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]Harbin Institute of Technology (Shenzhen)
`{xiaolinhui16, pengfang21}@mails.ucas.ac.cn`,
`{xiaoshan.yang, csxu}@nlpr.ia.ac.cn`, `wangyw@pcl.ac.cn`

Abstract

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. Simultaneously, the current mask visual language modeling (MVLM) fails to capture the nuanced referential relationship between image-text in referring tasks. In this paper, 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 and referring-aware mask language modeling. Both modules not only reconstruct modality-related content but also cross-modal referring content. Within MRefM, we propose a referring-aware dynamic image masking strategy that is aware of the referred region rather than relying on fixed ratios or generic random masking schemes. 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. Our method consistently surpasses existing approaches and achieves SoTA performance on both grounding and segmentation tasks, providing valuable insights for future research. Our code and models are available at `https://github.com/linhuixiao/OneRef`.

1 Introduction

Visual Grounding (VG) aims to ground a region referred by a expression query text in a specific image. The generalized VG / referring tasks include Referring Expression Comprehension (REC) [69, 62, 101, 31, 14, 91, 90, 48], Phrase Grounding (PG) [1, 74], and Referring Expression/Image Segmentation (RES/RIS) [69, 94, 89]. 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. Unlike object detection [57, 58] or instance segmentation [26], which usually relies on a close-set of categories to detect or segment multiple regions that satisfy the object label, visual grounding is not limited to fixed categories. It requires understanding the semantics of the query text and then grounding or segmenting specific areas. Therefore, visual grounding is a task that strongly relies on the multimodal interaction and alignment of visual and linguistic features.

Since the introduction of BERT [16] and ViT [19, 7], the state-of-the-art (SoTA) grounding works have widely adopted a pre-training and fine-tuning paradigm. As illustrated in Fig. 1, existing studies

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

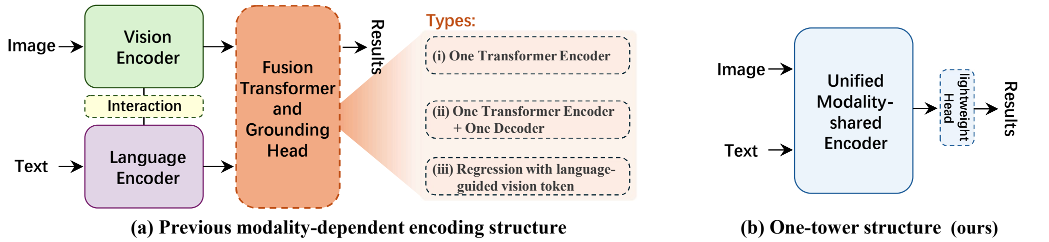


Figure 1: Comparison between our proposed approach and the mainstream REC/RES architectures.

employing pre-trained models, either utilizing uni-modal pre-trained models to separately transfer visual and language knowledge [14, 15, 98, 33, 55] or utilizing multimodal pre-trained models [91, 77, 89, 35], primarily fall into three typical architectures: ***(i)*** two modality encoders combined with a cross-modal fusion encoder, exemplified by TransVG *etc*. [11, 14, 98, 89, 91, 35, 92]; ***(ii)*** additionally incorporating a decoder, exemplified by MDETR *etc*. [33, 45, 97, 85, 54, 53, 77, 55]; ***(iii)*** direct regression based on language-guided visual features, such as LAVT, TransVG++, *etc*. [94, 15, 79, 99]. However, incorporating modality-dependent encoders in these studies presents a challenge for seamlessly integrating the two modalities into a unified feature space. Consequently, these works not only require an additional cross-modal Transformer-based [82] en-/decoder (*(i)* and *(ii)*), but also propose a variety of careful-designed interaction structures for modality-dependent encoders to facilitate early-stage fine-grained cross-modal alignment [15, 98, 92, 79, 54, 35, 55, 59], such as adapter [15, 35], cross-modal bridge [92], weight generation [79], image-text cross-attention [55, 54], *etc*. Therefore, these methods not only entail a large number of parameters but also involve intricate processes. Considering these critical limitations, we aim to explore simpler modality-shared grounding frameworks that can unify vision and language within a unified feature space, thereby obviating the necessity of the elaborate interaction modules, bulky fusion Transformer en-/decoders, as well as the special grounding tokens.

With the advancement of pre-training [70, 66], several studies have been conducted to explore unified modality-shared multimodal frameworks. YORO [29] implemented a shared encoder based on ViLT [37]. However, its modeling approach tends to overshadow the uni-modal knowledge and requires the encoder to incorporate additional query anchors, limiting its applicability for transfer with common pre-trained models. ONE-PEACE [86] has designed seven expert branches based on Mix-of-Expert (MoE) [5, 76, 22] to construct a three-modality foundation model to realize the integration of image, text, and audio modalities. However, their research employed extensive tri-modal data without exploring the potential utilization of MVLM for modeling the referring tasks. BEiT-3 [87] is built upon multi-way Transformer [5, 80], which adopts three MoE heads (*i.e*., vision, language, vision-language) and a modality-shared structure that effectively unifies vision and language within a shared feature space. It demonstrates notable advantages across various classification-like cross-modal fields (*e.g*., Retrieval, VQA *etc*.). However, no prior research has explored the utilization of BEiT-3 for achieving transfer in referring tasks. Consequently, our objective is to explore more concise and efficient referring grounding and segmentation transfer within a unified feature space on the one-tower model of BEiT-3. However, BEiT-3 model is pre-trained utilizing a generic Mask Vision Language Modeling (MVLM) approach, and this masking paradigm lacks fine-grained cross-modal referring ability and cannot effectively model the intricate referential relationship between images and text. As a result, there exists a significant gap when applying BEiT-3 to the regression-like referring tasks. Therefore, exploring how to incorporate fine-grained cross-modal referring capability into the mask modeling paradigm becomes an important research issue that has not been addressed yet.

In this paper, we propose a novel paradigm called ***Mask Referring Modeling (MRefM)***, as well as a unified and extremely concise grounding and referring segmentation framework named ***OneRef*** that no longer requires the fusion or interaction Transformer structure and the special grounding tokens.

Firstly, we propose MRefM paradigm to enhance the referring capability of BEiT-3 in a flexible manner. MRefM consists of two components: Referring-aware Mask Image Modeling (**Referring MIM**) and Referring-aware Mask Language Modeling (**Referring MLM**). The conventional MVLM is typically trained alternately or randomly with uni-modal MIM and MLM. In contrast, Referring MIM and Referring MLM are required to reconstruct two distinct types of content: their own modality-related content and cross-modal referring information. Specifically, ***(i) Referring MIM*** employs visual tokens after the dot product operation with the aggregated text token for reconstruction purposes. This not only entails reconstructing masked visual features itself but also necessitates reconstructing the visual target-relation score, which indicates the distance between the current token and the grounding region. The score encompasses four dimensions: horizontal and vertical distance to the grounding center, as well as width and height of the grounding region. In order to enhance the model's understanding capability for referred regions, we propose a referring-aware dynamic image masking strategy that replaces traditional ratio-fixed random masking so that referred regions are reconstructed with a relatively high mask ratio. ***(ii) Referring MLM*** employs text tokens after the dot product operation with the aggregated visual token for reconstruction purposes. This not only involves reconstructing masked text itself but also requires reconstructing semantic target-relation scores that represent the correlation degrees between current text tokens and referred image regions.

Secondly, existing grounding and segmentation models commonly employ a `[Region]` token and multiple query anchors to regress results. However, embedding the region token in backbone will disrupt the pre-trained model [15], and the query anchor also depends on the decoder [33]. With the unified feature space established by modality-shared encoder, we no longer need additional cross-modal en-/decoders to fuse uni-modal features, enabling us to more effectively leverage the knowledge acquired by pre-trained backbone. Benefiting from MRefM paradigm, the visual token inherently contains referring information. Consequently, we can discard special grounding token/anchors and directly construct lightweight and highly concise grounding and segmentation task heads based on the dot product operation within Referring MIM to unify the referring framework.

Contributions: Our contributions are threefold: ***(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.

2 Related work

2.1 Referring expression comprehension (REC) and segmentation (RES)

(i) REC. The recent supervised REC task, also known as visual grounding in a narrow sense, can be broadly categorized into **five main approaches**: **(1)** Fine-tuning with a uni-modal pre-trained language model and a closed-set detector. This setting is exemplified by TransVG [14], which builds upon the two-stage [102, 56, 52, 30] and one-stage [96, 95, 106] methods from the CNN era. It is considered the most conventional and extensively studied approach. **(2)** Fine-tuning with a pre-trained uni-modal language model and an open-set detection model pre-trained on box-level datasets mixed with multiple data sources. MDETR [33] represents this type of setting, where Fig. 1-(a)-(ii) plays a dominant role in its model structure. **(3)** Fine-tuning with multimodal self-supervised pre-trained models. CLIP-VG [91] serves as an example for this category, introduced primarily through the proposal of CLIP [70]. **(4)** Multimodal and multi-task mix-supervised pre-trained models. These methods typically combine multiple tasks while mixing datasets from each downstream task, employing mixed pre-training that incorporates both self-supervision and fine-grained supervision. UniTAB [97], OFA[85], *etc.* , represent such approaches where visual grounding often acts as one of the pre-training tasks. **(5)** Grounding multimodal large language models (GMLLMs). These methods influenced by works like GPT [6] or LLAMA [81] *etc*. These models integrate visual backbones into Large Language Models (LLMs) to generate grounding results rather than relying on regression techniques. Our approach mainly falls under type (3). ***(ii) RES.*** The development and approach categories of RES [49, 13, 35, 89, 79, 94, 93, 88] are generally similar to those of REC. However, the key distinction lies in the finer granularity of RES's output, which necessitates separate study from REC. In terms of model architecture, RES works predominantly employ two modality-dependent encoders and a decoder to generate the segmentation mask. Our work stands out as the first endeavor to explore RES within a unified multimodal feature space under a one-tower structure.

