Vision-Aware Text Features in Referring Image Segmentation: From Object Understanding to Context Understanding – Supplementary Materials –

Our supplementary has 5 sections. Section 1 shows additional information about datasets and training procedure. Section 2 explains how spaCy is used to extract main noun phrases from sentences and also explains how potential LLMs can be used to create diverse object descriptions to improve dataset annotations. Section 3 contains the additional experiments on RefCOCO(+/g), Ref-Youtube-VOS and Ref-DAVIS17. This section also illustrates and analyzes the performance of CLIP Prior, CMD and MCC in different situations as well as the runtime and.

1. Additional Implementation Details

1.1. Datasets

Image datasets. RefCOCO and RefCOCO+ [5] are two of the largest image datasets used for referring image segmentation. They contain 142,209 and 141,564 language expressions describing objects in images. RefCOCO+ is considered to be more challenging than RefCOCO, as it focuses on purely appearance-based descriptions. G-Ref [16], or RefCOCOg, is another well-known dataset with 85,474 language expressions with more than 26,000 images. The language used in G-Ref is more complex and casual, with longer sentence lengths on average.

Video datasets. Ref-YouTube-VOS [17] and Ref-DAVIS17 [6] are well-known datasets for referring video object segmentation. Ref-YouTube-VOS contains 3978 video sequences with approximately 15000 referring expressions, while Ref-DAVIS17 consists of 90 high-quality video sequences. These datasets are used to evaluate the performance of algorithms that aim to identify a specific object within a video sequence based on natural language expressions.

1.2. Metrics

In our work, we use mIoU and Precision@X to evaluate our method for image datasets, while $\mathcal{J}\&\mathcal{F}$ are used as evaluation metrics for video datasets. mIoU stands for mean Intersection over Union, which measures the average overlapping between the predicted segmentation masks and the ground truth annotations. Precision@X, on the other hand, measures the success rate of the referring process at a specific IoU threshold, and it focuses on the referring capability of the method.

In addition, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ are commonly used evaluation metrics for video object segmentation (VOS) datasets. The $\mathcal{J}$ is similar to the IoU score, while the $\mathcal{F}$ score is the boundary similarity measure between the boundary of the prediction and the ground truth. These two metrics together measure the performance of the predicted object mask over the entire video sequence. Higher $\mathcal{J}\&\mathcal{F}$ score indicates better RVOS performance.

Furthermore, to quantify the ability to consistently segment various expressions for the same object and further validate the effectiveness of our proposed Meaning Consistency Constraint, we leverage an Object-centric Intersection over Union (Oc-IoU) score, which calculates the overlap and union area between ground truth and all segmentation predictions of the same object. Specifically, consider the i-th object with K_i expressions referring to that object and the corresponding ground truth mask GT_i. Let P_i^j be the model's prediction for the j-th expression of the i-th object, where $j = \overline{1..\ K_i}$. The Object-centric IoU can be formulated as follows:

$$\text{Oc-IoU}(\mathrm{GT}_i, \mathrm{P}_i) = \frac{\mathrm{GT}_i \cap \mathrm{P}_i^1 \cap ... \cap \mathrm{P}_i^{\mathrm{K}_i}}{\mathrm{GT}_i \cup \mathrm{P}_i^1 \cup ... \cup \mathrm{P}_i^{K_i}}, \tag{1}$$

$$\text{Oc-IoU}_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} \text{Oc-IoU}(\mathrm{GT}_i, \mathrm{P}_i), \tag{2}$$

where N is the total number of objects/instances in the datasets.

1.3. Training Details

Our model is optimized using AdamW [15] optimizer with the initial learning rate of 10^{-5} for the visual encoder and 10^{-4} for the rest. Our model comprises a total of nine Masked-Attention Transformer Decoder layers followed [1]. We set the number of queries to 5 [22]. For

