VDSL 하계 세미나

Referring Expression Segmentation

Sogang University
Dept. of Artificial Intelligence

Presented by
조유빈

Outline

- Background
  - Referring Expression Segmentation
  - Transformer
- Referring Expression Segmentation
  - CGFormer [1]
    - Contrastive Grouping with Transformer for Referring Image Segmentation (CVPR 2023)
  - GRES [2]
    - GRES: Generalized Referring Expression Segmentation (CVPR 2023 Highlight)
- Conclusion

Background

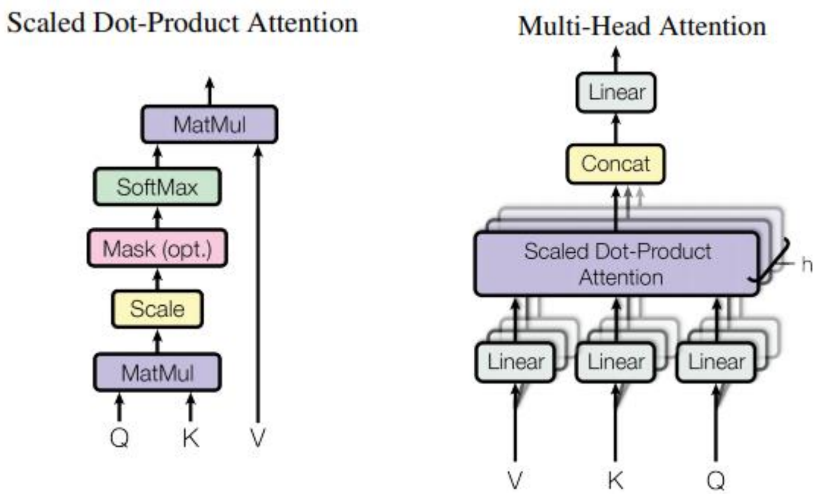
- Referring Expression Segmentation
  - Vision-Language multi-modal task
  - Target object를 지칭하는 language expression이 주어지면 이미지 내에서 해당 object만을 추출해내는 segmentation task
  - Challenging points
    - Target 객체와 다른 객체들과의 relationship 고려
    - 모호하고 복잡한 언어표현에 대한 알맞은 이해

Language expression & image input

Target 검출

Background

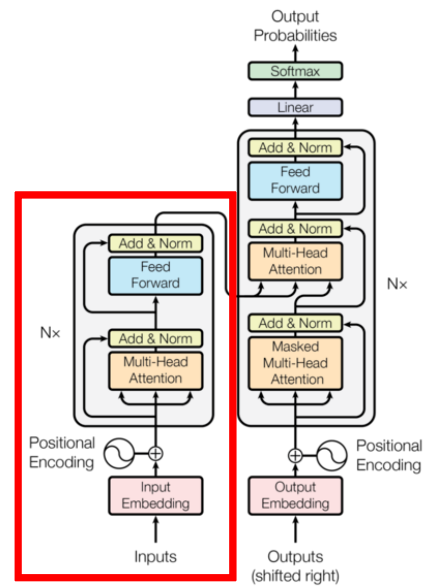
- Transformer [1]
  - Self-attention
    - 단일 sequence 내의 서로 다른 요소들을 관련시켜 한 position의 representation을 계산
  - Why self-attention
    - 병렬적으로 동시에 연산 가능
    - 멀리 떨어진 원소들 간의 path length 감소
      - Long-term dependency problem 해결
      - Global dependency 학습



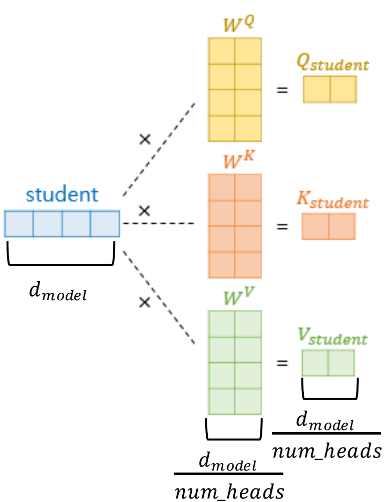
서강대학교 SOGANG UNIVERSITY

Background

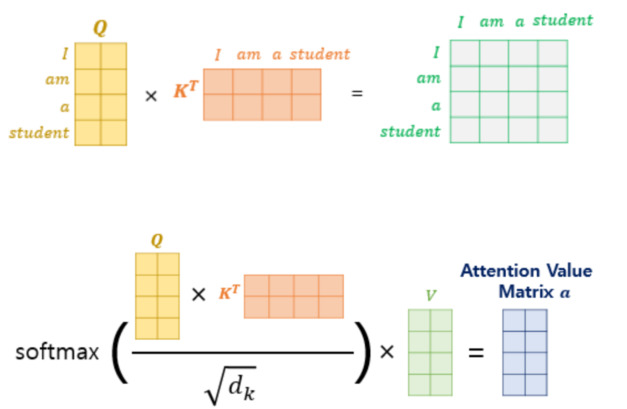
- Transformer [1]
  - Encoder
    - Multi-head self-attention
      - Self-attention : Q, K, V의 출처가 같음 (encoder vector)
      - Multi-head : 벡터의 차원을 축소시키고 attention을 병렬적으로 수행
        - ✓ 다른 관점에서 정보들을 수집
        - ✓ W^Q, W^K, W^V는 각 attention head마다 값이 다름

