NEURAL INFORMATION PROCESSING SYSTEMS

CoupAlign: Coupling Word-Pixel with Sentence-Mask Alignments for Referring Image Segmentation

Zicheng Zhang[1*] Yi Zhu[2*] Jianzhuang Liu[2] Xiaodan Liang[3] Wei Ke[1†]

[1]Xi'an Jiaotong University [2]Noah's Ark Lab, Huawei [3]Sun Yat-sen University

Background

- Referring image segmentation (RIS) aims at localizing all pixels of the visual objects described by a natural language sentence.
- Main challenges:
  - how to align the given language expression with visual pixels for highlighting the target.
  - how to distinguish the target from similar objects.

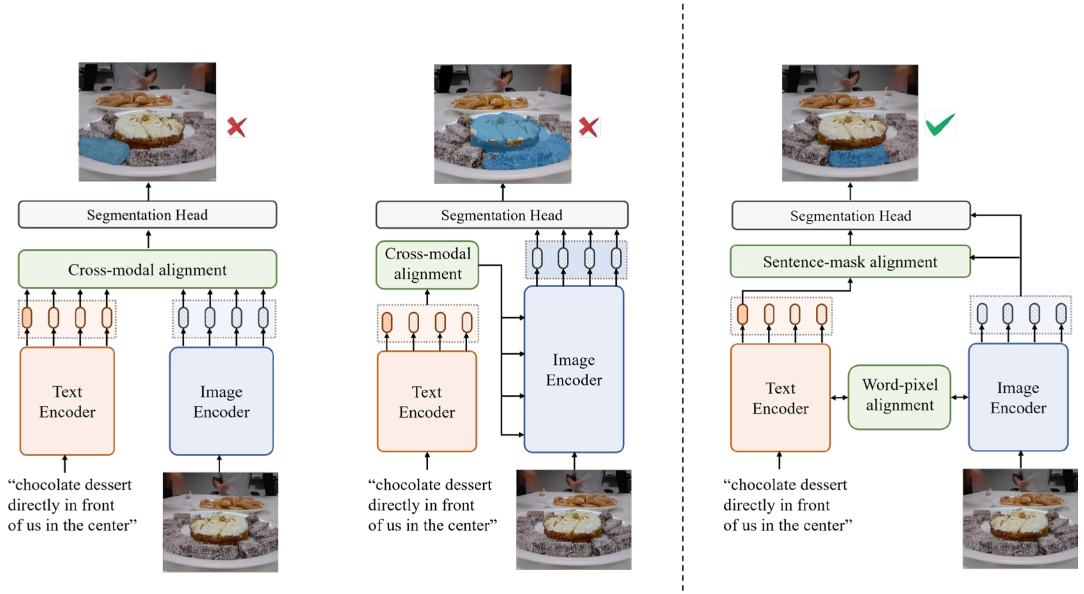
Input: image + texts

- white double decker bus
- gray small car
- the old man with T-shirt
- woman with pink dress