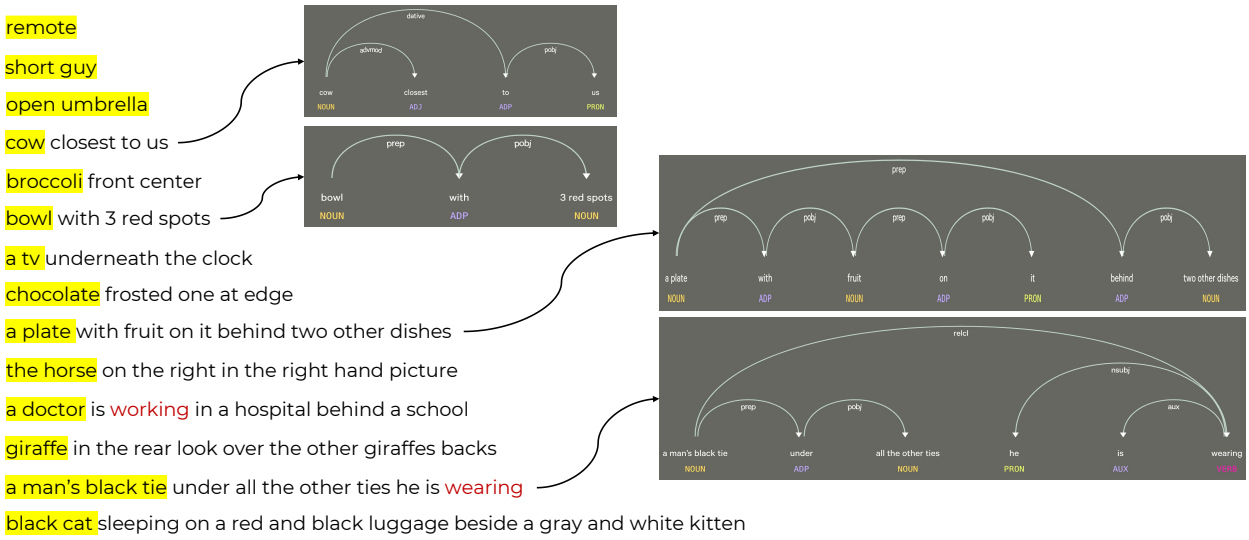


Figure 1. Examples of our main object extractor output. Given the expression, our algorithm will output the main noun phrase in the sentence. Typically, the root word of the sentence is a noun phrase, which we directly output as the main noun phrase. However, if the root word is not a noun phrase (*e.g.* working, wearing in the image), we instead focus on identifying its child noun. Additionally, we illustrate the dependency parsing tree for some representative sentences on the right.

the setting of training from classification weight from Imagenet on Ref-Youtube-VOS dataset, we train the model for $200,000$ iteration with the learning drop at $140,000$-th iteration. On Ref-DAVIS17 [6], we directly report the results using the model trained on Ref-YouTube-VOS without fine-tuning. In terms of coefficients in loss function, $\gamma_{cls} = 2$ and $\gamma_{mask} = 5$ are followed from Mask2Former. To maintain balance, we then choose $\gamma_{mcc} = 2$. We want to prioritize the mask loss with the highest weight because the IoU is the primary metric.

2. Additional Details of VATEX

2.1. Main Object Extractor

We use spaCy [3] to implement our main object extractor, leveraging its optimized, fast, and effective dependency parsing capabilities. First, spaCy extracts the root word of the sentence, also known as the head word, which has no dependency on other words (i.e., it has no parent word in the dependency tree). If this root word is a noun phrase, we directly output it as the main noun phrase of the sentence. If the root word is not a noun (e.g., a verb), we focus on its child noun to ensure it centers on the described object. Figure 1 shows some examples on the datasets and shows the output of our algorithm as well as the dependency parsing tree of some representative cases.

To handle complex sentence structures that lack a directly related noun phrase, we have implemented a rollback mechanism (in L27 of vatex/utils/noun_phrase.py) that returns the whole sentence, preventing information loss and mitigating potential errors from inaccurate main noun phrase extraction. As shown in Table 1, this rollback mechanism helps avoid poorly extracted nouns that could potentially cause incorrect segmentation masks.

Table 1. Rollback stats on the validation split of three RIS datasets.

Dataset	RefCOCO	RefCOCO+	G-Ref
Num expressions	10,834	10,758	4,896
Rollback rate(%)	10.7	10.8	3.1
mIoU w/o rollback	76.23	68.45	69.01
mIoU w. rollback	78.16	70.12	69.73

2.2. Enhancing Expression Diversity in Referring Image Segmentation Datasets through Prompting Techniques

Our method's utilization of diverse referring expressions for each object aligns with established best practices in text-image dataset annotation. This approach is widely accepted and implemented across several benchmark datasets. In scenarios where multiple expressions per object are unavailable, we have the flexibility to employ Large Language Models (LLMs) for enhancing expression diversity. This can be achieved either by augmenting existing expressions or generating new ones based on object masks, a technique successfully employed by datasets like RIS-CQ. Furthermore, we demonstrate a practical application of this approach through a sample that showcases how we can prompt

Table 2. Universality of VATEX. We conduct experiments to plug-and-play CLIP Prior and MCC in ReLA. † means we run experiment on their official code to get the mIoU score.

Method	RefCOCO	G-Ref
ReLA†	73.16	63.64
ReLA + CLIP Prior	74.32 +1.16	65.76 +2.12
ReLA + MCC	75.46 +1.16	65.12 +1.48
ReLA + CLIP Prior + MCC	76.33 +3.17	67.69 +4.05

ChatGPT to generate relevant expressions in Figure 2. This generation is based on factors like an object's position in the image, its relative position to other objects or people, and distinguishing attributes such as color or appearance.

Figure 2 showcases two innovative prompting techniques for generating object descriptions. On the left, we demonstrate how combining an original image with its masked version can effectively prompt GPT-4 to generate detailed descriptions. The right side of Figure 10 highlights the application of the SOTA 'Set of Mark' (SoM[1] [24]) technique to enhance the capability of GPT-4(V) in acquiring deeper knowledge. SoM involves creating masks for each object in the image using SAM, each distinguished by a unique identifier. This marked image then serves as an input for GPT-4V, enabling it to respond to queries necessitating visual grounding with greater accuracy and relevance.

3. Additional Results and Analysis

3.1. Universality of VATEX

VATEX employs CLIP Prior for Object Understanding and Meaning Consistency Constraint for Context Understanding. These two modules can be easily integrated into any DETR-based model (*e.g.* ReLA [11]) for RIS. We took ReLA as a representative work and reproduced the performance of ReLA on the validation sets of the RefCOCO and G-Ref datasets using mIoU metrics. As illustrated in Table 2, VATEX seamlessly integrates into current models, achieving significant performance gains of 3.17% on RefCOCO and 4.05% on G-Ref. This demonstrates the effectiveness of our approach in utilizing Vision-Aware text features for both object understanding and context understanding.