2.2 Mask vision language modeling

Motivated by the success of MLM [82] in BERT [16], MAE [27] and BEiT [4] have primary shifted their attention to MIM [21, 83, 3]. Subsequently, exemplified by BEiT-3 [87], numerous MVLM

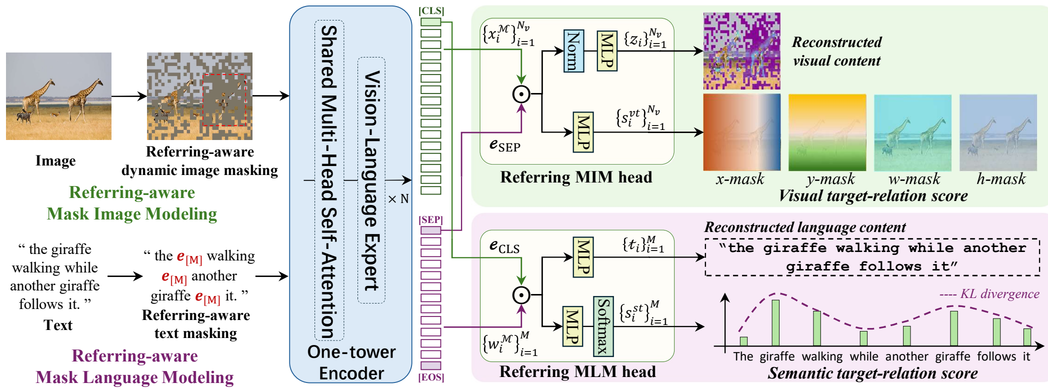


Figure 2: Illustration of our multimodal Mask Referring Modeling (MRefM) paradigm, which includes Referring-aware mask image modeling and Referring-aware mask language modeling.

works [61, 47, 36, 104, 2] have emerged, with most of these works implementing randomly alternating uni-modal MIM and MLM. Most relevant to our work are mask region modeling (known as MRM) [64, 83], which can be either unimodal MIM (*e.g.*, R-MAE [64]) or employ more fine-grained regional data and contrastive learning to reconstruct the alignment between regions and object labels (*e.g.*, ConLIP [61], VLT [17, 18] *etc.*). However, our work focuses on modeling the fine-grained referential relationship within image and text, so as to enhance the cross-modal referring capability, which is significantly different from these works.

3 Methodology

In this section, we propose our multimodal Mask Referring Modeling (***MRefM***) paradigm, which includes Referring MIM and Referring MLM, as well as a feature space unified grounding and segmentation framework ***OneRef***. We will introduce these methods in the following sections.

Following BEiT-3 [87], we employ a multimodal modality-shared Transformer [5] as the underlying backbone network. Initially, we perform mask-then-predict MRefM pre-training, and followed by transfer fine-tuning on the referring tasks. As shown in Fig. 2, the MRefM pre-training stage consists of two components: Referring-aware Mask Image Modeling (**Referring MIM**) and Referring-aware Mask Language Modeling (**Referring MLM**). Both modules aim to reconstruct two types of content: modality-related content within each modality and cross-modal fine-grained referring content.

3.1 Preliminaries

BEiT-3 [87] utilizes MIM, MLM, and MVLM for processing image, text, and image-text pairs respectively to facilitate the acquisition of general representations through MoE heads and shared multi-head self-attention. Notably, MVLM involves alternate training of MIM and MLM. Specifically:

(i) Vanilla mask image modeling. We denote $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ as the input image, and it is tokenized by a convolution projection to $N_v = HW/P^2$ patches $\{\boldsymbol{x}_i^p\}_{i=1}^{N_v}$, where $\boldsymbol{x}^p \in \mathbb{R}^{N_v \times D}$, H, W are the image size, and P is the patch size, D is the hidden dimension of the unified feature space. Then, we leverage a specific masking strategy to mask a specific number of image patches. The masked position is termed as $\mathcal{M}_v$. Thus, a shared learnable embedding $\boldsymbol{e}_{[\mathrm{M}]}$ is used to replace the masked image patch embeddings $\boldsymbol{x}_i^p$ if $i \in \mathcal{M}_v$. Subsequently, we prepend a learnable [CLS] token to the input, *i.e.*, $[\boldsymbol{e}_{\texttt{CLS}}, \{\boldsymbol{x}_i^p\}_{i=1}^{N_v}]$, and feed them to the one-tower Transformer. Next, we utilize a MIM head which consists of a linear projection and a softmax classifier to predict the visual tokens of the masked positions based on the corrupted image $\boldsymbol{x}^{\mathcal{M}}$. The visual tokens are obtained by the image tokenizer VQ-KD$_{\mathrm{CLIP}}$ proposed in BEiT v2 [67], which provides supervisions for the MIM self-supervised learning procedure. The visual tokens of the original image are denote as $\{\boldsymbol{z}_i\}_{i=1}^{N_v}$, and $\mathcal{I}$ denotes the pre-training images. Then, the training loss of MIM is defined as:

$$\mathcal{L}_{\mathrm{MIM}} = -\sum_{\boldsymbol{x} \in \mathcal{I}} \sum_{i \in \mathcal{M}_v} \log p(\boldsymbol{z}_i | \boldsymbol{x}_i^{\mathcal{M}}). \tag{1}$$

(ii) Vanilla mask language modeling. The input text is tokenized and projected to the word embeddings $\{\boldsymbol{w}_i\}_{i=1}^{M}$ by a SentencePiece tokenizer [41] with vocabulary size of 64010, where $\boldsymbol{w} \in \mathbb{R}^{M \times D}$, M is the length of tokenized text sequence. Then, following BEiT-3 [87], we randomly mask the text tokens with a fixed masking ratio δ. The masked position is termed as $\mathcal{M}_w$. Thus, a shared learnable embedding $\boldsymbol{w}_{[\mathrm{M}]}$ is used to replace the masked word tokens $\boldsymbol{w}_i$ if $i \in \mathcal{M}_w$. We prepend a learnable special tokens `[SEP]` and an end-of-sequence token `[EOS]` to the sequence, *i.e.*, $[\boldsymbol{e}_{\texttt{SEP}}, \{\boldsymbol{w}_i\}_{i=1}^{M}, \boldsymbol{e}_{\texttt{EOS}}]$, and feed them to the one-tower Transformer. Similarly, we utilize a MLM head which consists of a linear projection to predict the text tokens of masked positions based on the corrupted text data $\boldsymbol{w}^{\mathcal{M}}$. The original textual tokens are denoted as $\{\boldsymbol{t}_i\}_{i=1}^{M}$, and $\mathcal{T}$ denotes the pre-training text sequences. Then, the training loss of MLM is defined as:

$$\mathcal{L}_{\mathrm{MLM}} = -\sum_{\boldsymbol{x} \in \mathcal{T}} \sum_{i \in \mathcal{M}_w} \log p(\boldsymbol{t}_i | \boldsymbol{w}_i^{\mathcal{M}}). \tag{2}$$

3.2 Referring-aware mask image modeling

After concatenating the visual and text tokens and feeding them into the modality-shared encoder, the vanilla MVLM is commonly implemented through the alternating use of MIM and MLM [87]. Despite the multimodal features are interact within the modality-shared encoder, it fundamentally remains a unimodal information reconstruction. Additionally, MVLM acquires general knowledge by randomly masking images and texts, it fails to effectively model the referential relationship. Hence, we propose Referring MIM and Referring MLM methods. Specifically, as shown in Fig. 2, our proposed Referring MIM incorporates two additional components: the reconstruction of visual target-relation score and a referring-aware dynamic masking strategy.

In Referring MIM (Fig. 2), instead of using uni-modal visual tokens [87, 47, 2], we propose to employ visual tokens that dot product with the aggregated text token $\boldsymbol{e}_{\texttt{SEP}} \in \mathbb{R}^{1 \times D}$ for the reconstruction purpose. The reconstruction of Referring MIM involves not only the modality-related content $\{\boldsymbol{z}_i\}_{i \in \mathcal{M}_v}$ but also the visual target-relation scores $\{\boldsymbol{s}_i^{vt}\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times 4}$. We utilize a visual target-relation head which consists of a three-layer perceptron (MLP) to predict the scores. The scores represent the distance between each patch token $\{\boldsymbol{x}_i^{\mathcal{M}}\}_{i=1}^{N_v}$ and the referred region $\mathcal{B} = (x_c, y_c, w_r, h_r)$, where (x_c, y_c, w_r, h_r) denote the center coordinate and the width and height of the referred region. It encompasses four masks, *i.e.*, *x-, y-, w-, h- masks*, which represent the normalized horizontal and vertical distances from the referred center, *i.e.*, $((x - x_c)/W,\ (y - y_c)/H)$, and the proportion of width and height on the the referred region, *i.e.*, $(P/w_r,\ P/h_r)$, respectively, where (x, y) denote the center coordinate of each patch. We denote $\odot$ as dot product operation. Finally, the training loss of Referring MIM is defined as:

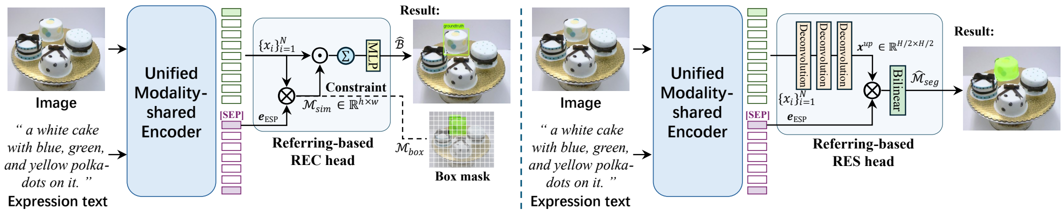
$$\mathcal{L}_{\text{Referring MIM}} = -\sum_{\boldsymbol{x} \in \mathcal{I}} \sum_{i \in \mathcal{M}_v} \log p(\boldsymbol{z}_i | (\boldsymbol{x}_i^{\mathcal{M}} \odot \boldsymbol{e}_{\texttt{SEP}})) - \sum_{\boldsymbol{x} \in \mathcal{I}} \sum_{i \in [1, N_v]} \log p(\boldsymbol{s}_i^{vt} | (\boldsymbol{x}_i^{\mathcal{M}} \odot \boldsymbol{e}_{\texttt{SEP}})). \tag{3}$$

Referring-aware dynamic image masking strategy. As shown in Fig. 4, among the existing masking strategies, MAE [27] adopts a high-ratio random masking while BEiT-3 [87] uses a low-ratio block-wise random masking, neither of which effectively directs attention to the referred region. SemMAE [43] proposes a semantic-guided masking that requires additional bulky semantic models and limits its generality. To enhance the model's understanding of the referred region through surrounding visual context and text semantics, we propose a referring-aware dynamic masking strategy as shown in Algo. 1.