Output: instance masks

Motivation

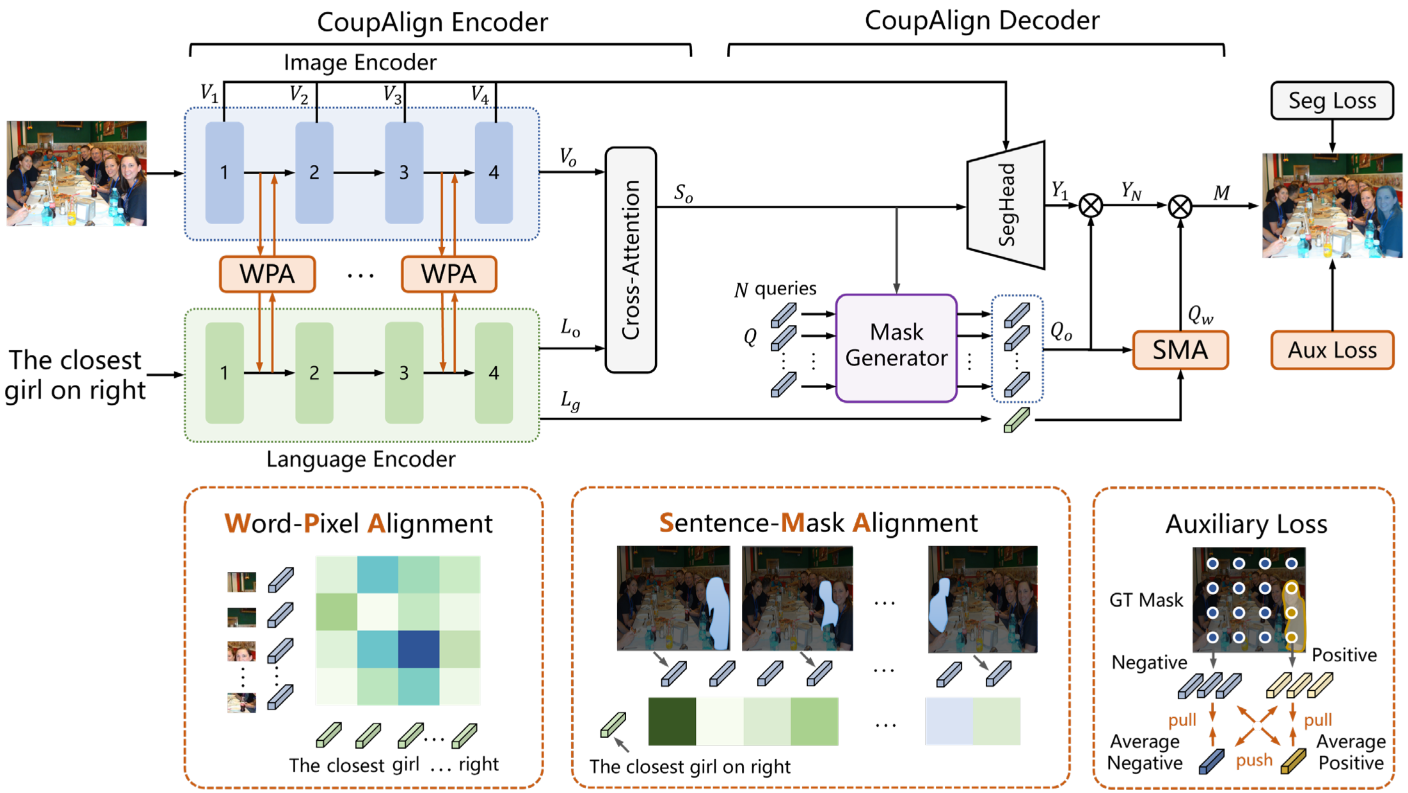
- Previous methods
  - only consider late fusion for vision and language features.
  - or consider early fusion but only have word-pixel level alignment.
- Our methods
  - adopts word-pixel level alignment in the early fusion stage.
  - and use sentence-mask level alignment to enhance the fused features in the late fusion stage.



(a) Existing method: Late Fusion (b) Existing method: Early Fusion (c) Ours: CoupAlign

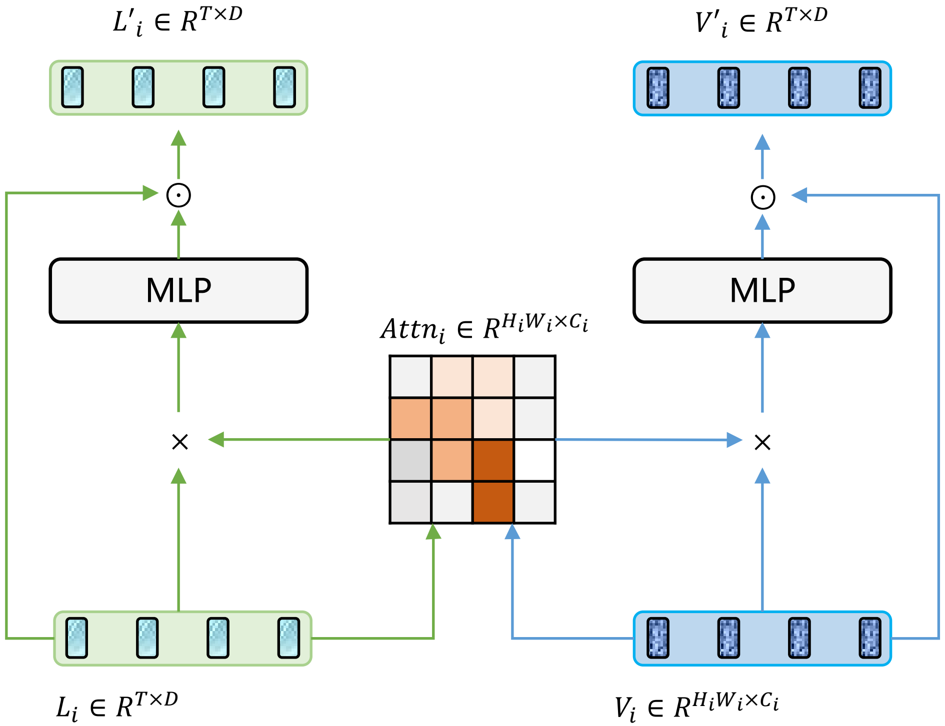
Method

- CoupAlign Architecture
  - WPA module enables cross-model interactions at each encoder stage.
  - Based on the aligned feature S_o, the mask generator produces N mask embeddings.
  - SMA module weights Q_o using L_g and projects the mask signals back to Y_N.



