Enriching Phrases with Coupled Pixel and Object Contexts for Panoptic Narrative Grounding

Tianrui Hui[1,2,5] , **Zihan Ding**[3,5] , **Junshi Huang**[5*] , **Xiaoming Wei**[5] , **Xiaolin Wei**[5] ,
Jiao Dai[1,2] , **Jizhong Han**[1,2] and **Si Liu**[3,4]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Institute of Artificial Intelligence, Beihang University
[4]Hangzhou Innovation Institute, Beihang University
[5]Meituan

Abstract

Panoptic narrative grounding (PNG) aims to segment things and stuff objects in an image described by noun phrases of a narrative caption. As a multimodal task, an essential aspect of PNG is the visual-linguistic interaction between image and caption. The previous two-stage method aggregates visual contexts from offline-generated mask proposals to phrase features, which tend to be noisy and fragmentary. The recent one-stage method aggregates only pixel contexts from image features to phrase features, which may incur semantic misalignment due to lacking object priors. To realize more comprehensive visual-linguistic interaction, we propose to enrich phrases with coupled pixel and object contexts by designing a Phrase-Pixel-Object Transformer Decoder (PPO-TD), where both fine-grained part details and coarse-grained entity clues are aggregated to phrase features. In addition, we also propose a Phrase-Object Contrastive Loss (POCL) to pull closer the matched phrase-object pairs and push away unmatched ones for aggregating more precise object contexts from more phrase-relevant object tokens. Extensive experiments on the PNG benchmark show our method achieves new state-of-the-art performance with large margins.

1 Introduction

Panoptic narrative grounding (PNG) [González *et al.*, 2021] is an emerging task whose goal is to predict the panoptic segmentation of things and stuff objects described by noun phrases in a narrative caption. Compared with relevant tasks like referring image segmentation [Yang *et al.*, 2022] and phrase grounding [Mu *et al.*, 2021], PNG not only extends the grounding granularity from sentences to phrases but also yields finer results of segmentation masks rather than bounding boxes, facilitating a more comprehensive visual-linguistic alignment. Benefiting from these properties, PNG enjoys a
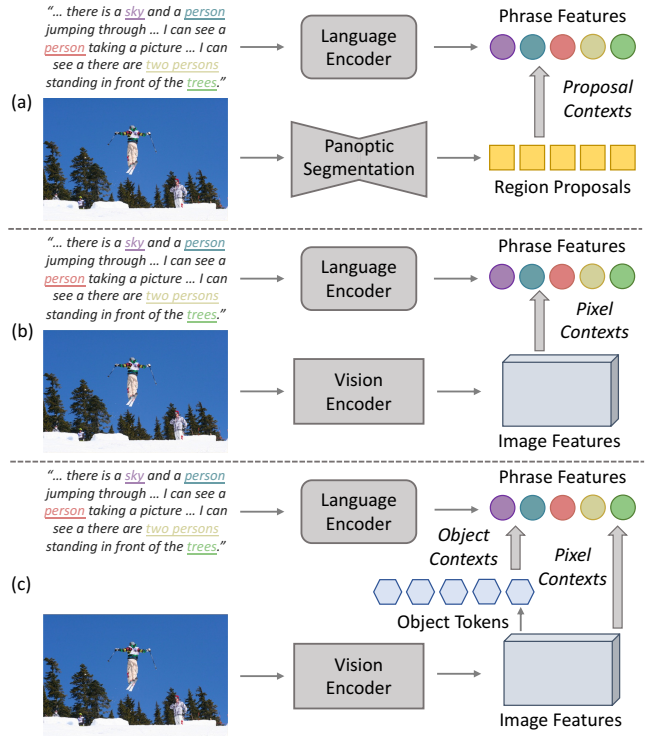
*Corresponding author

Figure 1: Comparison of visual-linguistic interaction schemes between previous methods and ours. (a) The previous two-stage method aggregates proposal contexts from offline-generated mask region proposals, which tend to be fragmentary and noisy. (b) The previous one-stage method aggregates pixel contexts directly from image features, but lacking object-level contexts inclines to incur semantic misalignment between phrases and pixels. (c) Our method proposes to enrich phrases with coupled pixel and object contexts containing both fine-grained part details and coarse-grained entity clues, forming more comprehensive visual-linguistic interaction.

wide range of potential applications such as image editing, intelligent robots, and human-computer interaction.

Considering the multimodal nature of the PNG task, one of the essential components of PNG methods is the visual-linguistic interaction between image and caption. As shown

in Figure 1 (a), the first work [González *et al.*, 2021], which proposes the PNG task, also designs a two-stage baseline method. It first leverages an off-the-shelf panoptic segmentation model [Kirillov *et al.*, 2019a] to generate mask region proposals for the input image, then performs matching between features of phrases and proposals. Herein, the visual-linguistic interaction is realized by cross-attention [Vaswani *et al.*, 2017] between phrases and region proposals, where proposal contexts are aggregated to the phrase features for better matching in the second stage. However, the region proposals are usually inaccurate segments, so the aggregated proposal contexts may provide fragmentary and noisy visual object information to the phrase features. Recently, PPMN [Ding *et al.*, 2022] proposes a one-stage method that lets each phrase directly find its matched pixels to avoid the offline proposal generation process. For visual-linguistic interaction, the one-stage method in Figure 1 (b) aggregate pixel-level visual contexts directly from the grid-form image features to the phrase features. Though pixel contexts contain fine-grained visual details, the absence of object-level visual contexts inclines to incur semantic misalignment between abstract noun phrases and concrete image pixels, yielding suboptimal object delineation.