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 $\mathcal{M}$.

The strategy avoids the drawbacks of the aforementioned methods and directs the model's attention to the referred region. Specifically, we denote the shape after patch reshaping of the image as (h, w), where $h = H/P$, $w = W/P$, and $N_v = h \times w$. To maximize the masking of the



(a) Referring Expression Comprehension task (b) Referring Expression Segmentation task

Figure 3: Illustration of the referring-based grounding and segmentation transfer.

referred region (x_s, y_s, w_r, h_r), where x_s, y_s represent the starting coordinates of the referred region, we introduce a margin m to its surroundings and denote its patch coordinates as $(x_{sp}, y_{sp}, w_{rp}, h_{rp})$, *i.e.*, $x_{sp} = \lfloor x_s/P \rfloor - m$, $w_{rp} = \lfloor w_r/P \rfloor + m$, y_{sp} and h_{rp} are similar to that of x_{sp} and w_{rp}, where $\lfloor \cdot \rfloor$ indicates rounding down to an integer. Thus, the number of referred patches is denote as $N_r = h_{rp} \times w_{rp}$. Then, as shown in Algo. 1, to ensure that the model allocates appropriate attention to the in-contextual information around the referred region, we utilize a random masking with a relatively low ratio β for its surroundings. Simultaneously, we employ a block-wise masking approach with a high ratio γ in the extended area of the region. Since referred regions vary across different image-text pairs, each sample's entire masking ratio α is dynamically determined:

$$\alpha = [\beta \cdot (N_v - N_r) + \gamma \cdot N_r]/N_v. \tag{4}$$

3.3 Referring-aware mask language modeling

Similarly, in Referring MLM, instead of using uni-modal linguistic tokens [87, 47, 2], we propose to employ linguistic tokens that dot product with the aggregated visual token $\boldsymbol{e}_{\texttt{CLS}} \in \mathbb{R}^{1\times D}$ for the reconstruction purpose. The reconstruction of Referring MLM involves not only the modality-related content $\{\boldsymbol{t}_i\}_{i\in\mathcal{M}_w}$ but also the semantic target-relation scores $\{\boldsymbol{s}_i^{st}\}_{i=1}^{M}$. The score represents the correlation between the referred target and the language token, which is obtained by a teacher model (*i.e.*, a BEiT-3 model with performed image-text contrastive intermediate tuning) with calculating the weighted sum of the normalized similarity between the language token $\{\boldsymbol{w}_i\}_{i=1}^{M}$ and the aggregated visual token $\boldsymbol{e}_{\texttt{CLS}}^{reg}$ of referred region, as well as the aggregated visual token $\boldsymbol{e}_{\texttt{CLS}}^{img}$ of entire image:

$$\boldsymbol{s}^{st} = \lambda_{reg} \cdot \sigma(< \boldsymbol{e}_{\texttt{CLS}}^{reg\top},\ \{\boldsymbol{w}_i\}_{i=1}^{M} >) + \lambda_{img} \cdot \sigma(< \boldsymbol{e}_{\texttt{CLS}}^{img\top},\ \{\boldsymbol{w}_i\}_{i=1}^{M} >), \tag{5}$$

where $< \cdot, \cdot >$ denotes cosine similarity operation, σ denotes the softmax normalization.As shown in Fig. 2, we utilize a semantic target-relation head which consists of a three-layer MLPs and a softmax normalization to predict the scores. Finally, the training loss of Referring MLM is defined as:

$$\mathcal{L}_{\text{Referring MLM}} = -\sum_{\boldsymbol{w}\in\mathcal{T}}\sum_{i\in\mathcal{M}_w} \log p(\boldsymbol{t}_i|(\boldsymbol{w}_i^{\mathcal{M}} \odot \boldsymbol{e}_{\texttt{CLS}})) - \sum_{\boldsymbol{w}\in\mathcal{T}}\sum_{i\in[1,M]} \log p_{kl}(\boldsymbol{s}_i^{st}|(\boldsymbol{w}_i^{\mathcal{M}} \odot \boldsymbol{e}_{\texttt{CLS}})), \tag{6}$$

where p_{kl} represents a probabilistic prediction with Kullback-Leibler divergence [28].

3.4 Referring-based grounding and segmentation transfer

The modeling of visual and language in a unified feature space eliminates the need for the commonly-used Transformer-based fusion en-/decoder [14, 54, 55] and various early-stage interaction techniques [15, 79, 92] to further uniform the visual and language features. Additionally, since the referential relationship is modeled by MRefM during pre-training, we can accurately regress the results of grounding and referring segmentation using the output tokens, without relying on the widely-used special grounding tokens (*e.g.*, `[Region]` token [14, 15, 98, 91, 92], query anchors [55, 54]).

Referring expression comprehension. As illustrated in Fig. 3-(a), based on Referring MIM, we initially perform a similarity operation between visual tokens $\{\boldsymbol{x}_i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v\times D}$ and aggregated language token $\boldsymbol{e}_{\texttt{SEP}} \in \mathbb{R}^{1\times D}$ to obtain a softmax-normalized similarity mask $\mathcal{M}_{sim} \in \mathbb{R}^{h\times w}$. This mask is then replicated and multiplied back to each hidden dimension of the visual tokens. Subsequently, the visual tokens are summed to yield reduced tokens, which are finally subjected to regress the prediction box $\hat{\mathcal{B}} = (\hat{x}_c, \hat{y}_c, \hat{w}_r, \hat{h}_r)$ using a 3-layer MLPs:

$$\hat{\mathcal{B}} = \text{MLP}(\sum_{i\in[1,N_v]} (\text{Repeat}(\sigma(< \boldsymbol{e}_{\texttt{ESP}}{}^{\top},\ \{\boldsymbol{x}_i\}_{i=1}^{N_v} >)) \odot \text{MLP}(\{\boldsymbol{x}_i\}_{i=1}^{N_v}))). \tag{7}$$

To enhance the accuracy of cross-modal similarity, we propose treating the similarity as a coarse-grained downsampling bounding box mask $\mathcal{M}_{box} \in \mathbb{R}^{h\times w}$ and imposing segmentation loss (*i.e.*,

Table 1: Comparison with **latest** SoTA methods on the five datasets for **REC/PG** tasks with single-dataset fine-tuning setting. We highlight best result of base model in red and **bold** for large model.

Methods	Venue	Visual Backbone	Language Backbone	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	ReferIt test	Flickr 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**

Table 2: Comparison with **latest** SoTA methods for **REC** task with dataset-mixed intermediate pre-training setting. 'RefC' represents the mixup of RefCOCO/+/g training data. † indicates RefC has been used during pre-training. 'G-DINO-L*' denotes 'O365,OI,GoldG,Cap4M,COCO,RefC'.

Methods	Venue	Visual/Language Backbone	Intermediate pretrain data	Data size	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg 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**

Focal loss [51] and Dice/F-1 loss [63]) on the sigmoid activated similarity mask $\mathcal{M}_{sim} \in \mathbb{R}^{h \times w}$ with coefficient λ_{f_box} and λ_{d_box} as the box mask constraints:

$$\mathcal{L}_{box\_mask\_constraints} = \lambda_{f\_box}\mathcal{L}_{focal}(\mathcal{M}_{sim}, \mathcal{M}_{box}) + \lambda_{d\_box}\mathcal{L}_{dice}(\mathcal{M}_{sim}, \mathcal{M}_{box}). \tag{8}$$

Consequently, the loss function for the REC task can be reformulated as the weighted sum of vanilla grounding loss (*i.e.*, smooth L1 loss [25] and Giou loss [73]) and box mask constraints:

$$\mathcal{L}_{REC} = \lambda_{L_1}\mathcal{L}_{\text{L1}}(\hat{\mathcal{B}}, \mathcal{B}) + \lambda_{giou}\mathcal{L}_{\text{giou}}(\hat{\mathcal{B}}, \mathcal{B}) + \mathcal{L}_{box\_mask\_constraints}. \tag{9}$$

Referring expression segmentation. As illustrated in Fig. 3-(b), the implementation of referring segmentation can be regarded as a simplified version of grounding. Initially, we employ a 3-layer deconvolution to up-sample the visual token to $\boldsymbol{x}^{up} \in \mathbb{R}^{H/2 \times W/2}$. Subsequently, cosine similarity operations are performed on the up-sampled visual tokens and the aggregated language token. The resulting similarity mask is then utilized as the final predicted mask $\hat{\mathcal{M}}_{seg} \in \mathbb{R}^{H \times W}$ after applying 1-layer bilinear interpolation. We denote the ground truth segmentation mask as $\mathcal{M}_{seg} \in \mathbb{R}^{H \times W}$, then the loss function for RES is defined as follows:

$$\mathcal{L}_{RES} = \lambda_{f\_seg}\mathcal{L}_{focal}(\hat{\mathcal{M}}_{seg}, \mathcal{M}_{seg}) + \lambda_{d\_seg}\mathcal{L}_{dice}(\hat{\mathcal{M}}_{seg}, \mathcal{M}_{seg}). \tag{10}$$

4 Experiments

4.1 Experimental setups

Datasets and evaluation metrics. Our method is validated in the REC, RES, and PG tasks with five widely used datasets, namely three REC/RES datasets (RefCOCO/+/g [101, 62]), as well as two PG datasets (ReferItGame [34] and Flickr30k Entities [68]). In PG, the query pertains to a specific phrase,

Table 3: Comparison with **latest** SoTA methods (**mIoU** metric) on the three datasets for **RES** task with both single-dataset fine-tuning setting and dataset-mixed intermediate pre-training setting.

Methods	Venue	Visual/Language Backbone	Intermediate pretrain data	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg 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**	**73.76**
Dataset-mixed intermediate pre-training setting:											
PolyFormer-B†[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†[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†[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† (unsupervised)		BEiT3-B / BEiT3-B	RefC	78.20	79.26	75.92	72.54	75.54	67.39	71.28	71.13
OneRef-B† (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† (ours)	NeurIPS'24	BEiT3-L / BEiT3-L	RefC	**81.26**	**83.06**	**79.45**	**76.60**	**80.16**	**72.95**	**75.68**	**76.82**

Table 4: Ablation of MRefM on mixup pre-training setting.

