boldsymbol{w}\in\mathcal{T}}\sum_{i\in\mathcal{M}_w}\log p(\boldsymbol{t}_i|(\boldsymbol{w}_i^{\mathcal{M}}\odot\boldsymbol{e}_{\texttt{CLS}})) - \sum_{\boldsymbol{w}\in\mathcal{T}}\sum_{i\in[1,M]}\log p_{kl}(\boldsymbol{s}_i^{st}|(\boldsymbol{w}_i^{\mathcal{M}}\odot\boldsymbol{e}_{\texttt{CLS}}))$$

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Methodology

In **Referring MIM**, the visual target-relation scores $\{s_i^{vt}\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times 4}$ represent the distance between each patch token $\{x_i^{\mathcal{M}}\}_{i=1}^{N_v}$ and the referred region $(x_c, y_c, w_r, h_r)$, where $(x_c, y_c, w_r, h_r)$ denote the center coordinate and the width and height of the referred region. By slicing the predicted score, four masks can be derived. The score represents the spatial distance and the relative size between the current patch region and the referred region.

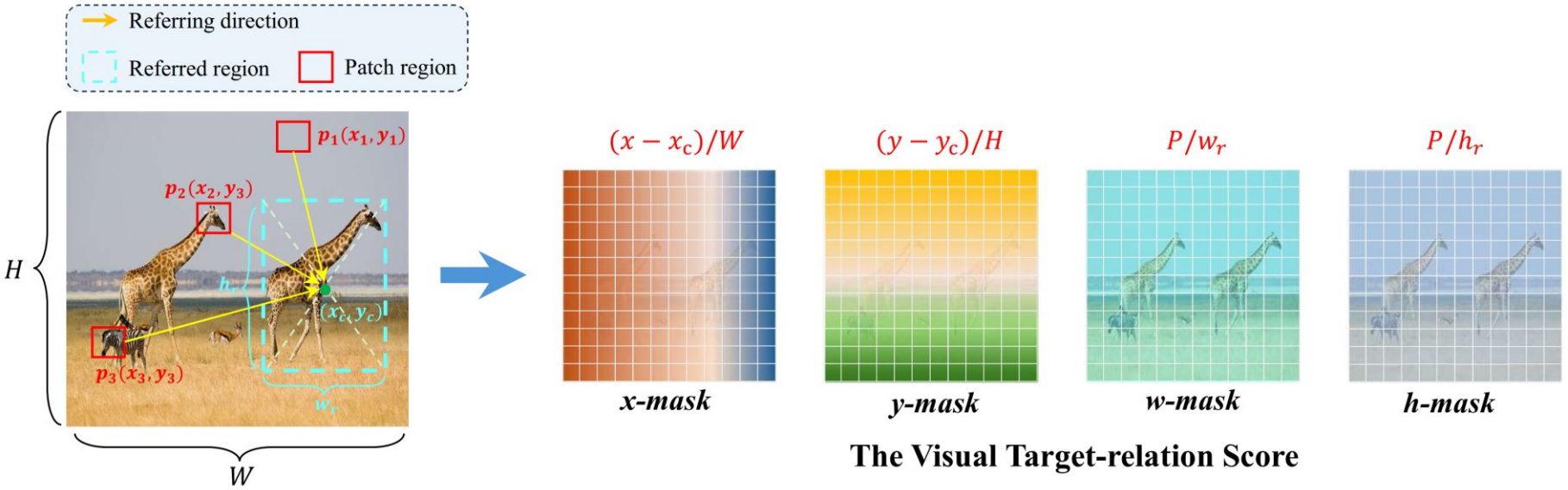



**Fig.6 The Visual Target-relation Score.**

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Methodology

In **Referring MIM**, we propose referring-aware dynamic image masking strategy with dynamic mask ratio $\alpha$ that is aware of the referred region rather than relying on fixed ratios or generic random masking schemes.

$$\alpha = [\beta \cdot (N_v - N_r) + \gamma \cdot N_r]/N_v.$$

**Algorithm 1** Referring-aware Dynamic Masking

**Input:** $N_v$ *image patches*, $N_r$ $(h_{rp} \times w_{rp})$ *referred patches*.
**Output:** *Dynamic masked positions* $\mathcal{M}$.