Given the above discussion, the visual-linguistic interaction schemes of previous methods still limit the discriminative ability of phrase features due to insufficient visual contexts, albeit their strong performances. Therefore, in this paper, we propose to enrich phrases with coupled pixel and object contexts for more comprehensive visual-linguistic interaction as illustrated in Figure 1 (c). On the one hand, pixel contexts aggregated directly from image features contain more fine-grained part details, which can assist phrases in predicting finer segmentation masks. On the other hand, object contexts, which are also derived from image features, focus more on coarse-grained entity perception and can facilitate the distinguishment of corresponding referred objects for different phrases. By coupling the pixel and object contexts, the discriminative ability of phrases could be further enhanced to excavate and match with more relevant visual objects. Concretely, our method first extracts image and phrase features by vision and language encoders. Then, phrases and a set of learnable object tokens are concatenated and fed into our proposed Phrase-Pixel-Object Transformer Decoder (PPO-TD) as queries, in which multi-scale image features from a pixel decoder serve as keys and values. Through consecutive masked cross-attention layers, phrases aggregate pixel contexts from image features to obtain fine-grained part details. Meanwhile, the representations of object tokens are progressively learned to perceive coarse-grained entity information as well. With interlaced self-attention layers in our PPO-TD, phrases are also enriched with object contexts by aggregating information from object tokens to further enhance their discriminative ability. Finally, the inner product between phrase and image features is conducted to generate a binary mask for each phrase based on the pixel-phrase matching similarities.

In order to provide more precise object contexts for phrases, we also design a Phrase-Object Contrastive Loss (POCL) to refine the learning process of object tokens repre-sentations. The Hungarian algorithm is first performed to obtain the bipartite matching between object tokens and ground-truth masks/classes. Since the ground-truth masks/classes corresponding to each phrase are also available in the PNG annotations, we can find which object tokens are matched with each phrase. Utilizing these matching relations as the supervision, we devise a contrastive loss on phrases and object tokens to increase the feature similarities between matched phrase-object pairs and decrease those between unmatched pairs. Constrained by this loss, phrase features can aggregate object contexts from more relevant object tokens to enhance their discriminative ability.

Our contributions are summarized as follows:

- We propose to enrich phrases with coupled pixel and object contexts by a Phrase-Pixel-Object Transformer Decoder (PPO-TD), where both fine-grained part details and coarse-grained entity clues are aggregated to phrase features for comprehensive visual-linguistic interaction.

- We also design a Phrase-Object Contrastive Loss (POCL) to increase feature similarities between matched phrase-object pairs and decrease those between unmatched pairs, thus providing more precise object contexts for phrases.

- Extensive experiments on the PNG benchmark demonstrate our method achieves new state-of-the-art performance with large margins.

2 Related Work

2.1 Referring Image Segmentation

The goal of referring image segmentation is to segment the object referred by the subject of an input sentence. Previous mainstream methods [Ye *et al.*, 2019; Huang *et al.*, 2020; Hui *et al.*, 2020; Liu *et al.*, 2021; Zhou *et al.*, 2021; Tan *et al.*, 2020; Feng *et al.*, 2021; Jing *et al.*, 2021] are based on FCN [Long *et al.*, 2015] architecture and design various cross-modal attention mechanisms to fuse image and sentence features. With the recent success of Transformer [Vaswani *et al.*, 2017], LAVT [Yang *et al.*, 2022] and ReSTR [Kim *et al.*, 2022] employ Vision Transformer [Wang *et al.*, 2022] to directly encode and fuse image and sentence features with additional channel attention or class seed embedding. CRIS [Wang *et al.*, 2022] transfers large-scale vision-language pretrained knowledge of CLIP for text-to-pixel alignment. PNG extends the grounding granularity of referring image segmentation from sentences to phrases, which could achieve fine-grained vision-language alignment. In this paper, we propose to enrich phrases with coupled pixel and object contexts for more comprehensive visual-linguistic interaction. The discussion of existing PNG methods has been included in Sec. 1, so we omit it here.

2.2 Panoptic Segmentation

Panoptic segmentation task is first proposed by [Kirillov *et al.*, 2019b], whose aim is to unify semantic segmentation and instance segmentation tasks for a coherent scene segmentation of things and stuff classes. Previous methods [Kirillov *et al.*, 2019a; Cheng *et al.*, 2020; Wang *et al.*, 2020;
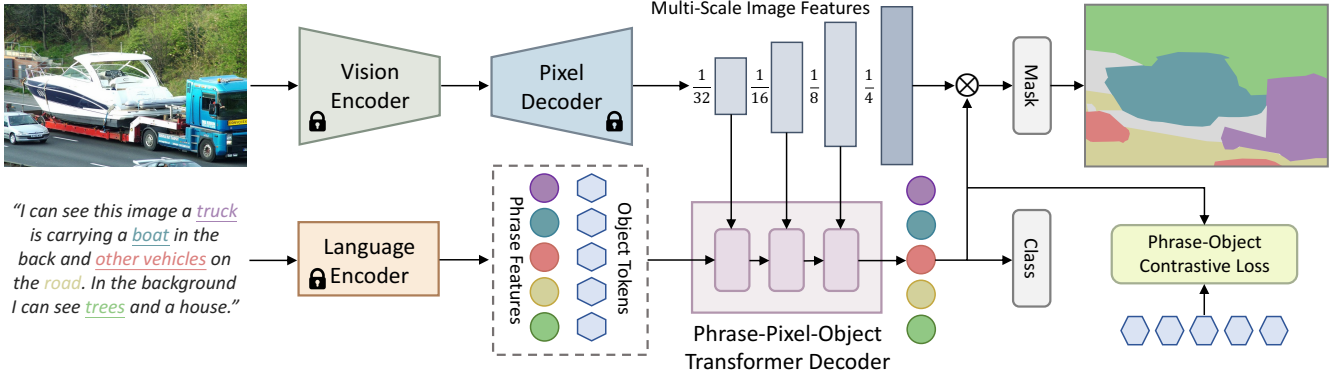
Figure 2: Overview of our method. Image and phrase features are extracted by vision and language encoders. A pixel decoder further refines the multi-scale image features. Our proposed Phrase-Pixel-Object Transformer Decoder takes the concatenation of phrase features and learnable object tokens as queries and multi-scale image features as keys and values, where phrases are enriched with coupled pixel and object contexts to form a more comprehensive visual-linguistic interaction. A Phrase-Object Contrastive Loss is also proposed to increase feature similarities between matched phrase-object pairs and decrease those between unmatched ones so that more phrase-relevant object contexts are aggregated. The final mask predictions are obtained by the inner product between phrase and image features. The parameters of the language encoder, vision encoder, and pixel decoder are fixed during training.