MIM	MLM	image masking strategy	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test
✗	✗	✗	78.56	83.36	71.72	80.41	80.52
vanilla	vanilla	random	79.68	84.59	72.11	81.35	81.11
vanilla	vanilla	referring-aware	80.06	85.77	73.96	81.96	82.16
Ref MIM	vanilla	referring-aware	83.64	88.26	76.58	83.55	85.86
vanilla	Ref MLM	referring-aware	81.52	86.87	75.93	82.88	84.32
Ref MIM	Ref MLM	random	85.08	89.12	78.56	85.57	86.89
Ref MIM	Ref MLM	referring-aware	**86.38**	**90.38**	**79.47**	**86.82**	**87.32**

Table 5: Ablation of the task heads.

Architecture (Fine-tuning setting)	RefCOCOg val	RefCOCOg test
full model in REC	**83.68**	**83.52**
full model *w/o* box mask loss	82.54	82.02
w. fusion encoder + region token	78.93	78.51
full model in RES	**69.37**	**69.70**
deconv after similarity operation	67.98	68.62
4-layer deconv *w/o* linear upsample	68.33	68.96
2-layer deconv *w.* 2-layer upsample	67.51	67.65

while in REC and RES, the query refers to a reference expression. The text of RefCOCO+/g exhibits greater length and complexity in comparison to that of RefCOCO. In REC/PG, we follow previous works [14, 96] that employs Intersection-over-Union (IoU) as the evaluation metric, *i.e.*, a prediction is deemed accurate only when its IoU exceeds or equals 0.5. We compute the prediction accuracy for each dataset as a performance indicator. While in RES, we follow previous works [79, 35] that employs mean IoU (mIoU) and overall IoU (oIoU) for each dataset as the indicators. The detailed statistics information regarding these five datasets are provided in the Appendix B.

Experimental details. Since MRefM is proposed on the basis of the traditional MVLM, considering the pre-training cost of MVLM, we adopt BEiT-3 [87] base and large model as our initial weights and then perform intermediate MRefM pre-training on the task-relevant dataset. Such intermediate pre-training is common in existing grounding works [55, 54, 33]. As described in Sec. 2.1, to verify the effectiveness of our MRefM approaches, **we conduct extensive experiments on three settings**: **(1) The basic single-dataset fine-tuning setting.** This setting does not require additional training data and aligns with existing supervised and self-supervised transfer approaches [91, 92, 89, 35]. In this setting, we perform supervised single-dataset intermediate MRefM pre-training before fine-tuning. **(2) The setting of fine-tuning with supervised dataset-mixed intermediate pre-training.** This setting aligns with existing grounded pre-trained approaches, such as Grounding-DINO [55], DQ-DETR [54] *etc*. we perform an MRefM intermediate pre-training before fine-tuning. **(3)** To verify the generality of MRefM, we perform **the setting of fine-tuning with unsupervised intermediate MRefM pre-training**. There are several ways to obtain the regions in the regional masking modeling works [64, 24, 78], such as Felzenswalb-Huttenlocher (FH) algorithm [23], SAM [38] *etc*. Thus, we adopt the unsupervised, fast, image-computable FH algorithm [23] to generate regions following R-MAE [64]. We then select the referred one using a BEiT-3 model with performed image-text contrastive intermediate tuning. More details about the selection of the unsupervised regions, network architecture, training and inference, model hyperparameters *etc*. are provided in the Appendix C.

4.2 Comparison with state-of-the-art methods

Referring expression comprehension. As shown in Tab. 1 and Tab. 2, we conducted experiments for the REC and PG tasks across **three settings**. **(1)** In the single-dataset fine-tuning setting, our base model surpasses the current SoTA method HiVG [92] by 2.07%(testB), 6.15%(testB), 4.73%(test),

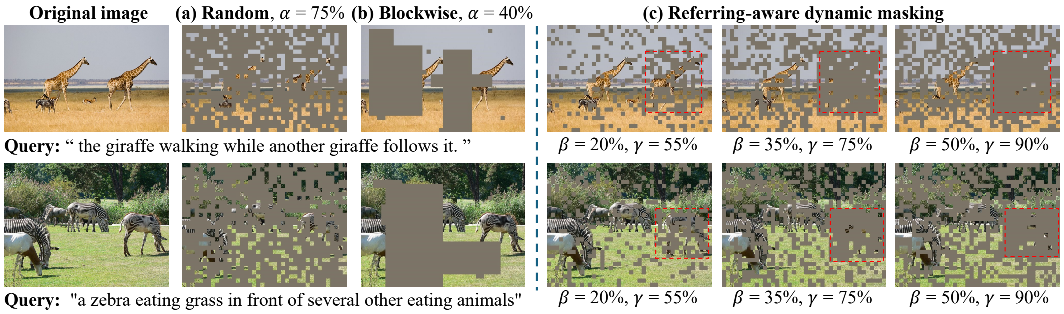


Figure 4: Illustrations of random masking (MAE) [27], block-wise masking (BEiT) [4], and our referring-aware dynamic masking. α denotes the entire masking ratio, while β and γ denote the masking ratio beyond and within the referred region.

Table 6: Generality study of MRefM on RefCOCOg.

Architecture	Backbone	Single-dataset		Mixup pretrain	
		val	test	val	test
TransVG	DETR / BERT-B	68.67	67.73	75.73	75.86
MRefM-TransVG	DETR / BERT-B	**71.51**	**70.84**	**78.71**	**78.69**
CLIP-VG	CLIP-B / CLIP-B	73.18	72.54	78.67	78.54
MRefM-CLIP-VG	CLIP-B / CLIP-B	**74.22**	**74.50**	**80.48**	**80.83**

Table 7: Generality of the task heads.

Architecture (Fine-tuning setting)	RefCOCOg	
	val	test
TransVG++ (Reproduced by us)	75.04	75.55
TransVG++ *w.* our REC head	**76.65**	**77.09**
LAVT [94]	63.34	63.62
LAVT *w.* our RES head	**64.84**	**65.35**

1.95%(test), and 1.50%(test) on the five datasets respectively, while also significantly outperforming the traditional uni-modal detector-based approach TransVG++ [15] by 4.37%(testB), 7.98%(testB), 7.22%(test), 2.47%(test), and 2.12%(test), respectively. **(2)** In dataset-mixed pre-training setting, our base model outperforms HiVG [92] by 1.35%, 2.79%, and 2.63% on RefCOCO/+/g testB/testB/test splits, outperforms Grounding-DINO [55] by 2.59%, 4.76%, and 2.38%, exceeds OFA by 5.28%, 5.18% ,and 5.01% , and even surpasses LION [9] - a GMLLM model that is 20-60 times larger than ours - by 3.76%, 2.13% ,and 1.69%. Note that among these works, UniTAB [97], OFA[85], LION [9] also utilize the MVLM on the pre-training stage. **(3)** Furthermore, we achieve competitive performance in the unsupervised setting, which shows the generality of MRefM paradigm. Additionally, our large-size model exhibits remarkable scalability with further substantial improvements in performance. More detailed results are provided in the Appendix E.

Referring expression segmentation. As presented in Tab. 3 (mIoU metric), we conducted experiments for RES task under **three settings**. **(1)** In the single-dataset fine-tuning setting, our base model surpasses the SoTA self-supervised method RISCLIP [35] by 2.65%, 4.77%, and 1.73% on RefCOCO/+/g testB/testB/test splits, respectively, while also significantly outperforming the traditional uni-modal detector-based approach VG-LAW [79] by 3.42%, 7.31%, and 4.57%, respectively. **(2)** In the dataset-mixed pre-training setting, our base model achieves superior performance compared to the SoTA method RISCLIP [35] with improvements of 4.53%, 8.21%, and 5.39%. **(3)** In the unsupervised pre-training setting, we also achieve competitive performance. Additionally, our large-size model also exhibits remarkable scalability and demonstrates a substantial improvement in performance. For oIoU metric, the results are presented in Appendix E.2 (Tab. 13).

4.3 Ablation study

The Mask Referring Modeling. In Tab. 4, we conducted ablation studies on MRefM, which included Referring MIM ('Ref MIM'), Referring MLM ('Ref MLM'), and referring-aware dynamic image masking ('referring-aware'). The 'vanilla' denotes the vanilla MVLM described in Sec. 3.1. As shown in Tab. 4, referring MIM, referring MLM, and dynamic masking strategy resulted in improvements of 3.70%, 2.16%, and 1.05% on the RefCOCOg-test dataset, and with an overall improvement of 6.21%, demonstrates the effectiveness of our methods. More results are provided in the Appendix E.4.

The referring-aware dynamic masking strategy. Fig. 4 presents a schematic of the three masking strategies. In our experiments, as illustrate in Fig. 4-(c), β and γ demonstrate optimal performance at values of 0.35 and 0.75, respectively. More detailed results are provided in the Appendix E.5.

The referring-based task heads. We conducted ablation studies on the design of two referring-based task heads. Tab. 5 reveals that our modeling method effectively captures referring information at the backbone stage, benefiting from the one-tower structure. This approach is significantly more efficient than the traditional fusion encoder and special token-based method. Additionally, our proposed box mask loss also contributes to a performance gain of 1.50%(test).

4.4 Generality study

The generality of MRefM. Firstly, we perform an unsupervised MRefM pre-training in Tab. 2 and Tab. 3, both of which achieve competitive performance. Secondly, we replace the backbone and apply MRefM on DETR and CLIP by using TransVG [14] and CLIP-VG [91] under the two settings. Since the two frameworks do not interact at backbone stage, we build MRefM on the fusion encoder. In Tab. 6, MRefM can effectively learn referring representation, resulting in an overall performance gain of about 2.0%. All these findings demonstrate the validity and generality of the MRefM paradigm.

The generality of referring-based task heads. Since both TransVG++ [15] and LAVT [94] have modality interactions at backbone stage, we attempted to apply our task heads to both frameworks. TransVG++ is reproduced by us since its code is not available. Tab. 7 shows that our proposed task heads achieve a 1.5+% improvement in both REC and RES, offering a new avenue for future research.

5 Conclusion

In this paper, 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 to capture the referential relationships between vision and text. We demonstrate the effectiveness and generality of MRefM across three settings on REC, PG, and RES tasks, consistently achieving groundbreaking results. Furthermore, leveraging unsupervised methods enables potential large-scale pre-training of MRefM in the future, presenting a new direction for referring tasks.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62036012, U23A20387, 62322212, 62072455, in part by Pengcheng Laboratory Research Project under Grant PCL2023A08, in part by National Science and Technology Major Project under Grant 2021ZD0112200, in part by Alibaba Innovative Research Program, and also in part by CAS Project for Young Scientists in Basic Research (YSBR-116).