$\boldsymbol{c} \leftarrow$ Rand Select $\beta \cdot N_v$ numbers in $[1, N_v]$
New $\mathcal{M} \in \mathbb{R}^{1 \times N_v}$, $\{\{\mathcal{M}_i\}_i^{N_v} \mid \mathcal{M}_i = 1$ if $i \in \boldsymbol{c}$, else $0\}$
$\mathcal{M} \leftarrow \mathcal{M}$ reshape as $\mathcal{M} \in \mathbb{R}^{h \times w}$ ▷ *In-context masking*
New $\mathcal{M}_r \in \mathbb{R}^{h_{rp} \times h_{rp}}$ with all as 0 ▷ *Referred masking*
**while** $|\mathcal{M}_r| \leq \gamma \cdot N_r$ **do**
  $s \leftarrow$ Rand$(1, \gamma \cdot N_r - |\mathcal{M}_r|)$ ▷ *Block size*
  $r \leftarrow$ Rand$(a, \frac{1}{a})$ ▷ *Aspect ratio of block*
  $w_b \leftarrow \sqrt{s/r}; h_b \leftarrow \sqrt{s \cdot r}$ ▷ *Width, height of block*
  $l \leftarrow$ Rand$(0, w_{rp} - w_b); t \leftarrow$ Rand$(0, h_{rp} - h_b)$
  $\{\mathcal{M}_r(i,j) = 1 \mid i \in [l, l + w_b), j \in [t, t + h_b)\}$
**end**
$\mathcal{M}(x_{sp} : x_{sp} + w_{rp}, y_{sp} : y_{sp} + h_{rp}) = \mathcal{M}_r$
**return** $\mathcal{M}$.

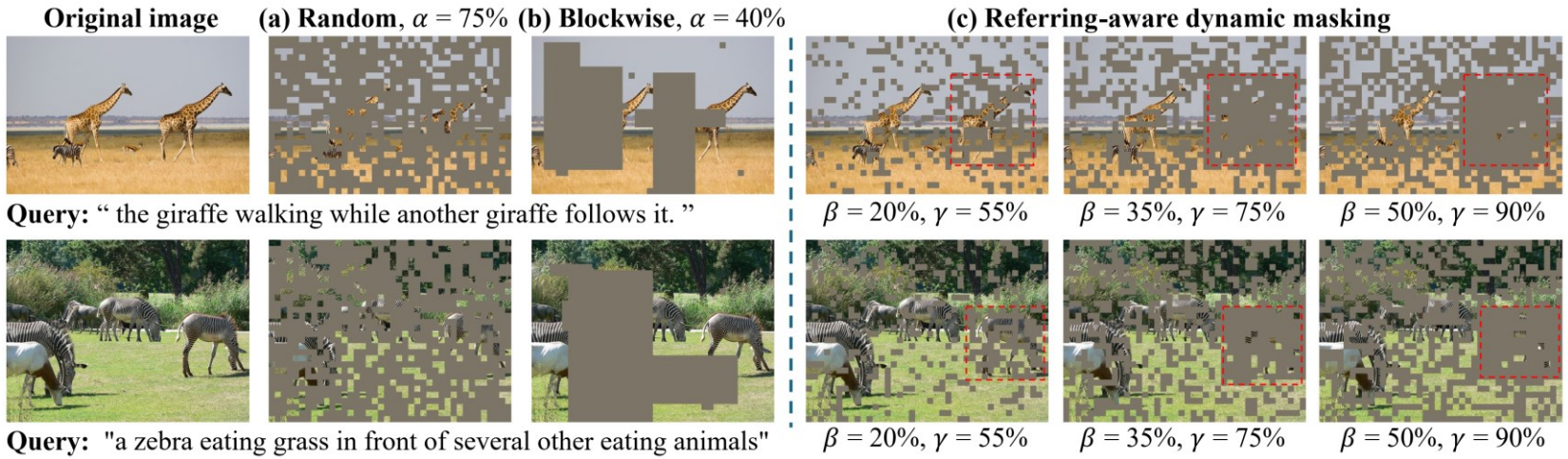



**Fig.7 Our proposed referring-aware dynamic masking strategy.**

In **Referring MLM**, the semantic target-relation score (referring weight distribution) between vocabularies and referred region are obtained as:

$$\boldsymbol{s}^{st} = \lambda_{reg} \cdot \sigma(< \boldsymbol{e}_{\texttt{CLS}}^{reg\top}, \{\boldsymbol{w}_i\}_{i=1}^{M} >) + \lambda_{img} \cdot \sigma(< \boldsymbol{e}_{\texttt{CLS}}^{img\top}, \{\boldsymbol{w}_i\}_{i=1}^{M} >)$$