Mohan and Valada, 2021] mainly focus on designing specialized branches to tackle semantic and instance segmentation in a single network. Recently, some Transformer-based methods [Cheng *et al.*, 2021; Zhang *et al.*, 2021; Cheng *et al.*, 2022] propose the mask classification paradigm to replace traditional per-pixel classification for more effective panoptic segmentation. For example, MaskFormer [Cheng *et al.*, 2021] exploits object queries to generate binary masks and their classifications simultaneously. Different from panoptic segmentation, PNG segments objects according to narrative captions for a comprehensive cross-modal understanding.

2.3 Phrase Grounding

Phrase grounding requires the model to localize objects referred by noun phrases with bounding boxes. Some early works [Wang *et al.*, 2018; Plummer *et al.*, 2018; Akbari *et al.*, 2019] first extract features of region proposals and phrases separately, then perform phrase-region matching in common feature space with specially designed loss functions. The following methods [Bajaj *et al.*, 2019; Mu *et al.*, 2021; Yu *et al.*, 2020; Dogan *et al.*, 2019] focus on designing effective interaction approaches between visual and linguistic features to excavate multimodal context for more accurate phrase-region matching. The recent GLIP model [Li *et al.*, 2022a] unifies object detection and phrase grounding into vision-language pretraining, which further boosts the grounding performance. The PNG task also extends phrase grounding from region-level localization to pixel-level segmentation, which facilitates accurate object perception.

3 Method

The overall framework of our method is shown in Figure 2, which follows the meta-architecture of Mask2Former [Cheng *et al.*, 2022]. The input image and caption are first encoded by the vision and language encoders to obtain the image and phrase features. Then, the pixel decoder extracts multi-scale

image features after the vision encoder. Our proposed Phrase-Pixel-Object Transformer Decoder (PPO-TD) takes the concatenation of phrase features and learnable object tokens as queries and multi-scale image features as keys and values. Through successive layers in PPO-TD, phrases are enriched with coupled pixel and object contexts to form more comprehensive visual-linguistic interaction. The inner product between phrase features and image features generates the final binary mask for each phrase. Moreover, we also design a contrastive loss on phrase-object pairs to aggregate more precise object contexts for phrases.

3.1 Multimodal Encoders and Pixel Decoder

For the input image $\mathcal{I} \in \mathbb{R}^{H^0 \times W^0 \times 3}$, we adopt ResNet-50 [He *et al.*, 2016] as the vision encoder to extract image features $\boldsymbol{F}^i \in \mathbb{R}^{H^i \times W^i \times C_f^i}, i \in \{2, 3, 4, 5\}$, where $H^i = \frac{H^0}{2^i}, W^i = \frac{W^0}{2^i}, C_f^i$ denote the height, width, and channel numbers of the i-th level of image features from the ResNet backbone. For the input narrative caption with N noun phrases, we adopt the BERT-base model [Devlin *et al.*, 2018] to extract phrase feature embeddings $\boldsymbol{R} \in \mathbb{R}^{N \times C_r}$, where C_r denotes the channel number of phrase features.

Following [Cheng *et al.*, 2022], we exploit multi-scale deformable attention Transformer (MSDeformAttn) [Zhu *et al.*, 2020] as the pixel decoder to fuse and refine multi-scale image features. The pixel decoder takes $\{\boldsymbol{F}^i\}_{i=3}^5$ as inputs and first projects the channel numbers of these image features into the same value C_h with 1×1 convolution layers. Then, the projected image features, their generated sinusoids positional embedding [Vaswani *et al.*, 2017] $\{\boldsymbol{P}^i\}_{i=3}^5 \in \mathbb{R}^{H^i \times W^i \times C_h}$, and learnable scale-level embedding $\boldsymbol{S} \in \mathbb{R}^{3 \times C_h}$ are fed into the pixel decoder for multi-scale feature refinement:

$$\{\bar{\boldsymbol{F}}^i\}_{i=3}^5 = \text{MSDeformAttn}(\{\boldsymbol{F}^i\}_{i=3}^5, \{\boldsymbol{P}^i\}_{i=3}^5, \boldsymbol{S}). \quad (1)$$