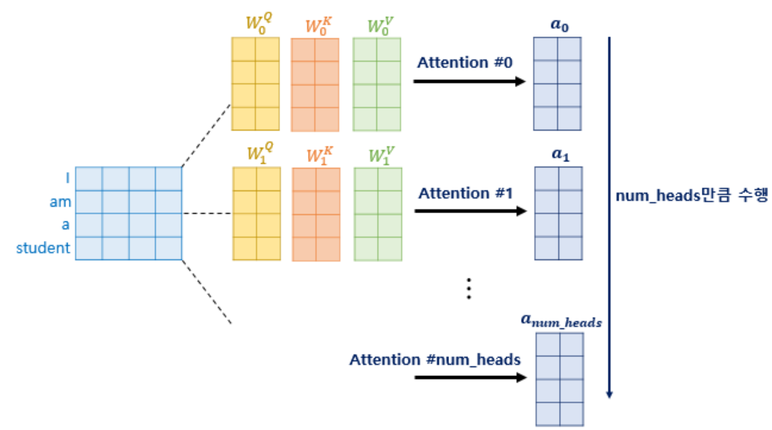


Query / Key / Value embedding



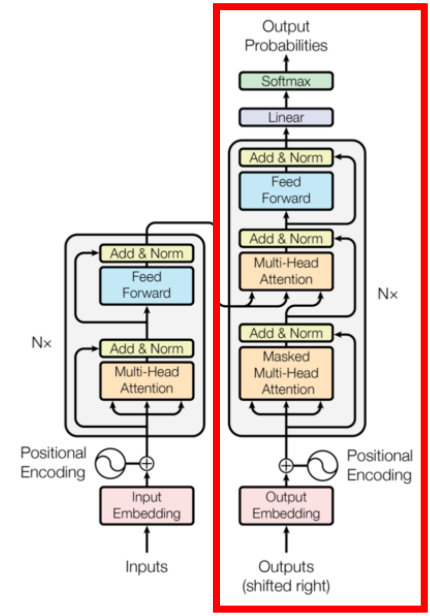
Attention score map

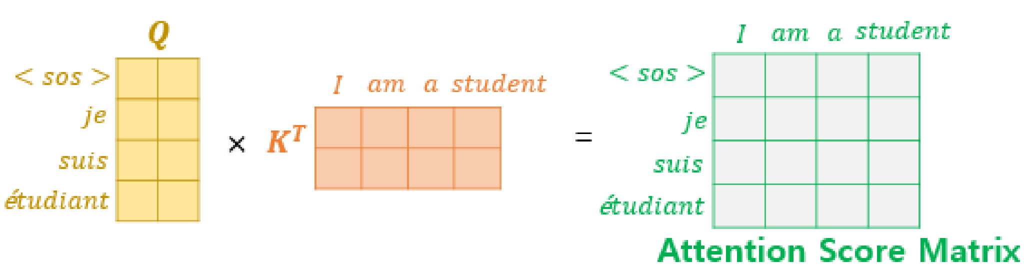


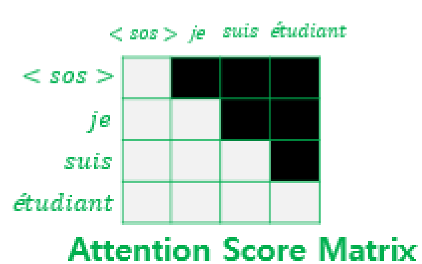


Background

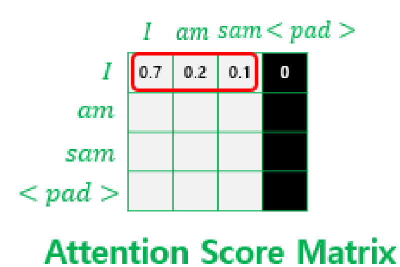
- Transformer [1]
  - Decoder
    - Masked multi-head self-attention
      - Self-attention : Q, K, V의 출처가 같음 (decoder vector)
      - 일부 원소는 매우 작은 음수 값을 곱해 masking
        - ✓ 실질적인 의미를 가진 단어가 아닌 <pad>인 경우
        - ✓ 현재 시점보다 미래에 있는 단어인 경우
    - Multi-head cross-attention (non self-attention)
      - Non self-attention : Q (decoder vector) / K, V (encoder vector)
      - Decoder출력을 위해 encoder의 어떤 정보를 참고하면 좋을지 attention 수행