References

[1] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12476–12486, 2019.

[2] Tarik Arici, Mehmet Saygin Seyfioglu, Tal Neiman, Yi Xu, Son Train, Trishul Chilimbi, Belinda Zeng, and Ismail Tutar. Mlim: Vision-and-language model pre-training with masked language and image modeling. *arXiv preprint arXiv:2109.12178*, 2021.

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021.

[5] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[9] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. *arXiv preprint arXiv:2311.11860*, 2024.

[10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[13] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. *arXiv preprint arXiv:2312.12198*, 2023.

[14] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[15] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

[17] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

[18] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916, 2022.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding (CVIU)*, 114:419–428, 2010.

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[22] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[23] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004.

[24] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[25] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050: 9, 2015.

[29] Chih-Hui Ho, Srikar Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. Yoro-lightweight end to end visual grounding. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 3–23. Springer, 2023.

[30] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE TPAMI*, 2019.

[31] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[32] Ziling Huang and Shin'ichi Satoh. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7753–7762, 2023.

[33] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[34] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[35] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip's image-text alignment to referring image segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4611–4628, 2024.

[36] Sungwoong Kim, Daejin Jo, Donghoon Lee, and Jongmin Kim. Magvlt: Masked generative vision-and-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23338–23348, 2023.

[37] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[41] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, page 66, 2018.

[42] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[43] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022.

[44] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[45] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34:19652–19664, 2021.

[46] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[47] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023.

[48] Zhitian Li, Wuhao Yang, Linhui Xiao, Xingyin Xiong, Zheng Wang, and Xudong Zou. Integrated wearable indoor positioning system based on visible light positioning and inertial navigation using unscented kalman filter. In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE, 2019.

[49] Zizhang Li, Mengmeng Wang, Jianbiao Mei, and Yong Liu. Mail: A unified mask-image-language trimodal network for referring image segmentation. *arXiv preprint arXiv:2111.10747*, 2021.

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

[51] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[52] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *ICCV*, pages 4673–4682, 2019.

[53] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023.

[54] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1728–1736, 2023.

[55] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[56] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, pages 1950–1959, 2019.

[57] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23776–23786, 2023.

[58] Yabo Liu, Jinghua Wang, Linhui Xiao, Chengliang Liu, Zhihao Wu, and Yong Xu. Foregroundness-aware task disentanglement and self-paced curriculum learning for domain adaptive object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[59] Yunfei Liu, Zhitian Li, Linhui Xiao, Shuaikang Zheng, Pengcheng Cai, Haifeng Zhang, Pengcheng Zheng, and Xudong Zou. Fdo-calibr: visual-aided imu calibration based on frequency-domain optimization. *Measurement Science and Technology*, 34(4):045108, 2023.

[60] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[61] Ziyang Luo, Yadong Xi, Rongsheng Zhang, GongZheng Li, Zeng Zhao, and Jing Ma. Conditioned masked language and image modeling for image-text dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 130–140, 2022.

[62] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[63] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[64] Duy Kien Nguyen, Yanghao Li, Vaibhav Aggarwal, Martin R Oswald, Alexander Kirillov, Cees GM Snoek, and Xinlei Chen. R-mae: Regions meet masked autoencoders. In *The Twelfth International Conference on Learning Representations*, 2023.

[65] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[66] Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Transactions on Multimedia*, 2023.

[67] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022. URL `https://api.semanticscholar.org/CorpusID:251554649`.

[68] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2641–2649, 2015.

[69] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

[70] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[71] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[72] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[73] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[74] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer, 2016.

[75] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[76] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2016.

[77] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[78] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[79] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2023.

[80] Jiajia Tang, Kang Li, Ming Hou, Xuanyu Jin, Wanzeng Kong, Yu Ding, and Qibin Zhao. Mmt: Multi-way multi-modal transformer for multimodal learning. In *IJCAI*, pages 3458–3465, 2022.

[81] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[83] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2122–2131, 2023.

[84] Ning Wang, Jiajun Deng, and Mingbo Jia. Cycle-consistency learning for captioning and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5535–5543, 2024.

[85] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[86] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

[87] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pre-training for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[88] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[89] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[90] Linhui Xiao, Jinge Wang, Xiaosong Qiu, Zheng Rong, and Xudong Zou. Dynamic-slam: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117:1–16, 2019.

[91] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 2023.

[92] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. HiVG: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 2024. URL https://openreview.net/forum?id=NMMyGy1kKZ.

[93] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023.

[94] 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*, pages 18155–18165, 2022.

[95] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, pages 4683–4693, 2019.

[96] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, pages 387–404, 2020.

[97] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.

[98] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15502–15512, 2022.

[99] Jiabo Ye, Junfeng Tian, Ming Yan, Haiyang Xu, Qinghao Ye, Yaya Shi, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, et al. Uniqrnet: Unifying referring expression grounding and segmentation with qrnet. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[100] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

[101] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[102] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[103] Heng Zhao, Joey Tianyi Zhou, and Yew-Soon Ong. Word2pix: Word to pixel cross-attention transformer in visual grounding. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[104] Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. Mamo: masked multimodal modeling for fine-grained vision-language representation learning. *arXiv preprint arXiv:2210.04183*, 2022.

[105] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[106] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. 2021.

[107] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

[108] Hong Zhu, Qingyang Lu, Lei Xue, Mogen Xue, Guanglin Yuan, and Bineng Zhong. Visual grounding with joint multi-modal representation and interaction. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[109] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Appendix

We provide an overview of the Appendix below:

- **Appendix A: Explanation of the task definition**
- **Appendix B: Introduction of the datasets**
  - Appendix B.1 The five referring datasets.
  - Appendix B.2 Explanation of the dataset abbreviations.
  - Appendix B.3 Comparison of datasets used in the pre-trained models.
- **Appendix C: Implementation details**
- **Appendix D: Technical remarks**
  - Appendix D.1 Further explanation for the effectiveness mechanism of the visual target-relation score.
  - Appendix D.2 The selection of the unsupervised regions.
  - Appendix D.3 Referring-aware text masking.
  - Appendix D.4 The difference of the task heads between ours with other frameworks.
- **Appendix E: Extra experimental results**
  - Appendix E.1 The results on phrase grounding task under mixup pre-training setting.
  - Appendix E.3 Computational costs analysis compared with SoTA methods.
  - Appendix E.4 Complete ablation study of MRefM.
  - Appendix E.5 Ablation study of the mask ratio in referring-aware dynamic masking.
- **Appendix F: Visualization of the results**
- **Appendix G: Further discussions**
  - Appendix G.1 Limitations.
  - Appendix G.2 Broader impacts.

A Explanation of the task definition

As explained in Sec. 1 of the main text, Visual Grounding (VG) aims to grounding a region referred by a query text in a specific image. The generalized visual grounding includes Referring Expression Comprehension (REC), Phrase Grounding (PG), and Referring Expression Segmentation (RES) tasks. However, in recent years, REC and RES have often been studied separately. Therefore, in numerous works [14, 15, 79, 92, 91, 98], visual grounding specifically refers to REC and PG tasks, which involve grounding a rectangular region. In this paper, we follow the mainstream and have not clearly separated the "grounding" from "generalized visual grounding" and "REC and PG tasks". When expressing the experimental task, "grounding" usually refers to REC or PG tasks, so as to discuss it parallelly with the RES task.

B Introduction of the datasets

B.1 The five referring datasets

We present the detailed descriptions of the five referring datasets used in our experimental study on the REC, PG, and RES tasks. Tab. 8 presents the detailed statistics.

RefCOCO/RefCOCO+/RefCOCOg. These three datasets belong to the Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES) tasks, and the images of these three datasets derived from MSCOCO [50]. Expressions in RefCOCO [101] and RefCOCO+ [101] are collected by the two-player game proposed in ReferitGame [34]. There are two test splits called "testA" and "testB". Images in "testA" only contain multiple people annotation. In contrast, images in "testB" contain all other objects. Expressions in RefCOCOg [62] are collected on Amazon Mechanical Turk in a non-interactive setting. Thus, the expressions in RefCOCOg are longer and more complex. RefCOCOg has "google" and "umd" splits. **The "google" split** does not have a public test set,

Table 8: The detailed statistics of RefCOCO [101], RefCOCO+ [101], RefCOCOg [62], ReferItGame [34] and Flickr30K Entities [68] datasets. We represent test split and testA split in the same column.

Dataset	Images	Instances	total queries	train queries	val queries	test(A) queries	testB queries
RefCOCO [101]	19,994	50,000	142,210	120,624	10,834	5,657	5,095
RefCOCO+[101]	19,992	49,856	141,564	120,191	10,768	5,726	4,889
RefCOCOg [62]	25,799	49,822	95,010	80,512	4,896	9,602	–
ReferItGame[34]	20,000	19,987	120,072	54,127	5,842	60,103	–
Flickr30k [68]	31,783	427,000	456,107	427,193	14,433	14,481	–

Table 9: Comparison of datasets used in the pre-trained models of the comparable methods.

Pretrained model	Uni-modal image / Data size	Uni-modal text / Data size	Image-text pairs / Data size	Total
CLIP [70]	–	–	LAION-400M [75] / 400M	400M
UniTAB [86]	image-text pairs: CC3M[8], *etc.* ; image-text-box pairs: COCO, VG, RefC, O365, SUB, *etc.* from multiple downstream task			>20M
OFA [85]	ImageNet-21K,*etc.* / 40M	filtered BookCorpus[109]*etc.* / 140GB	CC12M[8], SBU[65], COCO[50], VG[39],*etc.* / 21M	40M+140GB+21M
EVA-G [21]	–	–	Merged-2B(LAION-2B[75]+COYO-700M) / >2B	>2000M
FlanT5 [21]	–	filtered crawl data / >750GB	–	>750GB
ONE-PEACE [86]	image-text pairs: LAION-2B [75]; audio-text pairs: 2.4M + 8000 hours			>2000M
BEiT-3 [87]	ImageNet-21K[40] / 14M	BookCorpus[109],*etc.* / 160GB	CC12M[8], SBU[65], COCO[50], VG[39],*etc.* / 21M	14M+160GB+21M

and **exists an overlap between the training and validation image sets.** The “umd” split does not have this overlap. Therefore, **to prevent data leakage of the test set** and following previous studies [79, 102], **we exclude the “google” split in the fine-tuning settting and dataset-mixed pre-training setting**. Thus, we trained and tested the RefCOCOg dataset only on the “umd” split.