Within the pixel decoder, each pixel on the image features samples a small set of positions from features of all scales us-

ing learned offsets and aggregates their features using learned attention weights. $\{\bar{\boldsymbol{F}}^i\}_{i=3}^5$ denote image features after multiple deformable attention layers, which fuse rich multi-scale visual information. The resolutions of $\{\bar{\boldsymbol{F}}^i\}_{i=3}^5$ are $1/8$, $1/16$, and $1/32$ of the original image resolution, respectively. In addition, a bilinear interpolation layer is applied on $\bar{\boldsymbol{F}}^3$ to upsample its resolution by $2\times$, and the upsampled feature is summed and fused with input backbone feature $\boldsymbol{F}^2$ using lateral convolution layers:

$$\bar{\boldsymbol{F}}^2 = \text{Conv}_{3\times3}(\text{Conv}_{1\times1}(\boldsymbol{F}^2) + \text{Upsample}(\bar{\boldsymbol{F}}^3)) \quad (2)$$

The obtained feature $\bar{\boldsymbol{F}}^2$ has $1/4$ scale of the original image and serves as the per-pixel embedding to generate binary masks for phrases and object tokens.

3.2 Phrase-Pixel-Object Transformer Decoder

In order to enhance the discriminative ability of phrase features for distinguishing referred objects and predicting accurate segmentation masks, we propose a Phrase-Pixel-Object Transformer Decoder (PPO-TD) to enrich phrases with coupled pixel and object contexts for more comprehensive visual-linguistic interaction.

Each layer of our PPO-TD is mainly composed of cross-attention, self-attention, and feedforward network layers in a sequential manner. Given the effectiveness and popularity of cross-attention, previous PNG methods also leverage it to conduct visual-linguistic interaction. They utilize phrase features alone as the query and mask proposal features or image features as the key and value to aggregate proposal contexts only [González *et al.*, 2021] or pixel contexts only [Ding *et al.*, 2022] to the phrases. However, proposal contexts may contain fragmentary and noisy visual object information due to the inaccurate segments predicted by the off-the-shelf panoptic segmentation model. Meanwhile, only pixel contexts also tend to incur semantic inconsistency between abstract noun phrases and concrete image pixels since lacking object-level visual contexts may yield sub-optimal object delineation.

Therefore, we argue that both pixel contexts and object contexts are beneficial to enhancing the discriminative ability of phrase features. To introduce proper object-level visual contexts, we take the concatenation of phrase features and a set of learnable object tokens as the query of cross-attention instead of using phrase features alone. For key and value, we adopt multi-scale image features from the pixel decoder.

Take the l-th layer of our PPO-TD as an example. The N phrase features $\boldsymbol{R}_{l-1} \in \mathbb{R}^{N \times C_h}$ and M object tokens $\boldsymbol{O}_{l-1} \in \mathbb{R}^{M \times C_h}$ from the $(l-1)$-th layer are first summed with their learnable positional embeddings, and then concatenated and projected by a linear transformation to obtain the input query feature $\boldsymbol{Q}_l \in \mathbb{R}^{(N+M) \times C_h}$ of cross-attention layer:

$$\boldsymbol{Q}_l = [\boldsymbol{R}_{l-1} + \boldsymbol{P}_r; \boldsymbol{O}_{l-1} + \boldsymbol{P}_o]\boldsymbol{W}_q, \quad (3)$$

where $[;]$ is concatenation and $\boldsymbol{W}_q$ is linear transformation parameter. The key and value features are similarly obtained:

$$\boldsymbol{K}_l = (\bar{\boldsymbol{F}}_l + \boldsymbol{P}_l + \boldsymbol{S}_l)\boldsymbol{W}_k, \quad \boldsymbol{V}_l = (\bar{\boldsymbol{F}}_l + \boldsymbol{P}_l + \boldsymbol{S}_l)\boldsymbol{W}_v, \quad (4)$$

where $\bar{\boldsymbol{F}}_l \in \mathbb{R}^{H^l W^l \times C_h}$ is one of the flattened multi-scale image features $\{\bar{\boldsymbol{F}}^i\}_{i=3}^5$ from the pixel decoder, $\boldsymbol{P}_l$ and $\boldsymbol{S}_l$ are

learnable positional embedding and scale-level embedding. Following [Cheng *et al.*, 2022], we exploit masked cross-attention to constrain the attention scope of phrase features and object tokens as follows:

$$\boldsymbol{X}_l = \text{Softmax}(\boldsymbol{Q}_l \boldsymbol{K}_l^{\text{T}} + \boldsymbol{A}_{l-1})\boldsymbol{V}_l + \boldsymbol{X}_{l-1}, \quad (5)$$

where $\boldsymbol{X}_{l-1}, \boldsymbol{X}_l \in \mathbb{R}^{(N+M) \times C_h}$ are concatenated phrase features and object tokens in adjacent layers. $\boldsymbol{A}_{l-1}$ is the attention mask and its value on position (x, y) is:

$$\boldsymbol{A}_{l-1}(x, y) = \begin{cases} 0, & \text{if } \boldsymbol{M}_{l-1}(x, y) = 1 \\ -\infty, & \text{otherwise,} \end{cases} \quad (6)$$

where $\boldsymbol{M}_{l-1} = \text{Resize}(\boldsymbol{X}_{l-1}(\bar{\boldsymbol{F}}^2)^{\text{T}}) \in \mathbb{R}^{(N+M) \times H^l W^l}$ is the predicted binary masks of phrases and object tokens after normalization and binarization at threshold 0.5 in the $l-1$-th layer of PPO-TD. The attention mask constrains the cross-attention scope to predicted foreground regions, which facilitates the aggregation of more phrase-relevant pixel contexts to the phrase features, and also refines the representations of object tokens by suppressing other distractors.