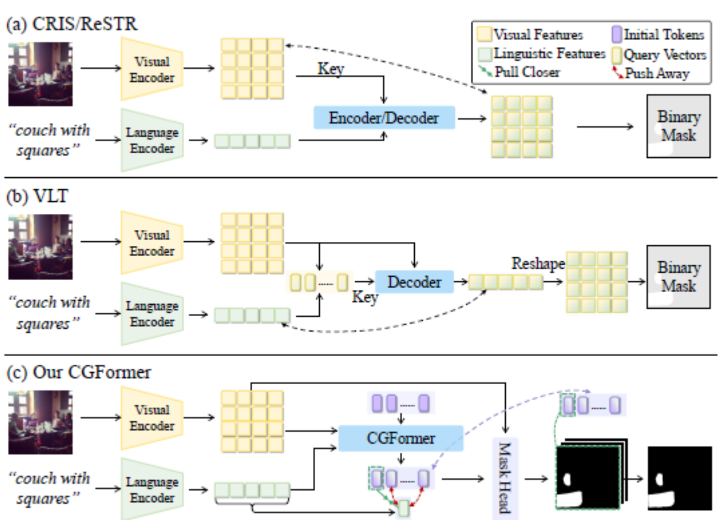
< Look-ahead mask >



< Padding mask >

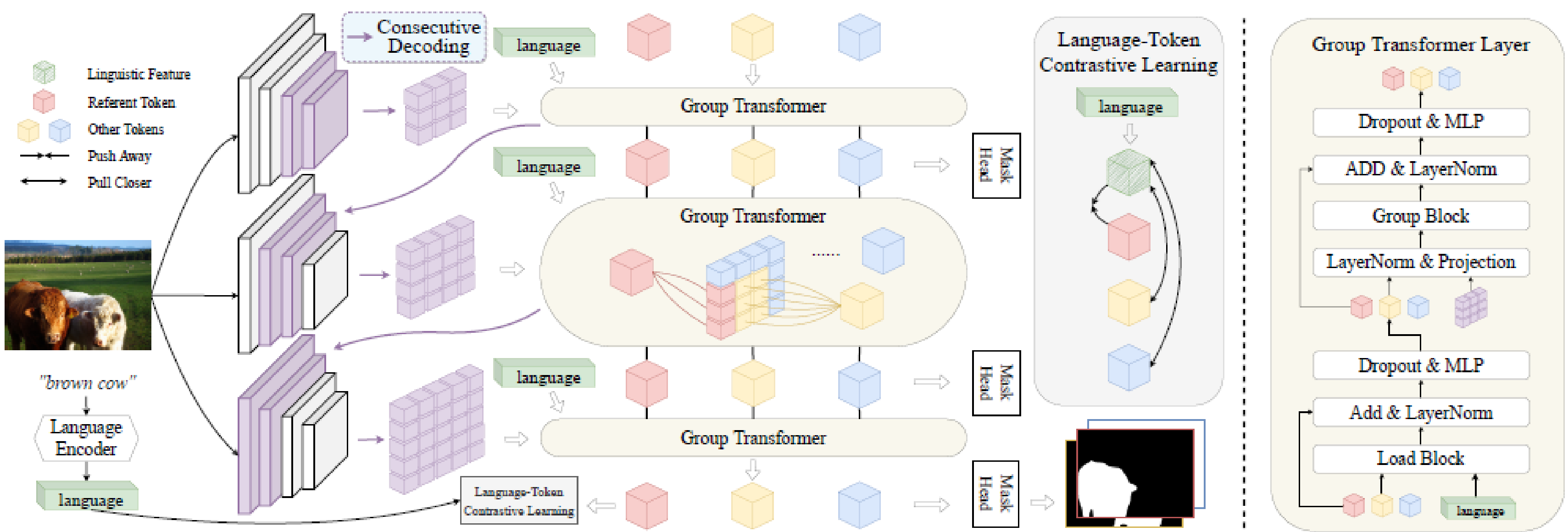
Referring Expression Segmentation

- CGFormer [1]
  - Existing methods fail to capture critical object-level information
    - Fail to focus on different regions and model their relations
    - Does not model the inherent differences between query vectors
      - Still focus on similar regions
  - Contrastive Grouping with Transformer (CGFormer) explicitly captures object-level information via token-based querying and grouping strategy
    - Different tokens focus on different visual regions without overlaps
    - Cooperate contrastive learning with the grouping strategy
  - Consecutive decoder achieve cross-level reasoning



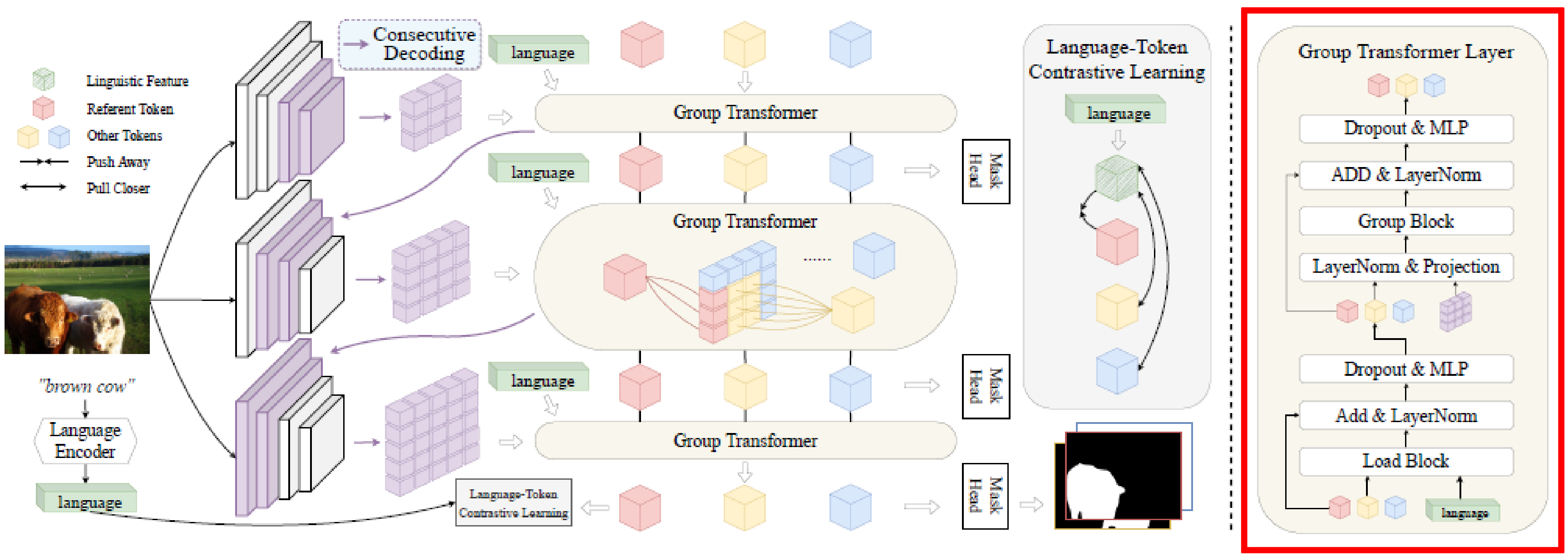
Referring Expression Segmentation

- CGFormer [1]
  - Group transformer
    - Use learnable query tokens to represent object-level information
    - Update query tokens by alternately querying the linguistic features and grouping visual features
      - Tokens capture the rich object characteristics relevant to the expression
  - Use contrastive learning to distinguish the referent token from other tokens
    - Maximizing the similarity between the referent token and the expression and minimizing the similarities between negative pairs



Referring Expression Segmentation

- CGFormer [1]
  - Group transformer layer
    - Load Block
      - Classical cross-attention block
      - Preload what linguistic information the query tokens should focus on at the current layer
    - Group Block
      - Interact between vision and language
      - Group visual features from the feature map into linguistic-enhanced query tokens