ReferItGame. ReferItGame [34] (short as ReferIt) belongs to the Phrase Grounding (PG) task, which contains images from SAIAPR12 [20] and collects expressions through a two-player game. In this game, the first player is shown an image with an object annotation and is asked to write a natural language expression referring to the object. The second player is then shown the same image along with the written expression and is asked to click on the corresponding area of the object. If the clicking is correct, both players receive points and swap roles. If not, a new image will be presented.

Flickr30k Entities. Flickr30k Entities (short as Flickr30k) [68] belongs to the phrase grounding task, which contains images in Flickr30k dataset. The query sentences are short noun phrases in the captions of the image. The queries are simpler and easier to understand compared to RefCOCO/+/g. Therefore, the ambiguity of the expression is heightened simultaneously, resulting in a relative increase in noise.

B.2 Explanation of the dataset abbreviations

In Tab. 2 of the main text, we provide abbreviations for the datasets used in intermediate pre-training. Specifically, ‘GoldG’ (proposed in MDETR [33]) is a mixed region-level fine-grained dataset created by combining three datasets - Flickr30k [68], MS COCO [50], and Visual Genome [39] - along with annotated text data for detection, REC and QGA tasks. It has a size of approximately 6.2M. ‘O365’ refers to the Object365 [105] dataset, ‘SBU’ stands for SBU caption [65], ‘VG’ represents the Visual Genome [39] dataset, and ‘OI’ stands for OpenImage [42] dataset.

B.3 Comparison of datasets used in the pre-trained models

As presented in Tab. 9, we conducted an analysis of the datasets employed by the backbone models compared in Tab. 2 within the dataset-mixed pre-training setting. From the Tab. 9, it is evident that BEiT-3 and OFA utilize comparable datasets for pre-training. Conversely, other compared works in Tab. 2, such as Shikra [10], Ferret [100], LION [9], and other models, such as ONE-PEACE [86] (a tri-modality foundation model), employ significantly larger amounts of data than BEiT-3. Consequently, our method does not possess any advantage concerning the volume of data used in pre-training.

C Implementation details

Network Architecture. The detailed network structure of our framework is shown in Tab. 10. We employ BEiT-B/16 and BEiT-L/16 as the backbone for our OneRef base and large version,

Table 10: Network structure of our proposed OneRef framework.

Model	Backbone	Input resolution	One-tower Transformer layers	dimension	heads	All parameters (include all MoE heads)
OneRef-B	BEiT-B/16	384	12	768	12	234M (REC), 267M (RES)
OneRef-L	BEiT-L/16	384	24	1024	16	639M (REC), 679M (RES)

Table 11: Hyperparameters of our framework during training. *lr.* denotes the learning rate.

Item	Value base model	large model
optimizer	AdamW	
Epoch for MRefM pre-training	110	
lr. for MRefM pre-training	0.5×10^{-4}	
weight decay	0.5×10^{-5}	
patch size	16×16	
Initial value of aspect ratio a in MIM	0.3	
mask ratio β in Referring MIM	0.75	
mask ratio γ in Referring MIM	0.35	
mask ratio δ in Referring MLM	0.40	
λ_{reg} and λ_{img} in Referring MLM	1, 1	
batch size for MRefM pre-training	32	8
Epoch for REC/RES transfer	20	
lr. for REC/RES transfer	0.3×10^{-4}	
λ_{l_1}, λ_{giou} in REC	2, 2	
λ_f, λ_d in REC	20, 2	
λ_{seg_f}, λ_{seg_d} in RES	20, 2	
batch size for REC/RES transfer	32 / 16	8 / 6

respectively. In the structure of OneRef-B, the one-tower encoder are 12-layer Transformers with the hidden embedding dimension of 768. In the structure of OneRef-L, the one-tower encoder is 24-layer Transformers with the hidden embedding dimension of 1024. The one-tower encoder encodes both textual and visual modalities. Due to the utilization of a 3-layer deconvolution, RES exhibits a slightly higher number of model parameters compared to the REC task.

Training Details. The batch size for pre-training the base model and large model are (32, 8), while they are (32, 8) and (16, 6) for transferring to the REC and RES tasks, respectively. Our model is optimized end-to-end by using the AdamW optimizer and a cosine learning scheduler with an initial learning rate of 0.5×10^{-4} for 110 epochs during the pre-training stage. During REC/RES transfer stage, the learning rates is 0.3×10^{-4} with 20 epochs. The framework and experiments in our study were conducted using PyTorch. For MRefM pre-training, the base model took 15 hours on 32 NVIDIA A100 GPUs, while the large model took 50 hours on the same number of GPUs. As for REC/RES transfer fine-tuning training, it took an average of 3 hours for the base model and 8 hours for the large model to process one dataset on 8 A100 GPUs.

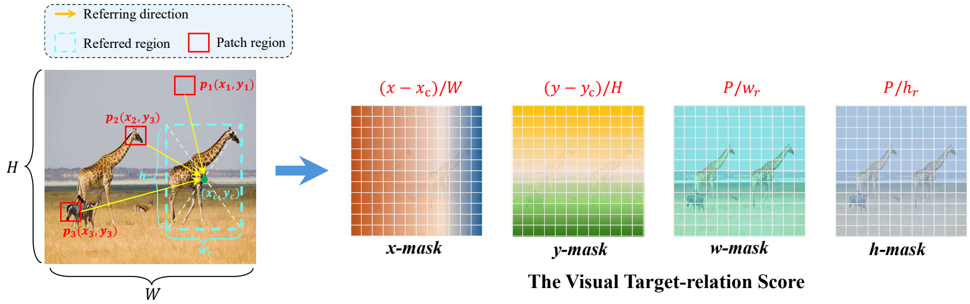
Inference Details. Unlike previous methods, such as TransVG++ [15], QRNet [98], *etc.* , which heavily rely on high-resolution images like 640×640, we adopt smaller resolution of 384×384. To ensure compatibility, we employ a long edge alignment and short edge pad filling scheme to the image. We include `[SEP]` and `[EOS]` token at the beginning and the end of the input text, and align it to a fixed length of 64 by padding empty tokens.

Model Hyperparameters. We summarize and report the hyperparameter settings of the OneRef framework in Tab. 11.

D Technical remarks

D.1 Further explanation for the effectiveness mechanism of the visual target-relation score

(1) The purpose of designing the Referring MIM algorithm. In the existing MIM paradigm, reconstruction is limited to solely relying on the visual features within the image. To enhance content reconstruction by leveraging cross-modal information as much as possible, our Referring MIM approach incorporates visual target-relation scores alongside visual modality content during reconstruction. This modeling approach presents increased difficulty as it necessitates reliance on



Referring text: " the giraffe walking while another giraffe follows it. "

Figure 5: The reconstruction of the visual target-relation score $\boldsymbol{s}^{vt} \in \mathbb{R}^{N_v \times 4}$. (x, y) represents the coordinates of a general image patch, and P is the patch size. 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.

textual information for reconstructing the two visual branch. Consequently, our model achieves a more comprehensive understanding of both visual and textual information. In this way, the model not only can perceive the information of the image modality itself but also have a more accurate understanding of the location and correlation of the key object features in different regions.

(2) How and why the visual target-relation score (*i.e.*, the *x*-, *y*-, *w*-, and *h-masks*) works. We provide a clearer illustration in Fig. 5 for further explanation. As mentioned in Sec. 3.2, this score represents the spatial distance between the current patch region and the referred region, it enables implicit deployment of grounding capability within each token of the model. When reconstructing the visual features and target-relation score of each local patch, the model actually needs to have an global and comprehensive understanding of the text modality information and the visual information. On this basis, the model needs to rely on the reconstructed visual features of the local patch to implicitly predict the specific location and size of the referred object, and then accurately predict the visual target-relation score. Finally, Referring MIM can enhance the model's global and multimodal understanding of textual and visual information, and then learn more general visual representations, which can have better generalization ability when deployed to downstream referring tasks.

The proposed Referring MIM is our own design, which is mainly used to improve the defects existing in MAE [27]/BEiT [4]. We can find the rationale of our method in some classic computer vision works, such as the YOLO series works [71], which predicts the location, size, confidence, and category of the object box corresponding to each grid cell based on the global understanding of the image. YOLO *etc*. [71] also confirmed that the object detection model obtained in this way has stronger generalization ability when transfer to detection tasks that differ greatly from the training data compared with other detectors.

D.2 The selection of the unsupervised regions

The process of selecting unsupervised regions bears resemblance to weakly-supervised visual grounding. Drawing inspiration from ALBEF's method [44] for weakly-supervised grounding, we employ a BEiT-3 model with performed image-text contrastive tuning to encode both the image and text, thereby obtaining a cross-modal text-to-image attention map for selection. Subsequently, leveraging the cross-modal attention and modular parsing of textual sentences provided by MAttNet [102] enables us to derive scores for each proposal. Finally, we select the region with the highest score as our objective in Referring MRefM.

D.3 Referring-aware text masking

In referring MLM, we utilize a referring-aware text masking strategy. Specifically, we preferentially mask out the referential subject of the expression text on the basis of a random mask, and the subject is obtained by the NLP parsing tool (*e.g.*, spaCy). Since this small technical point does not observe a significant performance gain as the referring-aware dynamic image masking strategy, we do not provide additional ablation experiments.

Table 12: Comparison with **latest** SoTA methods for PG task with dataset-mixed intermediate pre-training setting. 'RefC' represents the mixup of RefCOCO/+/g training data. † indicates RefC has been used during pre-training.