3.2. Additional Comparison on RefCOCO(+/g)

3.2.1 Fair backbone comparison

We have benchmarked our model, VATEX, using the ResNet-101 backbone, aligning it with CRIS and JMCELN for a more equitable comparison, as illustrated in Table 3. This adaptation demonstrates VATEX's superior performance, achieving a 1.26% improvement on RefCOCO val and a significant 1.93% on RefCOCO testB over the current state-of-the-art methods.

Table 3. Fair Backbone Comparison between CRIS, JMCELN, LAVT and VATEX.

Method	Backbone		RefCOCO		
	Visual	Textual	val	testA	testB
CRIS [21]	ResNet-101	CLIP	70.47	73.18	66.10
JMCELN [4]	ResNet-101	CLIP	74.40	77.69	70.43
VATEX (Ours)	ResNet-101	CLIP	**75.66**	**77.88**	**72.36**
LAVT [25]	Swin-B	BERT	74.46	76.89	70.94
LAVT [25]	Swin-B	CLIP	73.15	75.24	70.02
VATEX (Ours)	Swin-B	CLIP	**78.16**	**79.64**	**75.64**

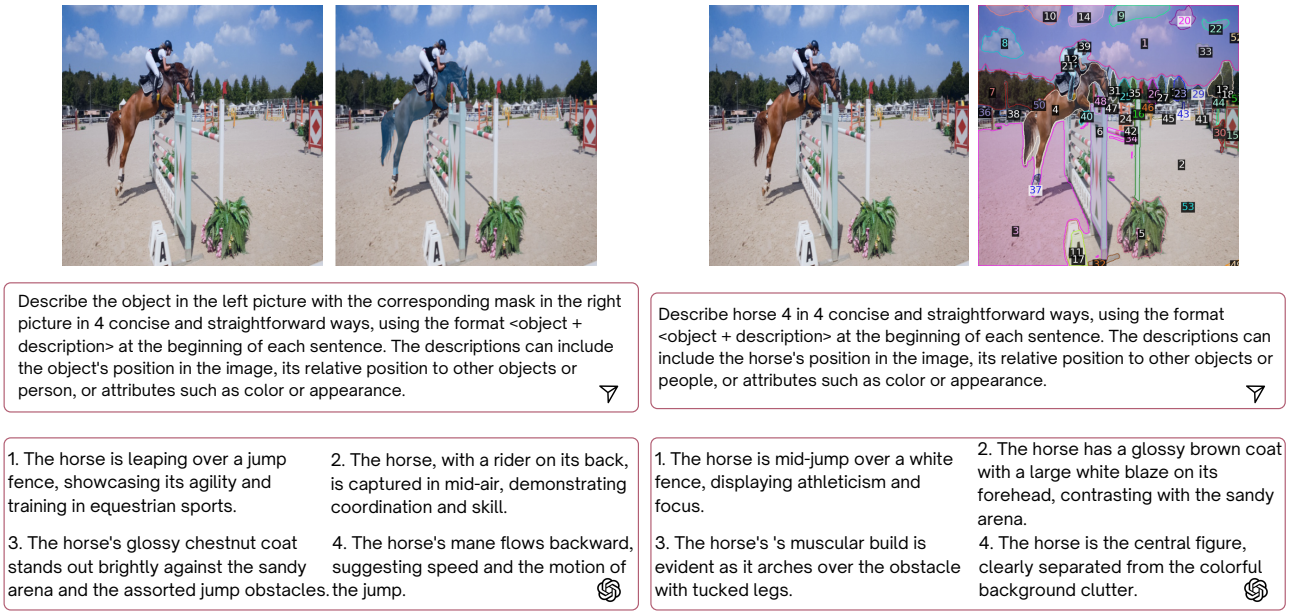
Further, to address comparisons with LAVT, we have experimented with CLIP as the text encoder, adhering to the official repository guidelines. This experiment revealed a performance decline of approximately 1% when substituting BERT with CLIP as the text encoder. This finding underscores the critical importance of using the CLIP Image Encoder together with the CLIP Text Encoder to maintain model performance. A similar trend was observed with ReferFormer, reinforcing our conclusion. Consequently, when compared to LAVT under the fair conditions in backbone, VATEX shows a substantial improvement, outperforming by 5.01%, 4.40%, and 5.62% on RefCOCO val, testA, and testB, respectively. This data confirms the effectiveness of our approach and the importance of consistent backbone usage for fair and accurate performance assessment.

3.2.2 External/Multiple Training dataset

We compare VATEX with other methods in RIS, which used external training data in Table 4. SeqTR [26], RefTR [9], and PolyFormer [12] enhance their performance on the RefCOCO dataset by incorporating external datasets—Visual Genome with 5.4M descriptions across over 33K categories, Flickr30k-entities with 158K descriptions, and the joint dataset RefCOCO(+/g) with 368K descriptions. Their papers indicate that using such external datasets for pretraining can improve performance by 8-10%.