Referring Expression Segmentation

- CGFormer [1]
  - Group transformer layer
    - Group Block
      - Embed the query tokens T_i and the vision feature map D_i into a common feature space
      - Calculate the similarities S_{pixel} between every pairwise features of the query tokens T_i' and vision features D_i' (eq.(1))
      - Compute the group to assign a segment token to by taking the one-hot operation of it argmax over all the groups (*hard assignment*)
        - ✓ Since the one-hot assignment operation via argmax is not differentiable, adopt a learnable Gumbel-softmax
        - ✓ Gradient of S_{mask} is equal to the gradient of S_{gumble}, which makes the Group Block differentiable and end-to-end trainable

$$S_{pixel} = \text{norm}_2(T_i')\text{norm}_2(D_i')^T, \qquad (1)$$

$$S_{gumbel} = \text{softmax}\big((S_{pixel} + G)/\tau\big),\ G\text{: Gumbel}(0,1)\text{ distribution} \qquad (2)$$

Hard assignment →

$$S_{onehot} = \text{onehot}\left(\text{argmax}_N\left(S_{gumbel}\right)\right), \qquad (3)$$

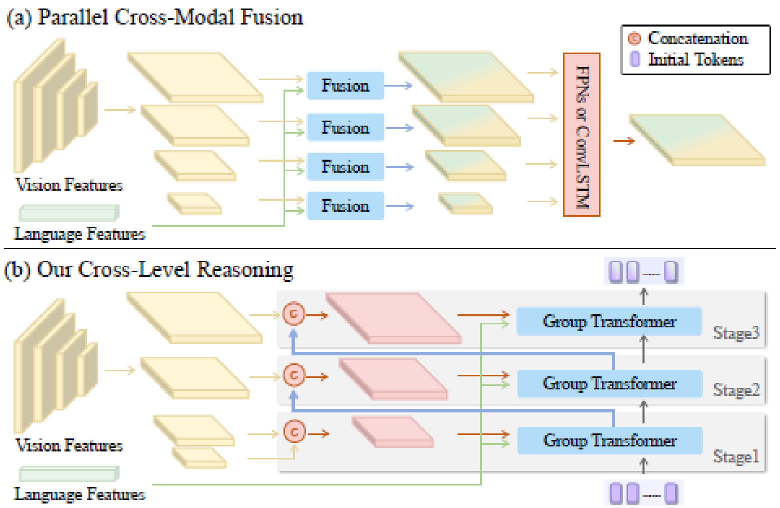
$$S_{mask} = (S_{onehot})^T - \text{sg}\big(S_{gumbel}\big) + S_{gumbel}\text{, sg : stop gradient} \qquad (4)$$

$$T_i = MLP(S_{mask}D_i') + T_i' \qquad (5)$$

서강대학교 SOGANG UNIVERSITY

Referring Expression Segmentation

- CGFormer [1]
  - Consecutive decoder
    - Previous works model the vision-language interaction at multiple levels in parallel and late integrate multi-level results
      - Fails to perform joint interaction across various levels
    - Consecutive decoder performs cross-level reasoning
      - Jointly updating the query tokens in every two consecutive decoder layers
      - The two-level cross-modal information will be consecutively propagated in multiple levels from bottom to up



Referring Expression Segmentation

- CGFormer [1]
  - Ablation study
    - The results of the method 2 and 3 suggest that simply adding tokens cannot boost performance
      - These tokens are likely to focus on similar information rather than distinct regions
    - Grouping strategy cooperated with contrastive loss to make tokens can focus on different regions
      - Method 4 delivers a 4.35% improvement
    - Hard assignment helps to obtain a more refined grouping
      - Method 5 achieves an improvement of 1.63%
    - Method 7 shows the effectiveness of the consecutive decoder
    - Compared to method 8, method 7 validates the necessity of the proposed contrastive grouping

	Method	P@0.5	P@0.7	P@0.9	oIoU
1	baseline	75.31	61.48	16.85	65.70
2	1+one token	77.28	64.94	19.47	66.39
3	1+N tokens	77.70	65.12	19.44	66.46
4	3+grouping	83.94	72.09	23.43	70.81
5	4+hard assignment	84.59	74.92	33.75	72.44
6	5+multi-scale	85.80	76.31	35.35	73.28
7	5+CD (ours full)	**87.23**	**78.69**	**38.77**	**74.75**
8	VLT(Swin-B+BERT)*	83.24	72.81	24.64	70.89
9	w/o cos	85.64	76.23	33.96	73.37
10	w/o learnable τ	86.14	76.99	36.48	73.50

Table 3. Ablation study on the validation set of RefCOCO. CD: Consecutive Decoder. cos: cosine similarity operation. τ: learnable parameter in Gumble Softmax. Results with * refer to [62].

서강대학교 SOGANG UNIVERSITY

Referring Expression Segmentation

- CGFormer [1]
  - Results

	Method	RefCOCO			RefCOCO+			G-Ref			ReferIt
		val	test A	test B	val	test A	test B	val-U	test-U	val-G	test
mIoU	DMN [40]	49.78	54.83	45.13	38.88	44.22	32.29	-	-	36.76	52.81
	MCN [37]	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-	-
	CGAN [36]	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	46.54	-
	LTS [23]	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-	-
	VLT [12]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	49.76	-
	CRIS [51]	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-	-
	Our CGFormer	**76.93**	**78.70**	**73.32**	**68.56**	**73.76**	**61.72**	**67.57**	**67.83**	**65.79**	**66.42**
oIoU	RRN [26]	55.33	57.26	53.93	39.75	42.15	36.11	-	-	36.45	63.63
	MAttNet [64]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	-	-
	CMSA [63]	58.32	60.61	55.09	43.76	47.60	37.89	-	-	39.98	63.80
	CMPC [19]	61.36	64.53	59.64	49.56	53.44	43.23	-	-	49.05	65.53
	LSCM [20]	61.47	64.99	59.55	49.34	53.12	43.50	-	-	48.05	66.57
	CEFNet [14]	62.76	65.69	59.67	51.50	55.24	43.01	51.93	-	-	66.70
	BUSNet [61]	63.27	66.41	61.39	51.76	56.87	44.13	-	-	50.56	-
	ReSTR [25]	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-	-
	LAVT [62]	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50	-
	Our CGFormer	**74.75**	**77.30**	**70.64**	**64.54**	**71.00**	**57.14**	**64.68**	**65.09**	**62.51**	**73.36**