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Methodology

By leveraging the unified visual language feature space and incorporating MRefM's ability to model the referential relations, our approach enables direct regression of the referring results without resorting to various complex techniques.

$$\hat{\mathcal{B}} = \mathrm{MLP}(\sum_{i\in[1,N_v]} (\mathrm{Repeat}(\sigma(< \boldsymbol{e}_{\mathrm{ESP}}{}^{\top}, \{\boldsymbol{x}_i\}_{i=1}^{N_v} >)) \odot \mathrm{MLP}(\{\boldsymbol{x}_i\}_{i=1}^{N_v}))).$$

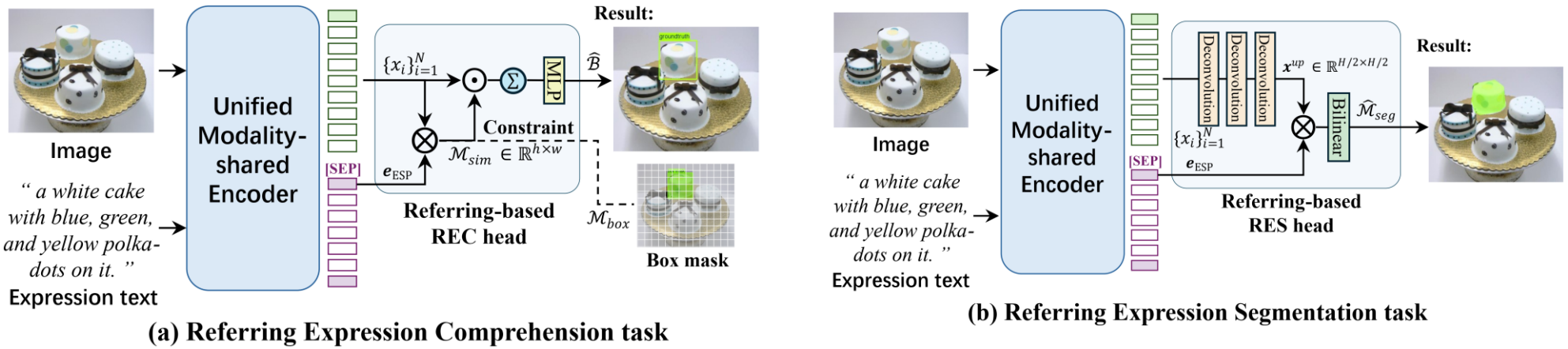



**Fig.8 Referring-based grounding and segmentation transfer.**

**Training Objectives for REC and RES transfer:**

REC:
$$\mathcal{L}_{REC} = \lambda_{L_1}\mathcal{L}_{\mathrm{L1}}(\hat{\mathcal{B}}, \mathcal{B}) + \lambda_{giou}\mathcal{L}_{\mathrm{giou}}(\hat{\mathcal{B}}, \mathcal{B}) + \mathcal{L}_{box\_mask\_constraints}$$

$$\mathcal{L}_{box\_mask\_constraints} = \lambda_{f\_box}\mathcal{L}_{focal}(\mathcal{M}_{sim}, \mathcal{M}_{box}) + \lambda_{d\_box}\mathcal{L}_{dice}(\mathcal{M}_{sim}, \mathcal{M}_{box}).$$

RES:
$$\mathcal{L}_{RES} = \lambda_{f\_seg}\mathcal{L}_{focal}(\hat{\mathcal{M}}_{seg}, \mathcal{M}_{seg}) + \lambda_{d\_seg}\mathcal{L}_{dice}(\hat{\mathcal{M}}_{seg}, \mathcal{M}_{seg})$$

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Contents

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Experiments

Our method consistently surpasses existing approaches and achieves SoTA performance on both grounding and segmentation tasks (Tab. 1, 2, 3) on RefCOCO/+/g datasets, providing valuable insights for future research.

In the single-dataset fine-tuning setting on REC task, our base model surpasses the current SoTA method HiVG by 2.07%(testB), 6.15%(testB), 4.73%(test), 1.95%(test), and 1.50%(test) on the five datasets respectively.