Compared to PolyFormer [12], without using external pretraining dataset, $\text{VATEX}_{\text{RefCOCO}}$ demonstrates superior performance over PolyFormer-B, while $\text{VATEX}_{\text{RefCOCO+}}$ and $\text{VATEX}_{\text{G-Ref}}$ achieve comparable results with [12] while using **42x** and **69x** smaller datasets respectively, with the exception of the RefCOCO+ test B. The performance's gap on RefCOCO+ Test B, which focuses on non-human objects described purely by their appearance (*e.g.* "the porcelain throne," "part of the bed occupied by a black pam-

[1] https://github.com/microsoft/SoM



(a) Prompting with Mask

(b) Prompting with SoM

Figure 2. Example of using GPT-4(V) with different prompting techniques to generate object description.

phlet"), could be attributed to the varied object categories covered during the pre-training phase with extensive external datasets.

On the otherhand, VATEX$_{\text{joint}}$ adopts a different strategy. By solely utilizing the RefCOCO(+/g) dataset, which is **16x smaller** than the datasets used by PolyFormer, VATEX$_{\text{joint}}$ with Swin-B backbone still achieves remarkable results. Specifically, VATEX$_{\text{joint}}$ outperforms PolyFormer by 4-6% across all benchmarks, setting a new state-of-the-art result on the RefCOCO dataset. UNINEXT [23] and HIPIE [20], while achieving strong results, rely on extensive pretraining and data leakage in finetuning (joint training with COCO for segmentation while RefCOCO images and annotations are a subset of COCO train split). In contrast, VATEX achieves competitive performance without relying on such extensive pretraining and removes all potential data leaking in the training phase.

3.2.3 Comparison with SOTA foundation models

Table 5 illustrates the quantitative performance between VATEX with generalist foundation models: Grounded-SAM [13] [8], SEEM [28] and X-Decoder [27] in Table 5. For Grounded-SAM, we first use Grounding DINO to extract the bounding box prediction from the text prompt, then we feed that bounding box to SAM to obtain the final segmentation mask. For X-Decoder and SEEM, we directly use the report number on their official github[2] with Focal-L backbones. While VATEX is trained on much smaller dataset sizes and smaller backbones, VATEX$_{\text{joint}}$ still significantly outperforms Grounded-SAM with 14.34%, 15.65%, and 16.4% improvements on RefCOCO, RefCOCO+ and G-Ref, respectively. Compared with X-Decoder and SAM, which are trained and finetuned on RefCOCO(+/g) datasets, we also outperform them with approximately 2% with VATEX and 7.7% with VATEX$_{\text{joint}}$.

3.3. Experimental results on Ref-YoutubeVOS and Ref-DAVIS17

The result for Ref-Youtube-VOS dataset is shown in Table 6. As can be seen, our method demonstrates superior performance, setting a new state-of-the-art for referring video object segmentation on the Ref-Youtube-VOS dataset with different backbones. In particular, our approach with the spatial-temporal backbone (e.g., Video-Swin [14]) and pre-trained weights from image dataset achieves the highest $\mathcal{J}\&\mathcal{F}$ score of 65.4% among all other methods on the Ref-Youtube-VOS dataset, including VLT and ReferFormer.

The results for Ref-DAVIS17 are shown in Table 7. Similarly, our approach achieves competitive performance compared to other state-of-the-art methods in referring video object segmentation. Specifically, with backbones ResNet-50, our proposed model outperforms ReferForme

[2] https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once/

Table 4. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref datasets with other SOTA methods using external training data. VATEX is trained with Swin-B backbone

Method	External Datasets	RefCOCO			RefCOCO+			G-Ref	
		val	testA	testB	val	testA	testB	val	test
SeqTR [26]	Visual Genome (5.4M) &	71.7	73.31	69.82	63.04	66.73	58.97	64.69	65.74
RefTR [9]	Flickr30k-entities (158K) &	74.34	76.77	70.87	66.75	70.58	59.4	66.63	67.39
PolyFormer-B [12]	RefCOCO(+/g) (368K)	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
UNINEXT-H [23]	Object365 (30M) &	82.2	–	–	72.5	–	–	74.7	–
HIPIE [20]	COCO + RefCOCO(+/g)	**82.6**	–	–	73.0	–	–	75.3	–
VATEX$_{\text{RefCOCO}}$	RefCOCO (142K)	78.16	79.64	75.64	-	-	-	-	-
VATEX$_{\text{RefCOCO+}}$	RefCOCO+ (141K)	-	-	-	70.02	74.41	62.52	-	-
VATEX$_{\text{G-Ref}}$	G-Ref (85K)	-	-	-	-	-	-	69.73	70.58
VATEX$_{\text{joint}}$	RefCOCO(+/g) (368K)	81.53	**82.75**	**79.66**	**74.61**	**78.75**	**68.52**	**75.54**	**76.4**

Figure 3. Our heatmap from CLIP Prior. Naive Implementation means feeding the whole sentence through CLIP Model, without the Main Object Extractor. By reducing the complexity of the text expression, it can be seen that the activation on the object of interest becomes more accurate. Best view in zoom.

Table 5. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref validation datasets with SOTA vision foundation models.

Method	RefCOCO	RefCOCO+	G-Ref
Grounded-SAM [13] [8]	67.19	58.96	59.14
X-Decoder [27]	-	-	67.5
SEEM [28]	-	-	67.8
VATEX	78.16	70.02	69.73
VATEX$_{\text{joint}}$	**81.53**	**74.61**	**75.54**

and achieves slightly better results than RRVOS. Moreover, our method achieves the best performance among all methods with the Video-Swin-B backbone with a $\mathcal{J}\&\mathcal{F}$ score of 65.4%, which is 3.8% higher than the closest competitor VLT.