Methods	Venue	Visual/Language Backbone	Intermediate pretrain data	Data size	ReferIt test	Flickr test
Dataset-mixed intermediate pre-training setting						
MDETR † [33]	ICCV'21	RN101/RoBERT-B	GoldG,RefC	6.5M	–	83.80
YORO† [29]	ECCV'22	ViLT [37] / BERT-B	GoldG,RefC	6.5M	71.90	–
UniTAB † [97]	ECCV'22	RN101/RoBERT-B	VG,COCO,*etc.*	>20M	–	79.38
HiVG-B† [92]	ACMMM'24	CLIP-B / CLIP-B	RefC,ReferIt,Flickr	0.8M	77.75	82.08
HiVG-L† [92]	ACMMM'24	CLIP-L / CLIP-L	RefC,ReferIt,Flickr	0.8M	78.16	82.63
OneRef-B† (0.2B)	NeurIPS'24	BEiT3-B / BEiT3-B	RefC,ReferIt	0.5M	79.66	84.01
OneRef-L† (0.6B)	NeurIPS'24	BEiT3-L / BEiT3-L	RefC,ReferIt	0.5M	**83.22**	**85.13**

Table 13: Comparison with **latest** SoTA methods (**oIoU** metric) on the three datasets for **RES** task with both single-dataset fine-tuning setting and dataset-mixed intermediate pre-training setting. † indicates RefC has been used during pre-training.

Methods	Venue	Visual/Language Backbone	Intermediate pretrain data	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test
Single-dataset fine-tuning setting *w.* uni-modal pre-trained close-set segmentation model: (traditional setting)											
LAVT [94]	CVPR'22	Swin-B / BERT-B	–	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
Single-dataset fine-tuning setting *w.* vision-language self-supervised pre-trained model:											
RISCLIP-B [35]	NAACL'24	CLIP-B / CLIP-B	–	73.57	76.46	69.76	65.53	70.61	55.19	64.10	65.09
RISCLIP-L [35]	NAACL'24	CLIP-L / CLIP-L	–	76.92	80.99	73.04	71.24	76.99	61.56	67.96	68.71
OneRef-B (ours)	NeurIPS'24	BEiT3-B / BEiT3-B	–	77.55	80.96	73.53	70.82	74.53	64.06	70.68	70.61
OneRef-L (ours)	NeurIPS'24	BEiT3-L / BEiT3-L	–	**80.48**	**82.78**	**78.27**	**74.25**	**78.41**	**69.85**	**74.91**	**77.36**
Dataset-mixed intermediate pre-training setting:											
HIPIE† [88]	NeurIPS'23	RN50,CLIP / BERT-B	RefC,O365,PACO	78.30	–	–	66.20	–	–	69.80	–
UNINEXT† [93]	CVPR'23	RN50 / BERT-B	RefC,O365,SOT	77.90	79.68	75.77	66.20	71.22	59.01	70.04	70.52
PolyFormer-B†[53]	CVPR'23	Swin-B / BERT-B	RefC	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05
PolyFormer-L†[53]	CVPR'23	Swin-L / BERT-B	RefC	75.96	78.29	73.25	69.33	74.56	61.87	69.20	70.19
OneRef-B† (ours)	NeurIPS'24	BEiT3-B / BEiT3-B	RefC	81.06	83.05	77.80	72.24	77.32	67.08	75.14	77.21
OneRef-L† (ours)	NeurIPS'24	BEiT3-L / BEiT3-L	RefC	**81.95**	**83.62**	**79.57**	**74.55**	**79.54**	**70.65**	**76.53**	**79.18**

D.4 The difference of the task heads between ours with other frameworks

Recently, several multi-task visual grounding studies [79, 45] have incorporated both grounding and segmentation task heads into their frameworks. Most relevance to our work is VG-LAW [79], which simplifies the implementation of grounding and segmentation heads by eliminating Transformer-based fusion encoders through visual adaptive weights generation. In contrast, for REC headers, we propose a box mask constraint based on cross-modal cosine similarity that significantly enhances the accuracy of such grounding approach. For the RES head, instead of employing adaptive weights generation, we directly obtain segmentation masks using cosine similarity for the visual tokens upsampled by a 3-layer deconvolution.

E Extra experimental results

E.1 The results on ReferIt and Flickr30k dataset under mixup pre-training setting

The results of our framework on the PG task (*i.e.*, ReferIt [34] and Flickr30k [68] datasets) under the mixup pre-training setting are presented in Tab. 12. It is worth noting that a majority of studies conducted under this setting have not provided these results, thus only several works are included in the table. As shown in Tab. 12, our base model outperforms HiVG by 1.91% and 1.93% on the two datasets, and also achieves SoTA performance.

E.2 The results for RES task under oIoU metric

The results for RES task under oIoU metric are presented in Tab. 13. oIoU is calculated as the ratio between the total intersection area and the total union area of all test samples, each of which is consists of both a text query and an image. This metric particularly favors larger objects. As indicated in Tab. 13, (1) in the single-dataset fine-tuning setting, our base model outperforms RISCLIP [35] by 3.77%, 8.87%, and 5.52% on RefCOCO/+/g testB/testB/test split, respectively. (2) Similarly, in the

Table 14: Comparison of the computational cost in REC task. The results are obtained on RefCOCO dataset. The testing environment is 1 NVIDIA A100 GPU. † indicates that the model's code is not publicly available, and the replicated estimation results are shown. The backbone parameters of our UniRef model only include the actual calculated parameters, specifically those of the V-L expert head in MoE, while excluding the parameters of unused visual and language expert heads and their uni-modal branches. We highlight the best result in **bold**. (*FPS: images / (GPU · second)*)

Model	Backbone param.↓	Fusion+head param.↓	Total param.↓	FLOPs (G)↓	Fine-tune FPS↑	Test FPS↑	testA time↓	testA Acc.↑
TransVG [14]	150M	21M	171M	214	22.8	59.6	95s	82.7
QRNet [98]	252M	21M	273M	540	9.4	50.9	111s	85.9
TransVG++† [15]	161M	10M	171M	396	2.6	8.7	644s	88.4
MDETR [33]	150M	135M	185M	642	4.7	19.9	283s	89.6
Grounding-DINO [55]	156M	15M	172M	464	–	8.3	681s	91.8
UniRef (Ours)	**147M**	**1.7M**	**149M**	**162**	**55.8**	**83.2**	**68s**	**94.3**

Table 15: Complete ablation study of MRefM using our OneRef-base model in REC task on both single-dataset fine-tuning setting and mixup intermediate pre-training setting.(Acc@0.5(%))

MIM	MLM	image masking strategy	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test
Single-dataset fine-tuning setting:										
✗	✗	✗	85.23	88.13	83.82	78.56	83.36	71.72	80.41	80.52
vanilla	vanilla	block-wise	85.75	88.86	83.43	78.47	84.27	71.66	81.06	81.19
Ref MIM	Ref MLM	referring-aware	**88.75**	**90.95**	**85.34**	**80.43**	**86.46**	**74.26**	**83.68**	**83.52**
Dataset-mixed intermediate pre-training setting: (main)										
✗	✗	✗	85.23	88.13	83.82	78.56	83.36	71.72	80.41	80.52
vanilla	vanilla	random	86.60	89.86	84.96	79.68	84.59	72.11	81.35	81.11
vanilla	vanilla	referring-aware	86.71	90.58	85.33	80.06	85.77	73.96	81.96	82.16
Ref MIM	vanilla	referring-aware	88.86	92.12	86.89	83.64	88.26	76.58	83.55	85.86
vanilla	Ref MLM	referring-aware	87.26	91.68	86.37	81.52	86.87	75.93	82.88	84.32
Ref MIM	Ref MLM	random	90.56	93.55	88.23	85.08	89.12	78.56	85.57	86.89
Ref MIM	Ref MLM	block-wise	90.07	93.32	88.21	84.55	88.83	77.98	84.71	86.69
Ref MIM	Ref MLM	referring-aware	**91.89**	**94.31**	**88.58**	**86.38**	**90.38**	**79.47**	**86.82**	**87.32**

dataset-mixed intermediate pre-training setting, our base model surpasses UNINEXT [93] by 2.03%, 8.07%, and 6.69% on RefCOCO/+/g testB/testB/test split, respectively. Furthermore, our large model exhibits remarkable scalability with additional performance enhancements.

E.3 Computational costs analysis compared with SoTA methods

In this paper, we highlight two significant advantages of our model architecture over other frameworks: (a) Instead of using a Transformer to fuse visual and language features, we only employ a simple lightweight task head; (b) Our one-tower architecture eliminates the need for early interaction techniques in the backbone network, thereby reducing the computational complexity of the model.

We compare the energy efficiency of our model with several well-known SoTA works on the REC task from various perspectives, including the number of parameters, computational complexity (FLOPs), inference speed (FPS), and test time (s). As can be seen from Tab. 14, due to the simplification of our model's structure, the number of parameters and the calculation complexity are significantly lower than other well-known models. Specifically, our feature fusion and grounding head module only require 1.7M parameters, while other methods use 20M, meaning we only have about 8.5% of their parameter count. Additionally, our computation is only 34.9% of Grounding-DINO and 25.2% of MDETR. Moreover, our inference speed is 10 $\times$ faster than Grounding-DINO and TransVG++ (the speed also related to the image size used by the model). Despite these advantages, thanks to the modality-shared feature space, we outperform all these well-known works.

E.4 Complete ablation study of MRefM on single-dataset fine-tuning and mixup pretrain settings

The complete ablation results of MRefM on both single-dataset fine-tuning and mixup pretrain settings are provided in Tab. 15, which serves as a supplement to Tab. 4 in the main text. In the table, the masking ratio is set to 0.4 when using block-wise or random masking strategies.