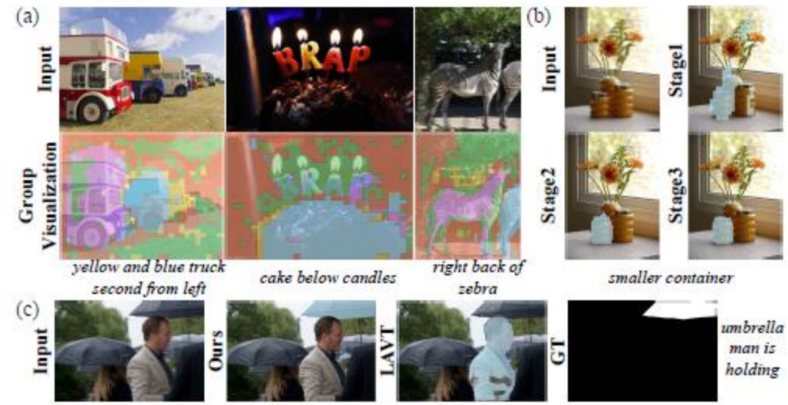
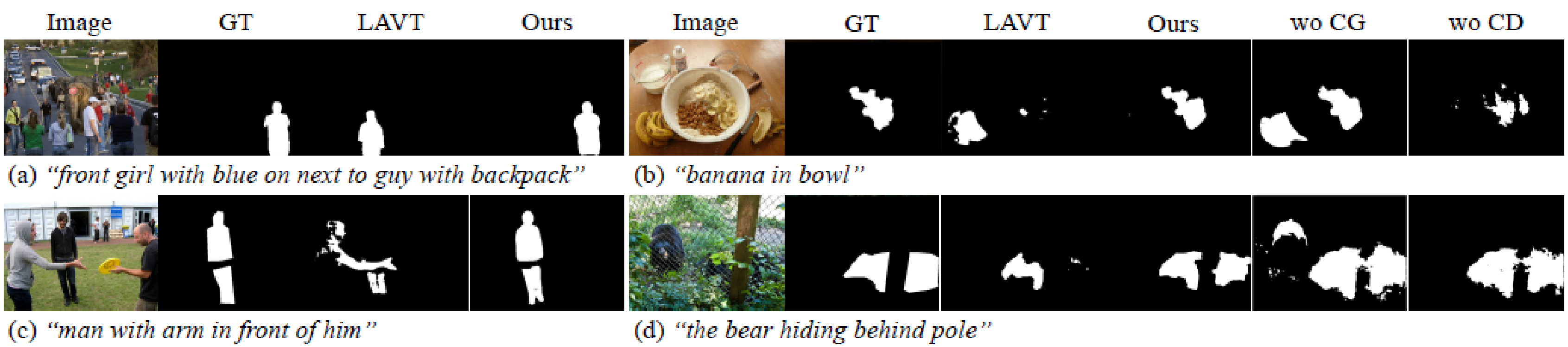


Figure 5. Visualization of grouping results for (a) different tokens (in different colors), (b) the referent token in three stages and (c) segmentation results of unseen objects.

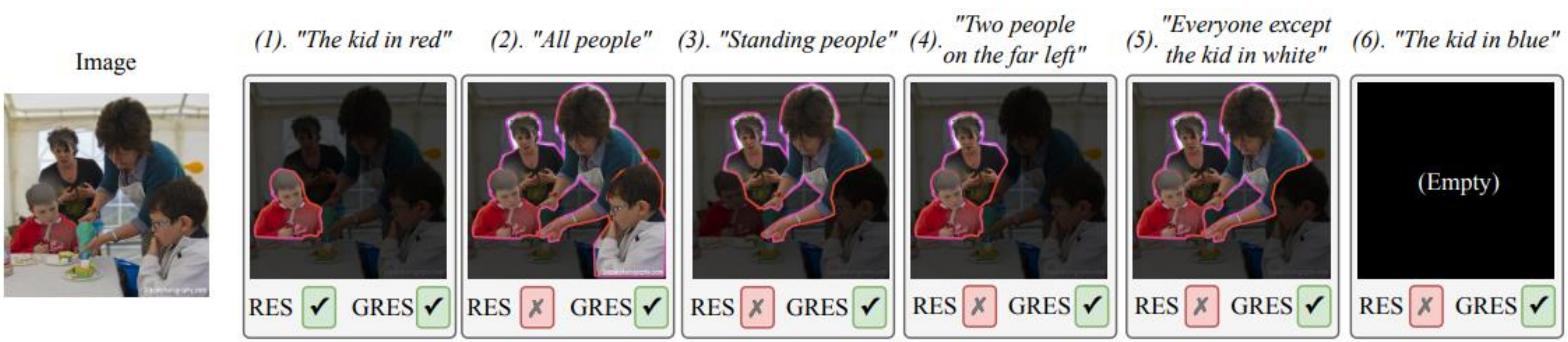


< Visualization results >

1) Liu, Ding, and Jiang, “GRES: Generalized Referring Expression Segmentation” (CVPR 2023)