3.4. Heatmap of CLIP Prior

To obtain the heatmap result, from the vector of shape $\left(\frac{H}{16} \times \frac{W}{16} + 1, 1\right)$, we remove "CLS" token and reshape it into 2D heatmap of $\frac{H}{16} \times \frac{W}{16}$. For visualization purposes, we resize the original image to 960×960, then pass it through CLIP-Image Encoder, resulting in a high-quality heatmap of size 60×60. Notably, we only use a default input size of 224×224 during training. Regarding the quality of the heatmap, Figure 3 demonstrates the comparison between the naive implementation and our prompt-based template. In the 3rd and 7th rows, it is evident that simplifying the sentence and employing prompt templates can aid in distinguishing the target object from the image, resulting in decreased localization errors.

While CLIP Prior excels at localizing objects of inter-

Table 6. Quantitative comparison with the SOTA on Ref-Youtube-VOS.

Methods	Backbone	Ref-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Train with Image segmentation weight from RefCOCO(+/g)				
ReferFormer [22]	ResNet-50	55.6	54.8	58.4
RR-VOS [10]	ResNet-50	57.3	56.1	58.4
VATEX (Ours)	**ResNet-50**	**58.5**	**57.1**	**59.9**
ReferFormer [22]	Swin-L	62.4	60.8	64.0
VATEX (Ours)	**Swin-L**	**64.2**	**61.4**	**67.0**
ReferFormer [22]	Video-Swin-B	62.9	61.3	64.6
VLT [2]	Video-Swin-B	63.8	61.9	65.6
VATEX (Ours)	**Video-Swin-B**	**65.4**	**63.3**	**67.5**

Table 7. Quantitative comparison with the SOTAs on Ref-DAVIS17 dataset.

Methods	Backbone	Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [22]	ResNet-50	58.5	55.8	61.3
RR-VOS [10]	ResNet-50	59.7	57.2	62.4
VATEX (Ours)	ResNet-50	**61.2**	**58.2**	**64.3**
ReferFormer [22]	Video-Swin-B	61.1	58.1	64.1
VLT [2]	Video-Swin-B	61.6	58.9	64.3
VATEX (Ours)	Video-Swin-B	**65.4**	**62.3**	**68.5**

est, it can struggle in complex cases where the expression describes multiple instances within the same category and their relative positions (*e.g.* bottom right of Figure 3). In these situations, the heatmap may encompass all objects within the category rather than the specific referred instances. However, CLIP Prior's core purpose is to narrow down the relevant region, not pinpoint the exact object. Identifying the precise instance will be handled later in the full-text prompt by the Transformer architecture, which can leverage additional context and relationships.

Moreover, CLIP Prior can also help the model in cases when the referring expression contains out-of-vocabulary objects. By transferring the knowledge from CLIP and embedding the heatmap into the query initialization, the model can obtain a good segmentation mask based on the cues from CLIP Prior. Figure 4 shows how CLIP Prior heatmap can help the model to localize the object in the early phase, thus improving the model's performance.

CLIP-based model in RIS. Adopting CLIP is a good practice taken by several previous methods, including CRIS, CM-MaskSD, and RIS-CLIP. However, to effectively use the aligned embedding from CLIP to obtain good results in referring segmentation is an open question. For example, although using powerful CLIP as the backbone, the SOTA CLIP-based method RIS-CLIP [7] has a comparable performance with the SOTA Non-CLIP model VG-LAW [18]. To analyze it, we take CRIS [21] as a baseline for CLIP-based model. CRIS directly used the well-aligned embedding space between text and vision for RIS. However, the performance of this work is not good compared to others, as there are two concerns with relying solely on CLIP for referring image segmentation tasks:

1. Frozen CLIP Model. CLIP model, trained on object-centric images, generates visual features focusing on semantic class meanings rather than instance-based details (see bird example in Figure 3). This limits the effectiveness of CLIP for instance-level tasks.

2. Fine-tuning CLIP Model. Fine-tuning the CLIP model risks overfitting on training samples, thereby diminishing its ability to generalize features to novel classes.

We found that learning from a visual backbone pre-trained on ImageNet and only utilizing frozen CLIP as a prior gave better performance on both instance-level segmentation and open-vocabulary segmentation nature of RIS task.

In Table 8, for a truly fair comparison, we provide our method w/o CLIP, which achieves 75.43, 67.38, and 68.12 mIoU, and we still outperform the SOTA LAVT (74.46, 65.81, and 63.34) and VG-LAW (75.05, 66.61 and 65.36) on RefCOCO(+/g) in the same setting.

Table 8. Quantitative results of referring image segmentation on Ref-COCO, Ref-COCO+, G-Ref validation datasets on CLIP-based and Non-CLIP model.