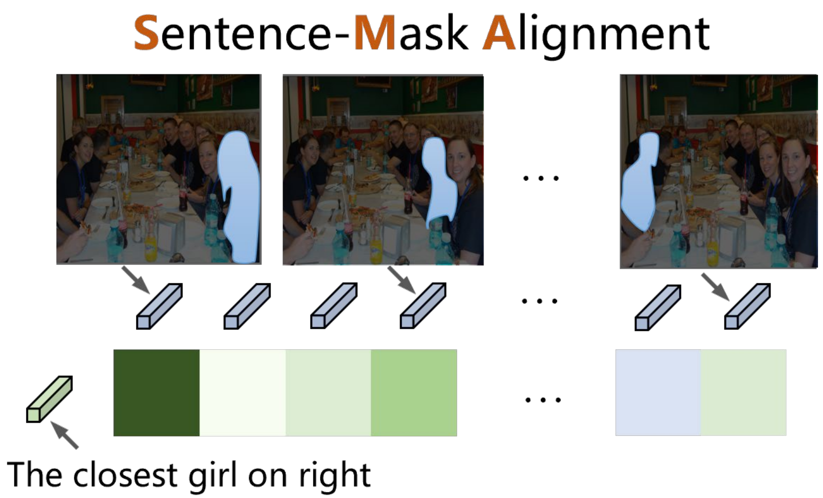
Method

- Word-Pixel Alignment
  - we use cross attention to align word tokens and pixel tokens
  - and design a gate to control the fused information flow.



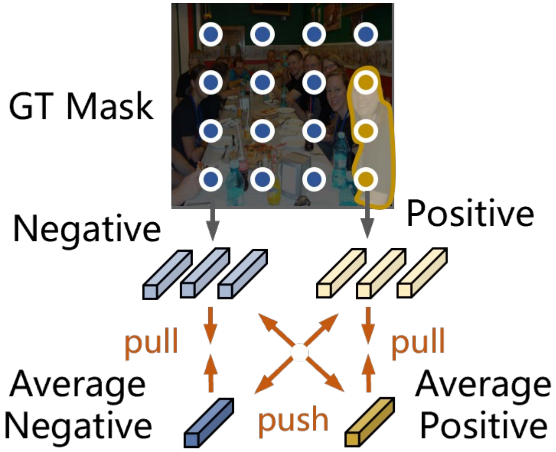
Method

- Sentence-Mask Alignment
  - we use global language feature and mask embeddings to compute attention weights.
  - Then we use the weights to aggregate the proposals to the final mask prediction.



Method

- Auxiliary loss
  - We adopt contrastive loss to enhance the ability of distinguishing the object from the background.



$$L_{P2N}^{\mathrm{Aux}} = -\frac{1}{|\mathcal{P}|}\sum_{y_i^+\in\mathcal{P}} \frac{\exp(y_i^+\cdot\hat{y}^+/\tau)}{\exp(y_i^+\cdot\hat{y}^+/\tau)+\sum_{y_k^-\in\mathcal{N}}\exp(y_i^+\cdot y_k^-/\tau)},$$

$$L_{N2P}^{\mathrm{Aux}} = -\frac{1}{|\mathcal{N}|}\sum_{y_i^-\in\mathcal{N}} \frac{\exp(y_i^-\cdot\hat{y}^-/\tau)}{\exp(y_i^-\cdot\hat{y}^-/\tau)+\sum_{y_k^+\in\mathcal{P}}\exp(y_i^-\cdot y_k^+/\tau)},$$