Referring Expression Segmentation

- GRES [1]
  - 기존 referring expression segmentation에서는 single target object만을 지칭하는 language expression으로 구성
    - Multi-target이나 no-target에 대한 expression은 고려되지 않음
  - 본 논문에서는 새로운 데이터셋인 generalized referring expression segmentation (GRES)을 제안
    - Single-target, multi-target, no-target에 대한 expressions를 포함
    - Enhances the model’s reliability and robustness to realistic scenarios where any type of expression can occur unexpectedly



< RES와 GRES 비교 >

1) Liu, Ding, and Jiang, "GRES: Generalized Referring Expression Segmentation" (CVPR 2023)

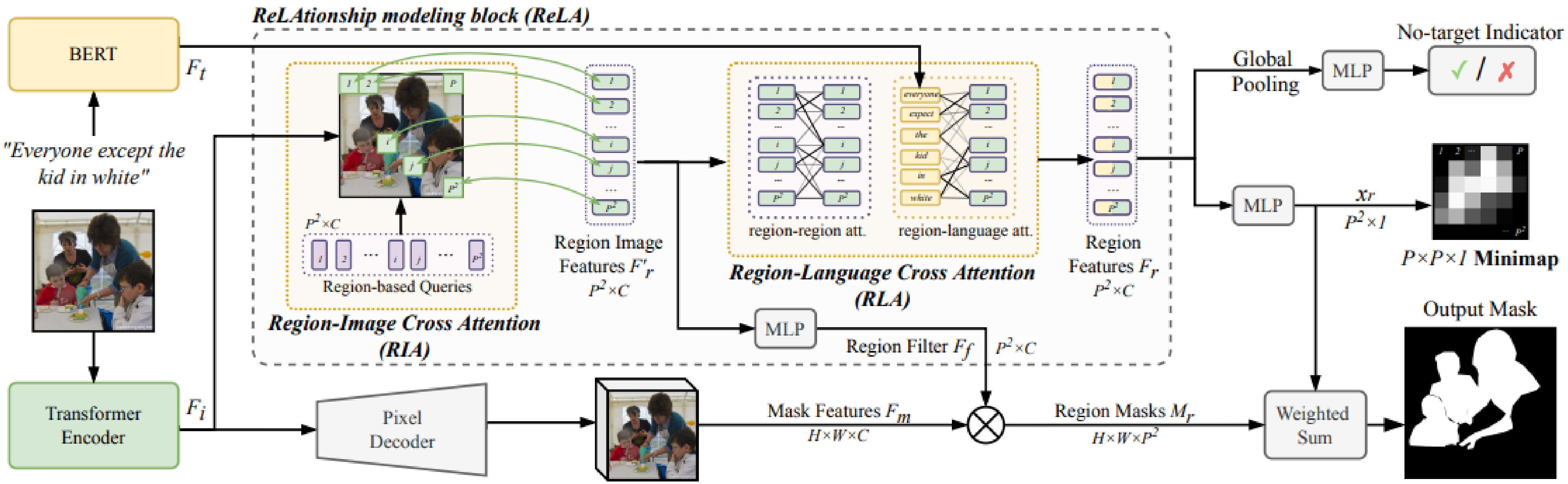
Referring Expression Segmentation

- GRES [1]
  - Features of multi-target samples
    - Usage of counting expressions (ex. *two* people)
      - The model must be able to differentiate cardinal numbers from ordinal numbers
    - Compound sentence structures without geometrical relation (ex. *and*, *except*, *with*, *or*)
      - Require the model to understand the long-range dependencies of both the image and sentence
    - Domain of attributes (ex. *Right* lady in *blue* and kid in *white*)
      - Require the model to have a deeper understanding of all the attributes and map the relationship of these attributes to their corresponding objects
    - More complex relationships
      - Require the model to have a deep understanding of all instances and their interactions in the image and expression
  - Rules for no-target samples to keep the dataset at a reasonable difficulty
    - The expression cannot be totally irrelevant to the image
    - The annotators could choose a deceptive expression drawn from other images

Image (a) | i. *"The **two** people on the far left"* | ii. *"Everyone **except** the kid in white"* | Image (b) | i. *"The bike **and** two passengers on **it**"* | ii. *"The bike **that has** two passengers and **its** driver"*

Referring Expression Segmentation

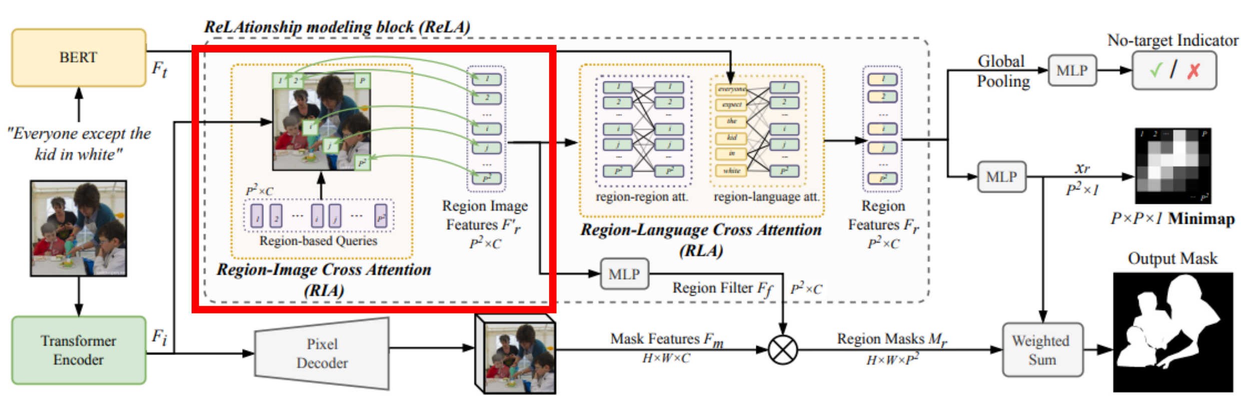
- GRES [1]
  - Overall architecture of the proposed baseline model for GRES
    - Modeling the interaction among regions in the image
      - Different from previous works using hard-split, regions are not predefined by using learnable queries
    - For the n^{th} regions, scalar x_r^n indicates its probability of containing targets
    - Region filter F_f is multiplied with the mask features F_m to generate the region mask M_r
    - Outputs : segmentation mask M & no-target label E
      - If E is predicted to be positive, the output mask M will be set to empty