Method	RefCOCO	RefCOCO+	G-Ref
CLIP-based Model			
CRIS [21]	70.47	62.27	59.87
CM-MaskSD [19]	72.18	64.47	62.67
RIS-CLIP [7]	75.68	69.16	67.62
Ours w/ CLIP Prior	**78.16**	**70.02**	**69.73**
Non-CLIP Model			
LAVT [25]	74.46	65.81	63.34
VG-LAW [18]	75.05	66.61	65.36
Ours w/o CLIP Prior	75.43	67.38	68.12

3.5. Full ablation study

Table 9 presents an ablation study conducted on the validation set of RefCOCO and Ref-Youtube-VOS, evaluat-

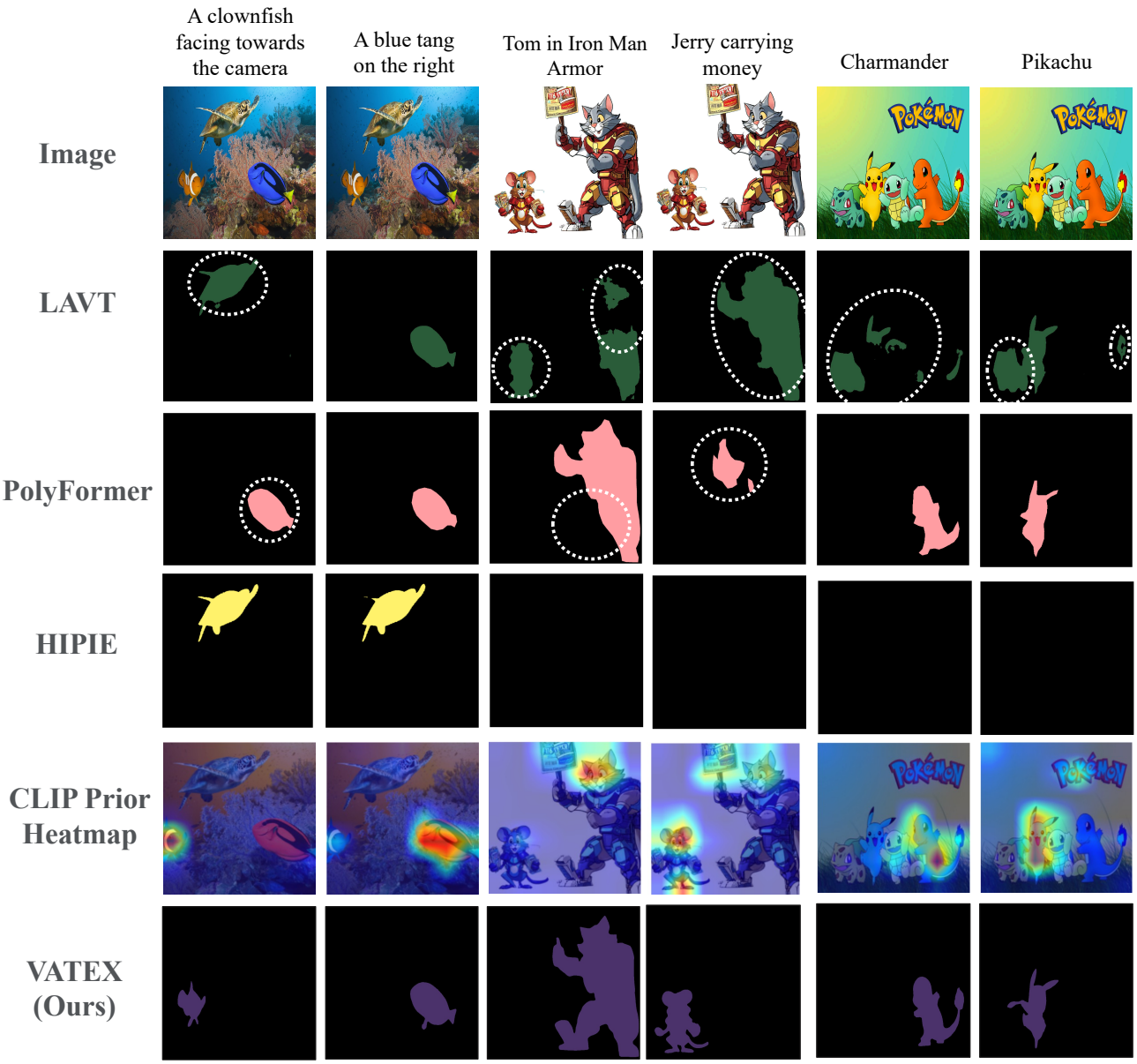


Figure 4. Comparison between VATEX with state-of-the-art methods on challenging out-of-vocabulary cases in referring image segmentation. LAVT's pixel-based approach results in imprecise masks with irrelevant pixel activation. PolyFormer, while creating instance-based masks, struggles with hard cases like "clownfish" or "Jerry" due to limited recognition of unfamiliar objects. HIPIE [20] fails completely due to its constrained pretraining on 365 categories from Objects365. Its high performance on RefCOCO may stem from overfitting and potential data leakage when joint training with COCO. In contrast, VATEX successfully segments correct objects in these difficult vocabulary situations by leveraging the CLIP Prior heatmap. This demonstrates VATEX's superior generalization to unseen objects and complex expressions, highlighting its effectiveness in real-world referring image segmentation tasks.

ing the mIoU (mean Intersection over Union) and $\mathcal{J}\&\mathcal{F}$, respectively of different model configurations. The study explores the impact of three components: CLIP Prior, CMD (Contextual Multimodal Decoder), and MCC (Meaning Consistency Constraint).