**Tab.1 Comparison with SoTA methods on REC task (Acc@0.5 metric).**

| Methods | Venue | Visual Backbone | Language Backbone | RefCOCO | | | RefCOCO+ | | | RefCOCOg | | ReferIt | Flickr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | val | testA | testB | val | testA | testB | val | test | test | test |
| **Single-dataset fine-tuning setting *w.* uni-modal pre-trained close-set detector and language model: (traditional setting)** | | | | | | | | | | | | | |
| TransVG [14] | ICCV'21 | RN101+DETR | BERT-B | 81.02 | 82.72 | 78.35 | 64.82 | 70.70 | 56.94 | 68.67 | 67.73 | 70.73 | 79.10 |
| Word2Pix [103] | TNNLS'22 | RN101+DETR | BERT-B | 81.20 | 84.39 | 78.12 | 69.74 | 76.11 | 61.24 | 70.81 | 71.34 | – | – |
| QRNet [98] | CVPR'22 | Swin-S [60] | BERT-B | 84.01 | 85.85 | 82.34 | 72.94 | 76.17 | 63.81 | 71.89 | 73.03 | 74.61 | 81.95 |
| VG-LAW [79] | CVPR'23 | ViT-Det [46] | BERT-B | 86.06 | 88.56 | 82.87 | 75.74 | 80.32 | 66.69 | 75.31 | 75.95 | 76.60 | – |
| TransVG++[15] | TPAMI'23 | ViT-Det [46] | BERT-B | 86.28 | 88.37 | 80.97 | 75.39 | 80.45 | 66.28 | 76.18 | 76.30 | 74.70 | 81.49 |
| **Single-dataset fine-tuning setting *w.* vision-language self-supervised pre-trained model:** | | | | | | | | | | | | | |
| CLIP-VG [91] | TMM'23 | CLIP-B | CLIP-B | 84.29 | 87.76 | 78.43 | 69.55 | 77.33 | 57.62 | 73.18 | 72.54 | 70.89 | 81.99 |
| JMRI [108] | TIM'23 | CLIP-B | CLIP-B | 82.97 | 87.30 | 74.62 | 71.17 | 79.82 | 57.01 | 71.96 | 72.04 | 68.23 | 79.90 |
| Dynamic-MDETR | TPAMI'23 | CLIP-B | CLIP-B | 85.97 | 88.82 | 80.12 | 74.83 | 81.70 | 63.44 | 74.14 | 74.49 | 70.37 | 81.89 |
| HiVG-B [92] | ACMMM'24 | CLIP-B | CLIP-B | 87.32 | 89.86 | 83.27 | 78.06 | 83.81 | 68.11 | 78.29 | 78.79 | 75.22 | 82.11 |
| HiVG-L [92] | ACMMM'24 | CLIP-L | CLIP-L | 88.14 | 91.09 | 83.71 | 80.10 | 86.77 | 70.53 | 80.78 | 80.25 | 76.23 | 82.16 |
| **OneRef-B (ours)** | NeurIPS'24 | BEiT3-B | BEiT3-B | 88.75 | 90.95 | 85.34 | 80.43 | 86.46 | 74.26 | 83.68 | 83.52 | 77.17 | 83.61 |
| **OneRef-L (ours)** | NeurIPS'24 | BEiT3-L | BEiT3-L | **92.87** | **94.01** | **90.19** | **87.98** | **91.57** | **83.73** | **88.11** | **89.29** | **81.11** | **84.75** |

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Experiments

In dataset-mixed pre-training setting on REC task, our base model outperforms HiVG by 1.35%, 2.79%, and 2.63% on RefCOCO/+/g testB/testB/test splits, outperforms Grounding-DINO by 2.59%, 4.76%, and 2.38%.