After the cross-attention layer, concatenated phrase features and object tokens are fed into the self-attention layer to communicate information with each other. By this means, phrase features can also aggregate object contexts from object tokens whose representations are adaptively extracted from multi-scale image features, while object tokens incorporate language priors to distinguish referred objects better. With coupled pixel and object contexts, the discriminative ability of phrase features is further enhanced to form a more comprehensive visual-linguistic interaction.

3.3 Phrase-Object Contrastive Loss

After L layers of PPO-TD, we take the output phrase features $\boldsymbol{R}_L$ to generate final mask prediction $\boldsymbol{M}_L \in \mathbb{R}^{N \times H^0 W^0}$ by inner product with the flattened per-pixel embedding $\bar{\boldsymbol{F}}^2$ followed by upsampling operation. Since each phrase is also associated with a ground-truth class to which the referred object belongs, we also use a linear layer to generate final class prediction $\boldsymbol{C}_L \in \mathbb{R}^{N \times K}$ to leverage class priors contained in phrases, where K is the number of classes. Then, the mask classification loss of phrases $\mathcal{L}_{\text{mask-cls}}^{\text{phr}}$, which consists of a cross-entropy classification loss and a binary cross-entropy mask loss, is formulated as follows:

$$\mathcal{L}_{\text{mask-cls}}^{\text{phr}} = \sum_{j=1}^{N} \left[-\log p(c_j^*) + \mathcal{L}_{\text{mask}}(\boldsymbol{m}_j, \boldsymbol{m}_j^*) \right], \quad (7)$$

where c_j^* is the ground-truth class of the j-th phrase, $m_j, m_j^* \in \mathbb{R}^{H^0 \times W^0}$ are the predicted and ground-truth masks of the j-th phrase respectively. For learnable object tokens, we optimize a bipartite matching [Cheng *et al.*, 2022] to find the matched ground-truth mask/class for each object token. Similar mask classification loss $\mathcal{L}_{\text{mask-cls}}^{\text{obj}}$ is also applied to object tokens, but the predicted masks of object tokens are not used as the final output results during inference.

Since phrases and object tokens are both matched with their own ground-truth masks/classes, we use these ground

truths as links to obtain the actual matching relations $G^* \in \mathbb{R}^{N \times M}$ between phrases and object tokens. Inner-product between features of phrases and object tokens is also performed to generate the matching predictions $G \in \mathbb{R}^{N \times M}$. Based on G^* and G, we design a Phrase-Object Contrastive Loss $\mathcal{L}_{\text{poc}}$ (POCL) to increase the feature similarities between matched phrase-object pairs and decrease those between unmatched pairs, which is implemented with binary cross-entropy loss:

$$
\begin{aligned}
\mathcal{L}_{\text{bce}}(G_{j,k}, G^*_{j,k}) = & - G^*_{j,k}\log(G_{j,k}) \\
& - (1 - G^*_{j,k})\log(1 - G_{j,k}),
\end{aligned} \tag{8}
$$

$$
\mathcal{L}_{\text{poc}}(G, G^*) = \frac{1}{NM} \sum_{j=1}^{N} \sum_{k=1}^{M} \mathcal{L}_{\text{bce}}(G_{j,k}, G^*_{j,k}). \tag{9}
$$

Supervised by $\mathcal{L}_{\text{poc}}$, more precise object contexts from more phrase-relevant object tokens can be aggregated to phrase features, which further enhances their discriminative ability. Our POCL is inspired by previous works [He *et al.*, 2020; Pang *et al.*, 2021; Wu *et al.*, 2022; Li *et al.*, 2022b] which leverage similar loss functions for supervised contrastive learning, but POCL contains intuitive designs tailored for the PNG task where richer contexts are naturally coupled.

Also noteworthy is that Eq. 9 can be rewritten as follows:

$$
\begin{aligned}
\mathcal{L}_{\text{poc}}(G, G^*) &= \frac{1}{N} \sum_{j=1}^{N} \frac{1}{M} \sum_{k=1}^{M} \mathcal{L}_{\text{bce}}(G_{j,k}, G^*_{j,k}) \\
&= \frac{1}{N} \sum_{j=1}^{N} \mathcal{L}_{\text{multi-cls}}(G_{j,:}, G^*_{j,:}) \\
&= \frac{1}{M} \sum_{k=1}^{M} \mathcal{L}_{\text{multi-cls}}(G_{:,k}, G^*_{:,k}).
\end{aligned} \tag{10}
$$

$\mathcal{L}_{\text{poc}}$ is equivalent to bi-directional multi-label classification between phrases and object tokens, which means each phrase can be classified into multiple object token labels (plural noun) and each object token can also be classified into multiple phrase labels (same object referred by different phrases). By this means, positive phrase-object pairs obtain high classification probabilities, while negative ones do the opposite. The final loss of our model is the combination of three losses:

$$
\mathcal{L} = \mathcal{L}_{\text{mask-cls}}^{\text{phr}} + \mathcal{L}_{\text{mask-cls}}^{\text{obj}} + \mathcal{L}_{\text{poc}}. \tag{11}
$$

4 Experiments

4.1 Dataset and Evaluation Metrics

We conduct experiments on the Panoptic Narrative Grounding (PNG) benchmark [González *et al.*, 2021]. PNG is constructed by transferring caption annotations of Localized Narratives dataset [Pont-Tuset *et al.*, 2020] to the panoptic segmentation annotations of MSCOCO dataset [Lin *et al.*, 2014]. For the linguistic domain, each narrative in PNG contains an average of 11.3 noun phrases, of which 5.1 noun phrases require to be grounded. The total number of noun phrases is $726,445$. For the visual domain, a total of $741,697$ segments ($659,298$ are unique) in the panoptic segmentation annotations of MSCOCO are matched with narrative captions.