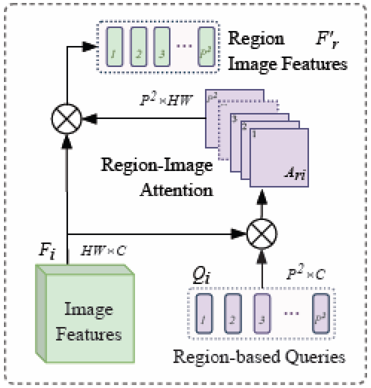


1) Liu, Ding, and Jiang, "GRES: Generalized Referring Expression Segmentation" (CVPR 2023)

Referring Expression Segmentation

- GRES [1]
  - ReLA: ReLAtionship modeling
    - Region-Image Cross Attention (RIA)
      - Flexibly collects region image features
      - Using P^2 learnable Region-based Queries supervised by the minimap
        - ✓ Each query corresponds to a spatial region in the image
      - The attention between image feature F_i and P^2 query embeddings Q_r is performed to generate P^2 attention maps
        - ✓ A_{ri} gives each query a $H \times W$ attention map indicating its corresponding spatial areas in the image
      - Making regions represent more fine-grained attributes at the sub-instance level
        - ✓ Sub-instance representations are desired for addressing the complex relationship and attribute descriptions in GRES

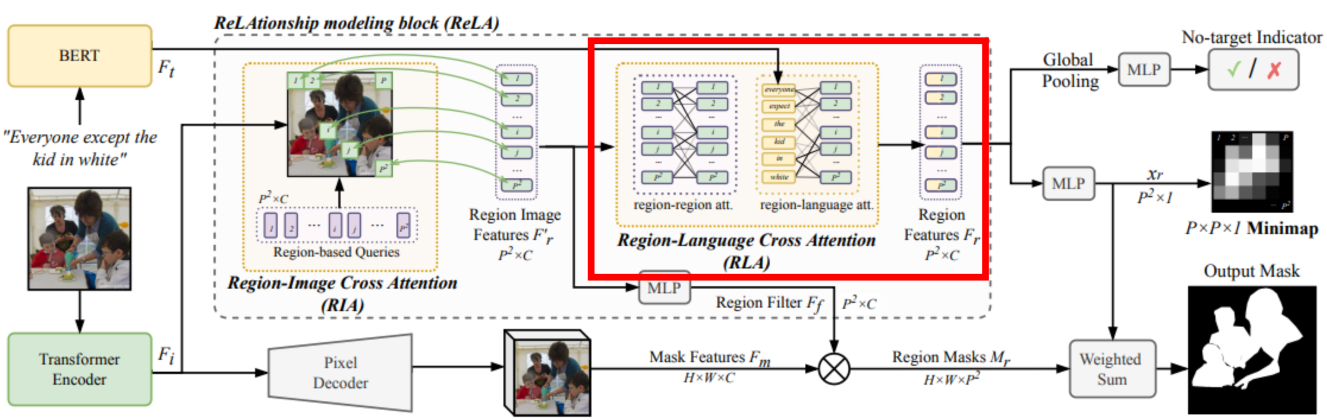


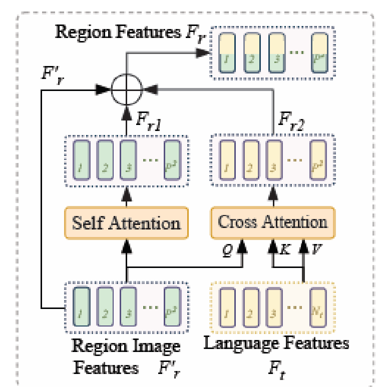


< Region-Image Cross Attention >

Referring Expression Segmentation

- GRES [1]
  - ReLA: ReLAtionship modeling
    - Region-Language Cross Attention (RLA)
      - RIA does not consider the relationship between regions and language information
      - Modeling the region-region and region-language interactions
      - Self-attention models the region-region dependency relationships
      - Cross-attention models the relationship between each word and each region
      - MLP fuses the interaction-aware region feature F_{r1}, language-aware region feature F_{r2}, and region image feature F_r'





< Region-Language Cross Attention >

서강대학교
SOGANG UNIVERSITY

1) Liu, Ding, and Jiang, “GRES: Generalized Referring Expression Segmentation” (CVPR 2023)

Referring Expression Segmentation

- GRES [1]
  - Ablation study
    - Fig. 6 shows the necessity and validity of gRefCOCO on the task of GRES
    - Design options of RIA in Table 2
      - Model #1 makes the global image information less pronounced
      - Compared to model #1, model #2 shows the importance of global context in visual feature encoding
      - Model #2 shows the effectiveness of the proposed adaptive region assigning
      - Model #3 shows that explicit correspondence between queries and spatial image regions is beneficial to ReLA

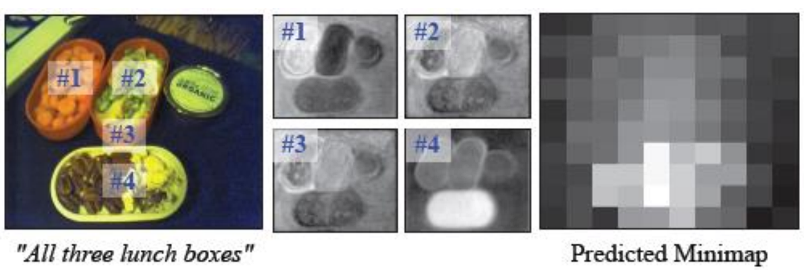
Figure 6. Example predictions of the same model being trained on RefCOCO *vs.* gRefCOCO.