**Tab.2 Comparison with SoTA methods on REC task (Acc@0.5 metric).**

| Methods | Venue | Visual/Language Backbone | Intermediate pretrain data | Data size | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | val | testA | testB | val | testA | testB | val | test |
| **Dataset-mixed intermediate pre-training setting (*w.* box-level dataset-mixed open-set detection pre-trained model)** | | | | | | | | | | | | |
| MDETR † [33] | ICCV'21 | RN101/RoBERT-B | GoldG,RefC | 6.5M | 86.75 | 89.58 | 81.41 | 79.52 | 84.09 | 70.62 | 81.64 | 80.89 |
| YORO† [29] | ECCV'22 | ViLT [37] / BERT-B | GoldG,RefC | 6.5M | 82.90 | 85.60 | 77.40 | 73.50 | 78.60 | 64.90 | 73.40 | 74.30 |
| DQ-DETR † [54] | AAAI'23 | RN101 / BERT-B | GoldG,RefC | 6.5M | 88.63 | 91.04 | 83.51 | 81.66 | 86.15 | 73.21 | 82.76 | 83.44 |
| Grounding-DINO-B† | arXiv'23 | Swin-T / BERT-B | O365,GoldG,RefC | 7.2M | 89.19 | 91.86 | 85.99 | 81.09 | 87.40 | 74.71 | 84.15 | 84.94 |
| Grounding-DINO-L† | arXiv'23 | Swin-L / BERT-B | G-DINO-L* | 21.4M | 90.56 | 93.19 | 88.24 | 82.75 | 88.95 | 75.92 | 86.13 | 87.02 |
| CyCo [84] | AAAI'24 | ViT[19]/ BERT-B | VG,SBU,CC3M,*etc.* | >120M | 89.47 | 91.87 | 85.33 | 80.40 | 87.07 | 69.87 | 81.31 | 81.04 |
| HiVG-B† [92] | ACMMM'24 | CLIP-B / CLIP-B | RefC,ReferIt,Flickr | 0.8M | 90.56 | 92.55 | 87.23 | 83.08 | 87.83 | 76.68 | 84.71 | 84.69 |
| HiVG-L† [92] | ACMMM'24 | CLIP-L / CLIP-L | RefC,ReferIt,Flickr | 0.8M | 91.37 | 93.64 | 88.03 | 83.63 | 88.16 | 77.37 | 86.73 | 86.86 |
| **Fine-tuning setting *w.* dataset-mixed multi-task mix-supervised pre-trained model:** | | | | | | | | | | | | |
| UniTAB † [97] | ECCV'22 | RN101/RoBERT-B | VG,COCO,*etc.* | >20M | 88.59 | 91.06 | 83.75 | 80.97 | 85.36 | 71.55 | 84.58 | 84.70 |
| OFA-B † [85] | ICML'22 | OFA-B / OFA-B | – | – | 88.48 | 90.67 | 83.30 | 81.39 | 87.15 | 74.29 | 82.29 | 82.31 |
| OFA-L † [85] | ICML'22 | OFA-L / OFA-L | – | – | 90.05 | 92.93 | 85.26 | 85.80 | 89.87 | 79.22 | 85.89 | 86.55 |
| **Fine-tuning setting *w.* grounding multimodal large language model (GMLLM):** | | | | | | | | | | | | |
| Shikra-7B† [10] | arXiv'23 | CLIP-L / Vicuna-7B[12] | RefC,VG | 0.5M | 87.01 | 90.61 | 80.24 | 81.60 | 87.36 | 72.12 | 82.27 | 82.19 |
| Ferret-7B† [100] | ICLR'24 | CLIP-L / Vicuna-7B[12] | GRIT [100] | >8M | 87.49 | 91.35 | 82.45 | 80.78 | 87.38 | 73.14 | 83.93 | 84.76 |
| LION-4B† [9] | CVPR'24 | EVA-G[21]/FlanT5-3B | VG,COCO,*etc.* | 3.6M | 89.73 | 92.29 | 84.82 | 83.60 | 88.72 | 77.34 | 85.69 | 85.63 |
| LION-12B† [9] | CVPR'24 | EVA-G[21]/FlanT5-11B | VG,COCO,*etc.* | 3.6M | 89.80 | 93.02 | 85.57 | 83.95 | 89.22 | 78.06 | 85.52 | 85.74 |
| **OneRef-B† (unsupervised)** | | BEiT3-B / BEiT3-B | RefC,ReferIt | 0.5M | 89.16 | 92.03 | 87.26 | 83.18 | 88.56 | 77.66 | 84.72 | 85.17 |
| **OneRef-B† (0.2B)** | NeurIPS'24 | BEiT3-B / BEiT3-B | RefC,ReferIt | 0.5M | 91.89 | 94.31 | 88.58 | 86.38 | 90.38 | 79.47 | 86.82 | 87.32 |
| **OneRef-L† (0.6B)** | NeurIPS'24 | BEiT3-L / BEiT3-L | RefC,ReferIt | 0.5M | **93.21** | **95.43** | **90.11** | **88.35** | **92.11** | **82.70** | **87.81** | **88.83** |

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Experiments

On RES task, in the single-dataset fine-tuning setting, our base model surpasses the SoTA VLP-based method RISCLIP by 2.65%, 4.77%, and 1.73% on RefCOCO/+/g. In the dataset-mixed pre-training setting, our base model also achieves superior performance compared to RISCLIP with improvements of 4.53%, 8.21%, and 5.39%.