Following prior works [González *et al.*, 2021; Ding *et al.*, 2022], we adopt Average Recall as the metric to evaluate our method. For each phrase, different Intersection over Union (IoU) thresholds between segmentation prediction and ground truth are considered to determine whether the prediction is a true positive or not. A curve of recall at different IoU thresholds is obtained to calculate the Average Recall, which is the area under the curve. For each plural noun phrase, all ground-truth instances are aggregated into a single segmentation to calculate IoU with similarly aggregated prediction.

4.2 Implementation Details

Our model is implemented with PyTorch. The vision encoder and pixel decoder are pretrained on MSCOCO panoptic segmentation task for a fair comparison with prior works [González *et al.*, 2021; Ding *et al.*, 2022]. The projected channel number C_h in the pixel decoder and our PPO-TD is 256. The input image is resized to 1024×1024. The language encoder is implemented with the base version of BERT [Devlin *et al.*, 2018] whose embedding dimension C_r is 768. The input caption has a maximum length of 230 words, among which at most $N = 30$ noun phrases are required to be grounded. The number of learnable object tokens $M = 100$. Our PPO-TD has $L = 9$ layers, and the mask classification losses for phrases and object tokens are applied before each layer as auxiliary losses. AdamW is used as the optimizer with a learning rate of $1e^{-4}$. We train our model with batch size 16 for $145K$ iterations on eight NVIDIA V100 GPUs. The parameters of the language encoder, vision encoder, and pixel decoder are fixed during training.

4.3 Comparison with State-of-the-Art Methods

We compare our method with previous state-of-the-art methods on the PNG benchmark. Since PNG [González *et al.*, 2021] is a pretty new task, only several previous methods could be compared. Table 1 summarizes the quantitative comparison results. It is observed that our method outperforms previous state-of-the-art ones by significant margins. Compared with the latest one-stage method PPMN [Ding *et al.*, 2022], our method achieves significant absolute performance gains of $6.7/6.8/8.2/9.1/4.3$ on Average Recall metric for *overall/things/stuff/singulars/plurals* splits, which demonstrates enriching phrases with coupled pixel and object contexts is superior to aggregating pixel contexts only due to the coarse-grained entity clues complemented by object tokens. Compared with the two-stage baseline method proposed by PNG, our performance boost becomes 10.7 on Average Recall for *overall* split, which shows the offline-generated mask proposals may introduce noisy and fragmentary object-level contexts while our learned object tokens can provide more precise object contexts to phrases. MCN [Luo *et al.*, 2020] tackles referring expression segmentation and comprehension with a collaborative network and its Average Recall is evaluated on only *things* split. Both our method and previous PNG methods largely outperform MCN, which is probably because sentence-mask/box alignment in referring tasks is insufficient to learn phrase-level segmentation well.

As illustrated in Figure 3, we also draw the recall curves of our methods and previous ones for detailed comparison.

(a) Overall performance   (b) Things and stuff categories   (c) Singulars and plurals
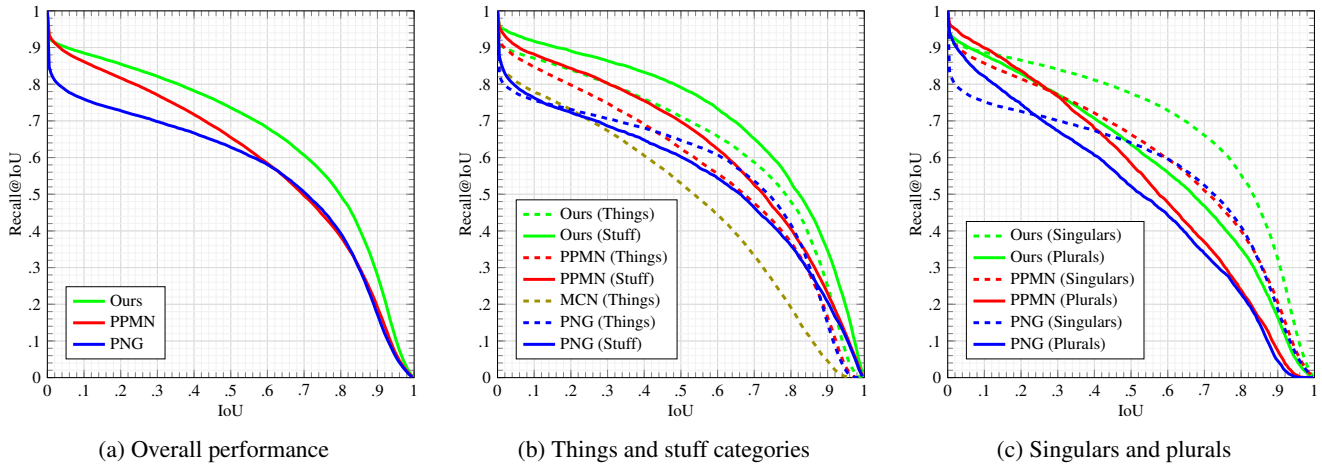
Figure 3: **Average Recall Curves** for our method performance (a) compared to the state-of-the-art methods, disaggregated into (b) things and stuff categories, and (c) singulars and plurals noun phrases.

(a) Things and stuff categories.