The first row represents the baseline model with none of the studied components incorporated. The mIoU for this configuration is 70.42% mIoU and 59.8 $\mathcal{J}\&\mathcal{F}$. In rows 2 to 4, the ablation study reveals that incorporating independently the CLIP Prior alone (row 2) and CMD (row 3) both contribute positively to the mIoU on the RefCOCO and $\mathcal{J}\&\mathcal{F}$ on Ref-YoutubeVOS validation set with an improvement of $1.53\%, 2.76\%$ mIoU and $1.7\%, 2.1\%$ $\mathcal{J}\&\mathcal{F}$, whereas the introduction of the Meaning Consistency Con-

Table 9. Ablation Study on the validation set of RefCOCO (mIoU) and Ref-Youtube-VOS ($\mathcal{J}\&\mathcal{F}$).

	CLIP Prior	CMD	MCC	RefCOCO	Ref-Youtube-VOS
1	-	-	-	70.42	59.8
2	✓	-	-	71.95 +1.53	61.5 +1.7
3	-	✓	-	73.18 +2.76	61.9 +2.1
4	-	-	✓	70.70 +0.30	60.2 +0.4
5	✓	✓	-	75.12 +4.72	63.1 +3.3
6	✓	-	✓	72.14 +1.74	61.3 +1.5
7	-	✓	✓	75.43 +5.01	63.6 +3.8
8	✓	✓	✓	78.16 +7.74	65.4 +5.6

Table 10. Ablation on the number of queries.

Number of queries	1	3	5	10	20	50
RefCOCO	77.23	77.84	78.16	78.02	78.11	77.91

straint (MCC) alone (row 4) leads to a modest increase (only 0.30% mIoU and 0.4 $\mathcal{J}\&\mathcal{F}$), emphasizing the individual significance of each component in enhancing model performance. Although MCC alone has a modest impact, when combined with the CMD in row 7, there is a notable improvement of 4.7% (mIoU of 75.1) and 3.3% ($\mathcal{J}\&\mathcal{F}$ of 63.1). This synergy demonstrates that while MCC alone may not perform exceptionally, its collaboration with CMD effectively enhances model performance, aligning with our approach of leveraging enriched text features conditioned by visual information for improved mutual interaction. The final row represents the model with all components (CLIP Prior, CMD, and MCC) combined, achieving the highest mIoU of 78.16 (+7.74) and $\mathcal{J}\&\mathcal{F}$ of 65.4 (+5.6).

Table 10 presents the impact of varying query numbers on VATEX's performance for the RefCOCO dataset. The results show that while a single query (N=1) achieves a respectable 77.23% mIoU, increasing the number of queries generally improves performance. The optimal performance is achieved with 5 queries, yielding 78.16% mIoU, while the performance slightly decreases for query numbers above 5 (78.02% for 10, 78.11% for 20, and 77.91% for 50 queries). The performance pattern is consistent with ReferFormer [22]'s findings.

3.6. Effect of MCC on Object segmentation mask.

To validate the effectiveness of our proposed MCC module, we propose to use a new Object-centric Intersection over Union (Oc-IoU) score. Unlike mIoU, which averages the overlap and union area for all segmentation predictions within the **same image**, Oc-IoU measures the overlap and union area between the ground truth and all segmentation predictions for the **same object** across different expressions, then averages these values across *all objects in the dataset*. This metric provides an evaluation of the consistency and accuracy of segmentation results across various expressions.

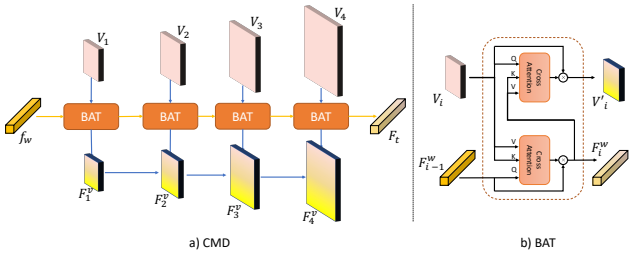


Figure 5. The architecture of Contextual Multimodal Decoder.

Table 11 provides the comparisons between our method and the state-of-the-art method LAVT in Oc-IoU on the validation set of three RIS benchmarks. As can be seen, our method outperforms LAVT in all three datasets. Comparing the last two rows of Table 11, we can see that the MCC helps the model, especially CMD to enhance mutual information between textual and visual features to further provide more consistent and accurate segmentation. These results underscore the compelling efficacy of our Meaning Consistency Constraint in resolving language ambiguities, thus improving the segmentation performance.

Table 11. Performance comparison between LAVT and VATEX on Oc-IoU metric.

Method	RefCOCO	RefCOCO+	G-Ref
LAVT [25]	62.51	50.79	56.01
Ours w/o MCC	66.42	54.92	59.25
Ours	**68.20**	**57.38**	**61.69**

3.7. Archiecture Figure of CMD

For a robust use of visual and text features in subsequent steps, we propose to fuse visual and text features using a Contextual Multimodal Decoder (CMD), which is designed to produce multi-scale text-guided visual feature maps while enhancing contextual information from the image into word-level text features in a hierarchical design as shown in Figure 5. The process on each level of CMD is achieved by a Bi-directional Attention Transfer(BAT), which incorporates two cross-attention modules.

3.8. Runtime and Computational Comparison of VATEX

We report the inference time in FPS and the number of parameters among VATEX, PolyFormer, and LAVT in Table 12. FPS is measured on an NVIDIA RTX 3090 with a

batch size of 1 by taking the average runtime on the entire RefCOCO validation set.