**Tab.3 Comparison with SoTA methods on RES task (mIOU metric).**

| Methods | Venue | Visual/Language Backbone | Intermediate pretrain data | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | val | testA | testB | val | testA | testB | val | test |
| **Single-dataset fine-tuning setting *w.* uni-modal pre-trained close-set segmentation model: (traditional setting)** | | | | | | | | | | | |
| RefTR [45] | NIPS'21 | RN101 / BERT-B | – | 74.34 | 76.77 | 70.87 | 66.75 | 70.58 | 59.40 | 66.63 | 67.39 |
| SeqTR [107] | ECCV'22 | DN53[72]/Bi-GRU | – | 71.70 | 73.31 | 69.82 | 63.04 | 66.73 | 58.97 | 64.69 | 65.74 |
| LAVT [94] | CVPR'22 | Swin-B / BERT-B | – | 74.46 | 76.89 | 70.94 | 65.81 | 70.97 | 59.23 | 63.34 | 63.62 |
| VG-LAW [79] | CVPR'23 | ViT-Det / BERT-B | – | 75.05 | 77.36 | 71.69 | 66.61 | 70.30 | 58.14 | 65.36 | 65.13 |
| **Single-dataset fine-tuning setting *w.* vision-language self-supervised pre-trained model:** | | | | | | | | | | | |
| CRIS [89] | CVPR'22 | CLIP-L / CLIP-L | – | 70.47 | 73.18 | 66.10 | 62.27 | 68.06 | 53.68 | 59.87 | 60.36 |
| JMCELN[32] | EMNLP'23 | CLIP-B / CLIP-B | – | 74.40 | 77.69 | 70.43 | 66.99 | 72.69 | 57.34 | 64.08 | 64.99 |
| RISCLIP-B [35] | NAACL'24 | CLIP-B / CLIP-B | – | 75.68 | 78.01 | 72.46 | 69.16 | 73.53 | 60.68 | 67.62 | 67.97 |
| RISCLIP-L [35] | NAACL'24 | CLIP-L / CLIP-L | – | 78.87 | 81.46 | 75.41 | 74.38 | 78.77 | 66.84 | 71.82 | 71.65 |
| **OneRef-B (ours)** | NeurIPS'24 | BEiT3-B / BEiT3-B | – | 77.57 | 79.05 | 75.11 | 71.25 | 75.41 | 65.45 | 69.37 | 69.70 |
| **OneRef-L (ours)** | NeurIPS'24 | BEiT3-L / BEiT3-L | – | **80.09** | **82.19** | **77.51** | **75.17** | **79.38** | **70.17** | **73.18** | **72.76** |
| **Dataset-mixed intermediate pre-training setting:** | | | | | | | | | | | |
| PolyFormer-B$^{\dagger}$[53] | CVPR'23 | Swin-B / BERT-B | RefC | 75.96 | 77.09 | 73.22 | 70.65 | 74.51 | 64.64 | 69.36 | 69.88 |
| RISCLIP-B$^{\dagger}$[35] | NAACL'24 | CLIP-B / CLIP-B | RefC | 75.68 | 78.01 | 72.46 | 72.46 | 74.30 | 61.37 | 69.49 | 69.53 |
| RISCLIP-L$^{\dagger}$[35] | NAACL'24 | CLIP-L / CLIP-L | RefC | 79.53 | 81.78 | 75.78 | 74.88 | 78.88 | 68.09 | 73.45 | 74.52 |
| **OneRef-B$^{\dagger}$ (unsupervised)** | | BEiT3-B / BEiT3-B | RefC | 78.20 | 79.26 | 75.92 | 72.54 | 75.54 | 67.39 | 71.28 | 71.13 |
| **OneRef-B$^{\dagger}$ (ours)** | NeurIPS'24 | BEiT3-B / BEiT3-B | RefC | 79.83 | 81.86 | 76.99 | 74.68 | 77.90 | 69.58 | 74.06 | 74.92 |
| **OneRef-L$^{\dagger}$ (ours)** | NeurIPS'24 | BEiT3-L / BEiT3-L | RefC | **81.26** | **83.06** | **79.45** | **76.60** | **80.16** | **72.95** | **75.68** | **76.82** |

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Experiments

We present the qualitative grounding and referring segmentation results with several relatively challenging examples. These results demonstrate the strong semantic comprehension capability of our OneRef model in complex text understanding and cross-modal grounding. Our approach providing valuable insights for future research.