Table 2. Ablation study of RIA design options.

#	Methods	P@0.7	P@0.8	P@0.9	cIoU	gIoU
#1	Hard split, input	63.02	59.81	19.26	54.43	55.34
#2	Hard split, decoder	70.34	65.23	21.47	60.08	60.93
#3	w/o minimap	72.19	66.02	21.07	61.30	62.06
#4	**ReLA** (ours)	**74.20**	**68.33**	**24.68**	**62.42**	**63.60**

Referring Expression Segmentation

- GRES [1]
  - Ablation study
    - Design options of RLA in Table 3
      - #1 : RLA is replaced by point-wise multiplying region features and globally averaged language features
      - #2 shows the validity of region-word interaction modeling
      - #3 shows the importance of the region-region relationship
      - #4 : use the region-region and region-word relationship modeling together
    - Number of region P in Table 4
      - Smaller P leads to coarser regions, which is not good for capturing fine-grained attributes
      - Larger P costs more resources and decreases the are of each region, making relationship learning difficult
      - In Fig.7, each region mask contains not only the instance of this region but also other instances with strong relationships



< Visualization of region masks & predicted minimap >

Table 3. Ablation study of RLA design options.

#	Methods	P@0.7	P@0.8	P@0.9	cIoU	gIoU
#1	Baseline	69.94	61.10	19.38	57.24	58.53
#2	+ language att.	72.03	65.42	21.04	59.86	60.53
#3	+ region att.	73.52	67.01	23.43	61.00	62.38
#4	**ReLA** (ours)	**74.20**	**68.33**	**24.68**	**62.42**	**63.60**

Table 4. Ablation study of Number of Regions

# Regions	P@0.7	P@0.8	P@0.9	cIoU	gIoU
4×4	68.48	60.25	20.33	56.57	57.01
8×8	72.36	66.85	23.56	59.74	61.23
10×10	**74.20**	**68.33**	**24.68**	**62.42**	**63.60**
12×12	74.14	67.56	23.90	62.02	63.50

Referring Expression Segmentation

- GRES [1]
  - Results
    - Comparison with SOTA RES methods on gRefCOCO in Table 5
      - Training previous methods on gRefCOCO
      - For previous networks, output masks with less than 50 positive pixels are cleared to all-negative, for better no-target identification
      - Explicit relationship modeling greatly enhances model's performance
    - No-target identification performance in Table 6
      - The gRefCOCO does not significantly affect the model's targeting performance while being generalized to no-target samples
      - A dedicated no-target classifier of ReLA is desired
        - ✓ ReLA-50pix : ReLA with the no-target classifier disabled
      - There are around 40% of no-target samples are missed
        - ✓ Many no-target expressions are very deceptive and similar with real instances in the image

Table 5. Comparison on gRefCOCO dataset.

Methods	val		testA		testB	
	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU
MattNet [46]	47.51	48.24	58.66	59.30	45.33	46.14
LTS [18]	52.30	52.70	61.87	62.64	49.96	50.42
VLT [5]	52.51	52.00	62.19	63.20	50.52	50.88
CRIS [39]	55.34	56.27	63.82	63.42	51.04	51.79
LAVT [44]	57.64	58.40	65.32	65.90	55.04	55.83
VLT+ReLA	58.65	59.43	66.60	65.35	56.22	57.36
LAVT+ReLA	61.23	61.32	67.54	66.40	58.24	59.83
ReLA (ours)	**62.42**	**63.60**	**69.26**	**70.03**	**59.88**	**61.02**

Table 6. No-target results comparison on gRefCOCO dataset.

Methods	val		testA		testB	
	N-acc.	T-acc.	N-acc.	T-acc.	N-acc.	T-acc.
MattNet [46]	41.15	96.13	44.04	97.56	41.32	95.32
VLT [5]	47.17	95.72	48.74	95.86	47.82	94.66
LAVT [44]	49.32	96.18	49.25	95.08	48.46	95.34
ReLA-50pix	49.96	96.28	51.36	96.35	49.24	95.02
ReLA	**56.37**	**96.32**	**59.02**	**97.68**	**58.40**	**95.44**

Referring Expression Segmentation

- GRES [1]
  - Results
    - In Table 7, ReLA outperforms other methods on classic RES
    - Qualitative results
      - Multiple targets of the same category or different categories in Image (a)
        - ✓ Showing the strong generalization ability
      - Counting words and shared attributes in Image (b)
      - Compound sentence in Image (c)
        - ✓ Model can understand the excluding relationship

Table 7. Results on classic RES in terms of cIoU. U: UMD split. G: Google split.

Methods	Visual Encoder	Textual Encoder	RefCOCO			RefCOCO+			G-Ref		
			val	test A	test B	val	test A	test B	$\text{val}_{(U)}$	$\text{test}_{(U)}$	$\text{val}_{(G)}$
MCN [32]	Darknet53	bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
VLT [5]	Darknet53	bi-GRU	67.52	70.47	65.24	56.30	60.98	50.08	54.96	57.73	52.02
ReSTR [21]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-	54.48
CRIS [39]	CLIP-R101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
LAVT [44]	Swin-B	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
VLT [6]	Swin-B	BERT	72.96	75.96	69.60	63.53	68.43	56.92	63.49	**66.22**	**62.80**
ReLA (ours)	Swin-B	BERT	**73.82**	**76.48**	**70.18**	**66.04**	**71.02**	**57.65**	**65.00**	65.97	62.70

서강대학교
SOGANG UNIVERSITY

Conclusion

- CGFormer [1]
  - Contrastive Grouping with Transformer (CGFormer) achieves object-aware cross modal
  - Consecutive decoder achieves cross-level reasoning
- GRES [2]
  - A new benchmark, called Generalized Referring Expression Segmentation (GRES), allows an arbitrary number of targets in the expressions
  - A baseline ReLA for GRES explicitly model the relationship between different image regions and words