Method	Average Recall		
	overall	things	stuff
PNG [2021]	55.4	56.2	54.3
MCN [2020]	-	48.2	-
PPMN [2022]	59.4	57.2	62.5
Ours	**66.1(+6.7)**	**64.0(+6.8)**	**70.7(+8.2)**

(b) Singulars and plurals noun phrases.

Method	Average Recall		
	overall	singulars	plurals
PNG [2021]	55.4	56.2	48.8
PPMN [2022]	59.4	60.0	54.0
Ours	**66.1(+6.7)**	**69.1(+9.1)**	**58.3(+4.3)**

Table 1: Comparison with previous state-of-the-art methods on the PNG benchmark, disaggregated into (a) things and stuff categories, and (b) singulars and plurals noun phrases.

When the IoU thresholds are around 0, the recall values of all methods approach 1. However, the curves of our method are consistently above those of previous counterparts at most IoU thresholds, showing that our method consistently grounds more noun phrases under different evaluation conditions. In particular, the areas under our curves are still larger than those of other methods when the IoU threshold is beyond 0.9, which indicates our method can produce more accurate segmentation masks upon very strict metrics.

PPO-TD			POCL	Average Recall				
PC	MA	OC		overall	things	stuff	singulars	plurals
✓				64.7	62.4	69.8	67.6	57.3
✓	✓			64.9	62.6	70.2	67.7	57.6
✓	✓	✓		65.5	63.3	70.5	68.6	57.7
✓	✓	✓	✓	**66.1**	**64.0**	**70.7**	**69.1**	**58.3**

Table 2: Verifying the effectiveness of components in our proposed PPO-TD and POCL. "PC": pixel contexts. "MA": masked attention. "OC": object contexts.

4.4 Ablation Studies

We also conduct ablation studies on the PNG benchmark to verify the effectiveness of our network designs.

Component Analysis. In Table 2, we analyze the effects of our proposed components. The 1-st row denotes aggregat-

ing pixel contexts (PC) to phrases by standard cross-attention in our proposed PPO-TD, which has achieved considerably higher performance than previous methods. The result of PC shows that pixel contexts are necessary to PNG models for generating masks of high quality. The 2-nd row represents applying masked cross-attention (MA) between phrases and image features to aggregate pixel contexts from foreground regions. The improvement indicates that MA can suppress background noises to provide more phrase-relevant pixel contexts. In the 3-rd and 4-th row, OC denotes introducing learnable object tokens to provide object contexts for phrases, and POCL is the phrase-object contrastive loss which makes phrases obtain more precise object contexts. Combining OC and POCL can achieve a further performance boost of 1.2 Average Recall for *overall* split over the particularly strong model in the 2-nd row, which is very challenging. This result well demonstrates that enriching phrases with coupled pixel and object contexts can further improve the strong performance of pixel contexts only, which shows the effectiveness of introducing object contexts.

Number of PPO-TD layers. Table 3 shows the results of our PPO-TD with different numbers of layers. Since three scales of image features are grouped into a stage in PPO-TD for three layers, we evaluate the choices of 3, 6, and 9 layers. we can find that more layers in PPO-TD yield higher performance, which shows that iteratively enriching phrases with
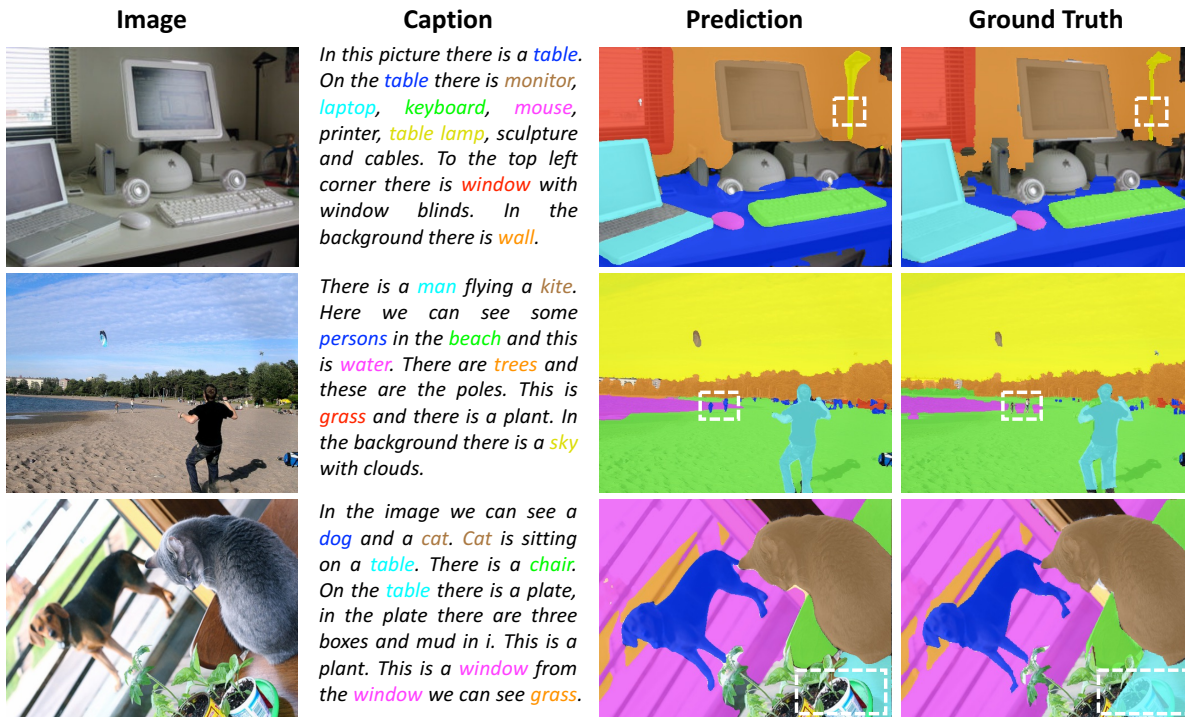
Image	**Caption**	**Prediction**	**Ground Truth**

Figure 4: Qualitative results of our method. The colors of phrases correspond to those of segments.