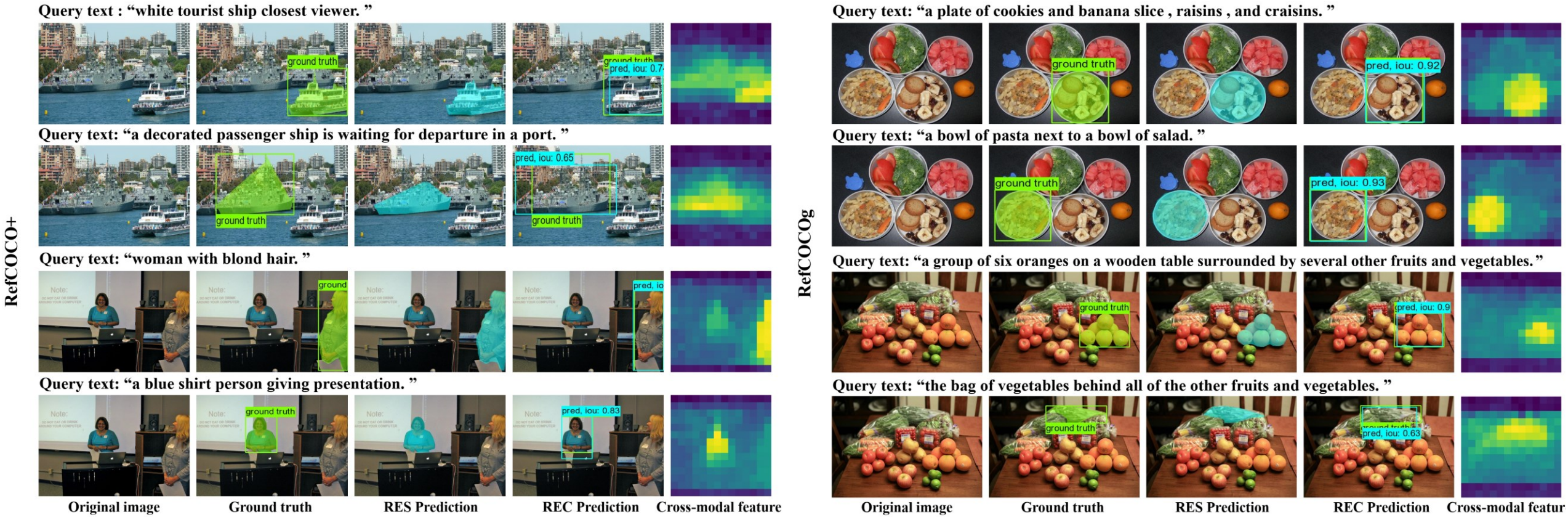



**Fig.9 Part of the qualitative results in REC and RES tasks.**

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Contents

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

# Conclusion

**Contributions:**

- (i) We pioneer the application of mask modeling to referring tasks by introducing a novel paradigm called mask referring modeling. This paradigm effectively models the referential relation between visual and language.
- (ii) Diverging from previous works, we propose a remarkably concise one-tower framework for grounding and referring segmentation in a unified modality-shared feature space. Our model eliminates the commonly used modality interaction modules, modality fusion en-/decoders, and special grounding tokens.
- (iii) We extensively validate the effectiveness of MRefM in three referring tasks on five datasets. Our method consistently surpasses existing approaches and achieves SoTA performance across several settings, providing a valuable new insights for future grounding and referring segmentation research.

**In a word**, we propose a novel, highly concise, and feature space unified one-tower referring framework. Additionally, we pioneer the exploration of mask modeling in referring tasks by introducing MRefM paradigm. MRefM enables potential large-scale pre-training of grounding in future, presenting a new direction for referring tasks.

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID: 8873

Paper Code Contact

# That's all.
# Thank you!