Table 16: Ablation study of the mask ratio in referring-aware dynamic masking strategy on RefCOCOg(val) dataset.

mask ratio β	mask ratio γ	RefCOCOg Acc@0.5(%)	mask ratio β	mask ratio γ	RefCOCOg Acc@0.5(%)
0.20	0.75	83.40	0.35	0.50	85.98
0.30	0.75	85.01	0.35	0.60	86.32
0.35	**0.75**	**86.82**	0.35	0.70	86.62
0.40	0.75	86.62	**0.35**	**0.75**	**86.82**
0.50	0.75	86.23	0.35	0.80	86.19
0.60	0.75	84.72	0.35	0.85	85.56
0.70	0.75	83.39	0.35	0.90	84.87

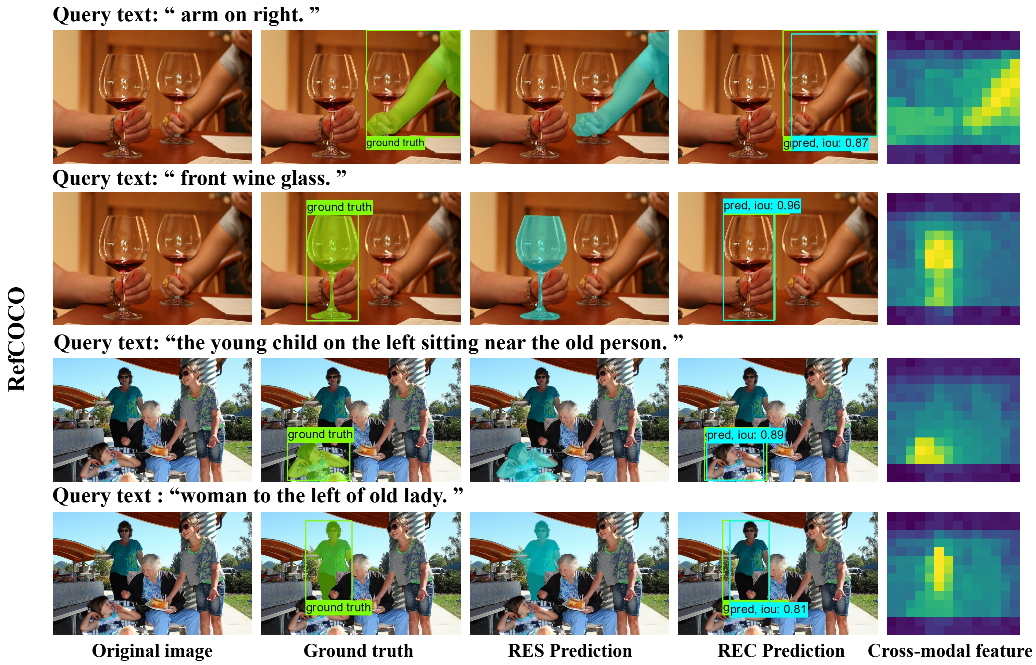


Figure 6: Qualitative results of our OneRef framework on the RefCOCO-val split. Each example shows two different query texts. From left to right: the original input image, the ground truth with box and segmentation mask (in green), the RES prediction of OneRef (in cyan), the REC prediction of OneRef (in cyan), and the cross-modal feature.

E.5 Ablation study of the mask ratio in referring-aware dynamic masking strategy

As depicted in Tab. 16, we conducted ablation experiments on the mask ratio within our proposed referring-aware dynamic image masking strategy. It is observed that while a high mask ratio of 0.75 is employed for the pixel reconstruction of MAE [27], achieving better results for BEiT's feature reconstruction requires a mask ratio ranging from approximately 0.4 to 0.45. In our proposed approach, favorable outcomes can be attained by setting β and γ to 0.35 and 0.75, respectively; where β represents the mask ratio beyond the referred region and γ denotes the mask ratio within it. Experimental statistics show that our entire mask rate α in each sample is about $0.4 \sim 0.5$.

F Visualization of the results

As shown in Fig. 6, Fig. 7, and Fig. 8, we present the qualitative grounding and referring segmentation results with several relatively challenging examples. Each example shows two different query texts. The cross-modal features are obtained by the cosine similarity between the `[SEP]` language token and the vision tokens on REC transfer model of OneRef-B. These results demonstrate the strong semantic comprehension capability of our OneRef model in complex text understanding and cross-modal grounding.

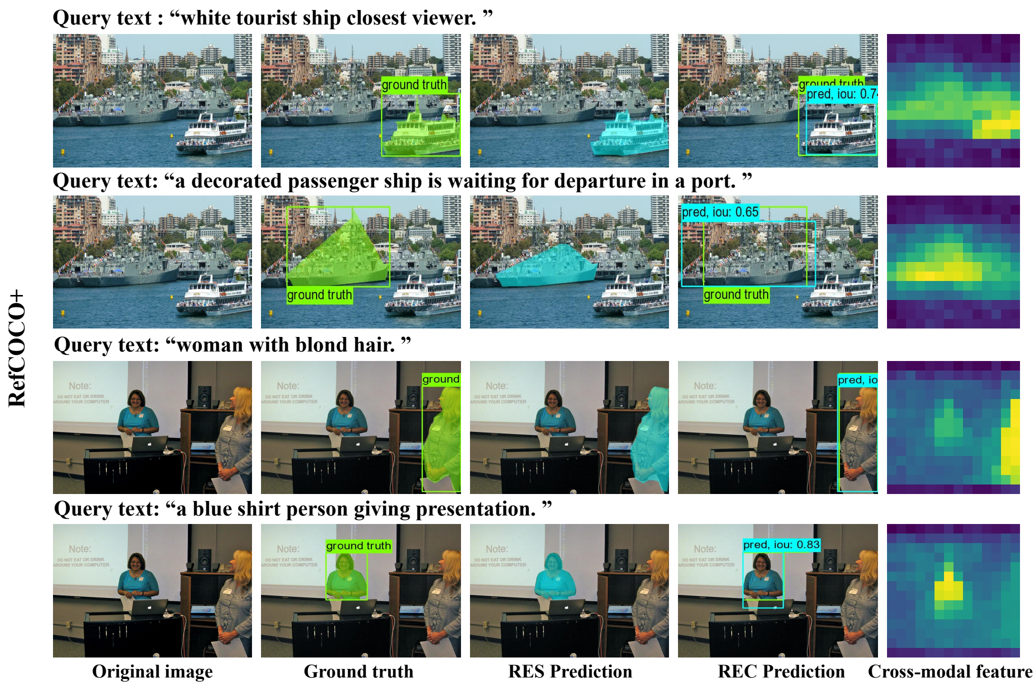


Figure 7: Qualitative results on the RefCOCO+-val dataset. The annotation is the same as Fig. 6.

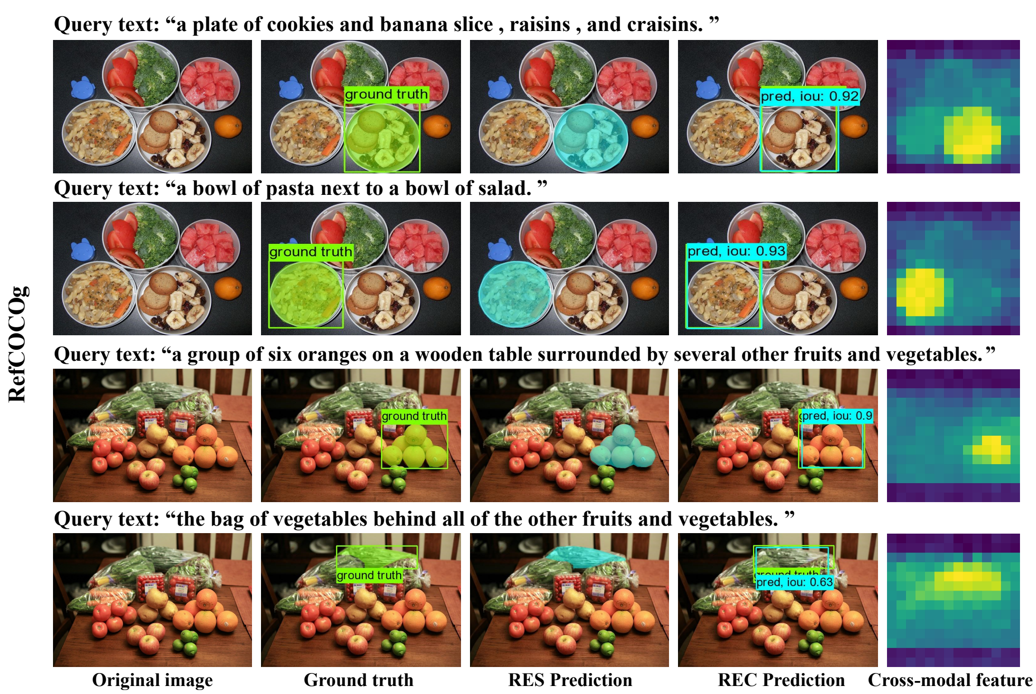


Figure 8: Qualitative results on the RefCOCOg-val dataset. The annotation is the same as Fig. 6.

G Further discussions

G.1 Limitations

Firstly, despite achieving remarkable grounding and segmentation results, the pre-training in this paper solely relies on the comparatively limited RefC dataset, as opposed to other studies with larger datasets.

Secondly, the MRefM paradigm necessitates additional referential bounding boxes as supervised data compared to self-supervised pre-training. Therefore, we explore the potential of unsupervised pre-training for MRefM. However, when utilizing image-text pairs obtained from web crawling, there is no guarantee that the referred regions will exhibit strong correlation with the text due to many texts describing the entire image. This aspect introduces certain challenges and biases during large-scale pre-training of MRefM. Consequently, this paper should serve as an inspiration for subsequent researchers to propose more convenient plug-and-play modeling methods.

G.2 Broader impacts

OneRef demonstrates strong grounding and referring segmentation capabilities, while MRefM represents a novel modeling paradigm for referential relationship. This facilitates users to easily utilize our model (*e.g.*, OneRef-L) for their own needs by simply providing some appropriate text queries. However, this also raises concerns about how our OneRef models with a strong understanding capabilities could be used inappropriately in the community, such as for large-scale illegal video surveillance. The open-set grounding capabilities could be manipulated through specialized textual cues to facilitate targeted detection or human tracking instead of generic ones. This manipulation could introduce biases in the detector and result in unfair predictions.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have discussed the limitations of this work in the Appendix G.1.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: We have provided the full set of assumptions and a complete proof in Sec. 3 of the main text.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: We have fully disclosed all the information needed to reproduce the main experimental results of the paper in Sec. 4 of the main text and Appendix C.

   Guidelines:

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

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

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

Answer: [No]

Justification: We answer No mainly for the following acceptable reasons: (1) The data we use are all publicly available and have been detailedly introduced in the paper. Researchers can acquire the data according to the provided reference information. (2) Due to time constraints, we were unable to compile and submit the anonymous code at the time of submission. However, as stated in the abstract, all models and code will be promptly made public after the decision is reached on this paper. (3) The implementation details of our work have been thoroughly explained in both the main text and supplementary materials. Even without publicly available code, researchers can reproduce it based on the information given in this paper.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We have specified all the training and test details in the Sec. 4 of the main text and Appendix B.1.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [No]

   Justification: The error bars are not reported because it would be too computationally expensive. Besides, it is not crucial for interpreting the experimental results in this task topic.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources in Appendix C.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We have conformed in every respect with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

   Answer: [Yes]

Justification: We have discussed both potential positive societal impacts and negative societal impacts of the work performed in Appendix G.2.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: Our paper poses no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We have cited the original paper (*e.g.*, BEiT-3 [87]) that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.