Table 12. Comparison in inference time and parameters on the validation set of RefCOCO dataset.

Method	mIoU	FPS	#params	#trainable params
LAVT	74.46	13	217M	217M
PolyFormer	75.96	3.5	295M	295M
VATEX(**Ours**)	78.16	11	251M	165M

3.9. Additional Visual Results

In Figure 6 and Figure 7, we present additional visualization results for our approach. These results demonstrate that VATEX can successfully segment referred objects in a variety of scenarios, including complex expressions or scenes containing multiple similar objects or rapidly changing shapes. To further illustrate our method's capabilities, we have also created a video demo that compares our approach to ReferFormer on Ref-Youtube-VOS. This video demo is provided as an attachment.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1

[2] 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*, 2022. 6

[3] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 2

[4] Ziling Huang and Shin'ichi Satoh. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7753–7762, Singapore, Dec. 2023. Association for Computational Linguistics. 3

[5] 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, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 1

[6] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1, 2

[7] Seoyeon Kim, Minguk Kang, and Jaesik Park. Risclip: Referring image segmentation framework using clip. *arXiv preprint arXiv:2306.08498*, 2023. 6

[8] 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. 4, 5

[9] 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. 3, 5

[10] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Towards robust referring video object segmentation with cyclic relational consensus, 2023. 6

[11] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 3

[12] 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. 3, 5

Figure 6. Qualitative results of VATEX according to different language expressions for each image on the validation split of G-Ref.

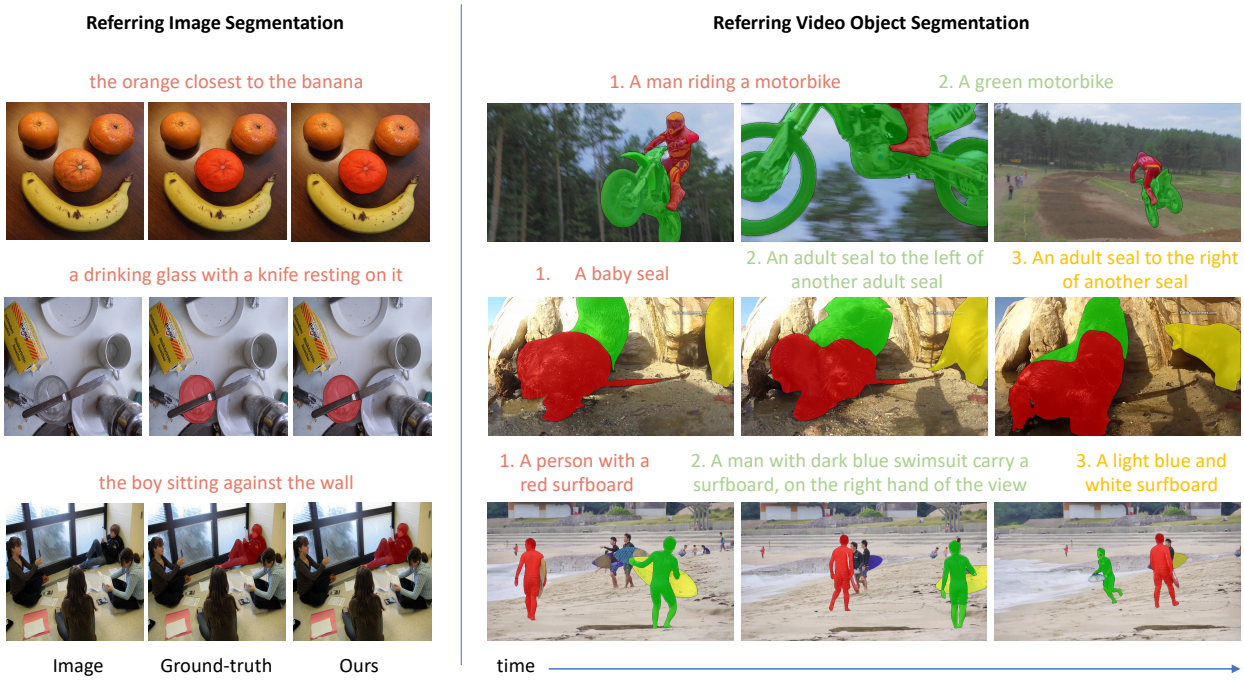


Figure 7. Visualization of VATEX's results. VATEX performs well in complex scenarios such as rapidly changing (*motorbike*), and distinguishing from multiple highly similar objects (*people, seal*). The last row of the video results shows a failure case: PDF segments the wrong man in the last column who has similar attributes when the correct one(green) disappears in the video sequences. Best viewed in color.

[13] 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. 4, 5

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 4

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 1

[16] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. 1

[17] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1

[18] 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. 6

[19] Wenxuan Wang, Jing Liu, Xingjian He, Yisi Zhang, Chen Chen, Jiachen Shen, Yan Zhang, and Jiangyun Li. Cmmasksd: Cross-modality masked self-distillation for referring image segmentation. *arXiv preprint arXiv:2305.11481*, 2023. 6

[20] 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. 4, 5, 7

[21] 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. 3, 6

[22] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 6, 8

[23] 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. 4, 5

[24] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes

extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 3

[25] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 3, 6, 8

[26] 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. 3, 5

[27] Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee*, and Jianfeng Gao*. Generalized decoding for pixel, image and language. 2022. 4, 5

[28] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 5