$$L_j^{\mathrm{Aux}} = L_{P2N}^{\mathrm{Aux}} + L_{N2P}^{\mathrm{Aux}},$$

Result

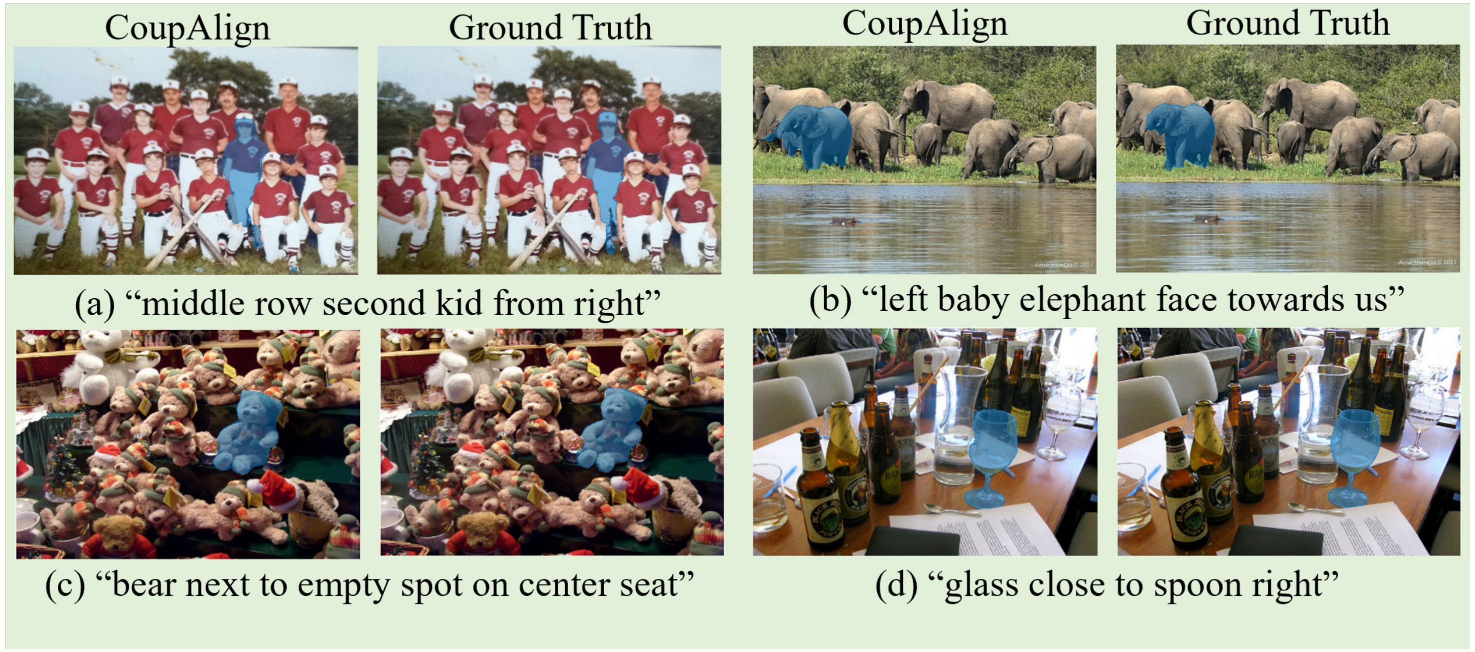
NEURAL INFORMATION PROCESSING SYSTEMS

- Comparing with SOTA
  - We evaluate our method on four referring image segmentation datatset and outperform previous method.

	Backbone	RefCOCO			RefCOCO+			G-Ref		ReferIt
		val	test A	test B	val	testA	testB	val (U)	test (U)	test
MCN [31]	Darknet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
BRINet [12]	ResNet-101	60.98	62.99	59.21	48.17	52.32	42.11	-	-	63.46
CMPC [13]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-	65.53
LSCM [14]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-	66.57
CGAN [30]	ResNet-101	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	-
BUSNet [43]	ResNet-101	63.27	66.41	61.39	51.76	56.87	44.13	-	-	-
EFN [11]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-	66.70
LTS [16]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
VLT [7]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	-
ReSTR [19]	ViT-B-16	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	70.18
CRIS [41]	ResNet-101	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36	-
LAVT [44]	Swin-B	72.73	75.82	68.79	62.14	**68.38**	55.10	61.24	62.09	-
CoupAlign (ours)	Swin-B	**74.70**	**77.76**	**70.58**	**62.92**	68.34	**56.69**	**62.84**	**62.22**	**73.28**

Visualization

- Visualization
  - CoupAlign works well in the scenes where crowded objects have similar color and context.



Conclusion

- Conclusion
  - CoupAlign captures both visual and semantic coherence of pixels within the referred object, and significantly outperforms state-of-the-art RIS methods.
  - Especially, CoupAlign has great ability in localizing the target from similar objects, showing great potential in segmenting natural language referred objects in real-world scenarios.

Thank you!