L	Average Recall				
	overall	things	stuff	sigulars	plurals
3	65.0	62.7	70.3	68.0	57.6
6	65.5	63.3	70.5	68.5	57.8
9	**66.1**	**64.0**	**70.7**	**69.1**	**58.3**

Table 3: Results of different numbers of PPO-TD layers L.

pixel and object contexts is beneficial.

Lang. Enc.	Average Recall				
	overall	things	stuff	sigulars	plurals
BERT [2018]	66.1	64.0	**70.7**	69.1	**58.3**
CLIP [2021]	**66.2**	64.3	70.5	**69.3**	58.1
T5 [2020]	**66.2**	**64.4**	70.5	**69.3**	**58.3**

Table 4: Results of different language encoders whose parameters are all fixed during training.

Different Language Encoders. Since our language encoder is fixed during training, we also explore the effects of using different language encoders to extract the phrase features. As shown in Table 4, both the large-scale vision-language pretraining model CLIP [Radford *et al.*, 2021] and large-scale language pretraining model T5 [Raffel *et al.*, 2020] yield similar performance to BERT, which indicates our method is insensitive to the structure of language encoder if its parameters are fixed during training.

4.5 Qualitative Results

We present the qualitative results of our method in Figure 4. Our method can generate high-quality segmentation according to dense phrases. We use white dashed boxes to mark some regions where our predictions are inconsistent with the ground truths. For example, in the 2-nd row, our method segments two people near the water while the ground-truth annotation is missing. In the 3-rd row, the ground truth annotates part of the plant pot as "table" while our method yields the correct prediction on the actual table area.

5 Conclusion

In this paper, we focus on the panoptic narrative grounding (PNG) task whose goal is to segment objects according to noun phrases in a narrative caption. An essential factor of PNG is visual-linguistic interaction, where previous methods aggregate proposal contexts or pixel contexts to phrases for realizing it. However, fragmentary visual information or semantic misalignment may be introduced by these methods and their performances are hence limited. Therefore, we propose to enrich phrases with coupled pixel and object contexts with a Phrase-Pixel-Object Transformer Decoder (PPO-TD) for a more comprehensive visual-linguistic interaction. In addition, we also propose a Phrase-Object Contrastive Loss (POCL) to aggregate more precise object contexts from more phrase-relevant object tokens. Extensive experiments on the PNG benchmark show that our method outperforms previous state-of-the-art methods by large margins.

Acknowledgments

References

[Akbari *et al.*, 2019] Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *CVPR*, 2019.

[Bajaj *et al.*, 2019] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3raphground: Graph-based language grounding. In *ICCV*, 2019.

[Cheng *et al.*, 2020] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020.

[Cheng *et al.*, 2021] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021.

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

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018.

[Ding *et al.*, 2022] Zihan Ding, Zi-han Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Xiaolin Wei, and Si Liu. Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In *ACM MM*, 2022.

[Dogan *et al.*, 2019] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *CVPR*, 2019.

[Feng *et al.*, 2021] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021.

[González *et al.*, 2021] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *ICCV*, 2021.

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

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[Huang *et al.*, 2020] 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.

[Hui *et al.*, 2020] 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.

[Jing *et al.*, 2021] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 2021.

[Kim *et al.*, 2022] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, 2022.

[Kirillov *et al.*, 2019a] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

[Kirillov *et al.*, 2019b] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[Li *et al.*, 2022a] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022.

[Li *et al.*, 2022b] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022.

[Lin *et al.*, 2014] 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.

[Liu *et al.*, 2021] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *TPAMI*, 2021.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[Luo *et al.*, 2020] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020.

[Mohan and Valada, 2021] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *IJCV*, 2021.

[Mu *et al.*, 2021] Zongshen Mu, Siliang Tang, Jie Tan, Qiang Yu, and Yueting Zhuang. Disentangled motif-aware graph learning for phrase grounding. In *AAAI*, 2021.

[Pang *et al.*, 2021] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021.

[Plummer *et al.*, 2018] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *ECCV*, 2018.

[Pont-Tuset *et al.*, 2020] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 2020.

[Tan *et al.*, 2020] Zhenxiong Tan, Tianrui Hui, Jinyu Chen, and Si Liu. Multi-granularity multimodal feature interaction for referring image segmentation. In *PRCV*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[Wang *et al.*, 2018] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *TPAMI*, 2018.

[Wang *et al.*, 2020] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020.

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

[Wu *et al.*, 2022] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022.

[Yang *et al.*, 2022] 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.

[Ye *et al.*, 2019] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019.

[Yu *et al.*, 2020] Tianyu Yu, Tianrui Hui, Zhihao Yu, Yue Liao, Sansi Yu, Faxi Zhang, and Si Liu. Cross-modal omni interaction modeling for phrase grounding. In *ACM MM*, 2020.

[Zhang *et al.*, 2021] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, 2021.

[Zhou *et al.*, 2021] Qianli Zhou, Tianrui Hui, Rong Wang, Haimiao Hu, and Si Liu. Attentive excitation and aggregation for bilingual referring image segmentation. *TIST*, 2021.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.