Extending CLIP's Image-Text Alignment to Referring Image Segmentation

Seoyeon Kim[1] **Minguk Kang**[1] **Dongwon Kim**[1] **Jaesik Park**[2] **Suha Kwak**[1]
[1]POSTECH, [2]Seoul National University
[1]{syeonkim07,mgkang,kdwon,suha.kwak}@postech.ac.kr
[2]jaesik.park@snu.ac.kr

- Problem / objective
  - Referring Image Segmentation
- Contribution / Key idea
  - Cross-modal Feature Extraction (CFE) module
  - Shared-space Knowledge Exploitation (SKE) module

- **Referring Image Segmentation (RIS)**

  자연어 표현이 가리키는 객체에 해당하는 이미지 내 영역 segment 하는 task

- **Previous research**

  이미지와 텍스트 독립적으로 처리 후, 후속적인 modal fusion 기법들을 통해 RIS 수행함.

- **Motivation**

  1. Cross-modal task인 RIS에 unimodal backbone을 사용하는 것이 과연 최선일까?
  2. MaskCLIP [1]: CLIP의 cross-modal nature가 RIS에 굉장히 적합하다!

- **Starting Point**

  CLIP 모델의 사용을 시작점으로 하여, CLIP의 image-text alignment 능력 충분히 이용하겠다.

- **Overview**

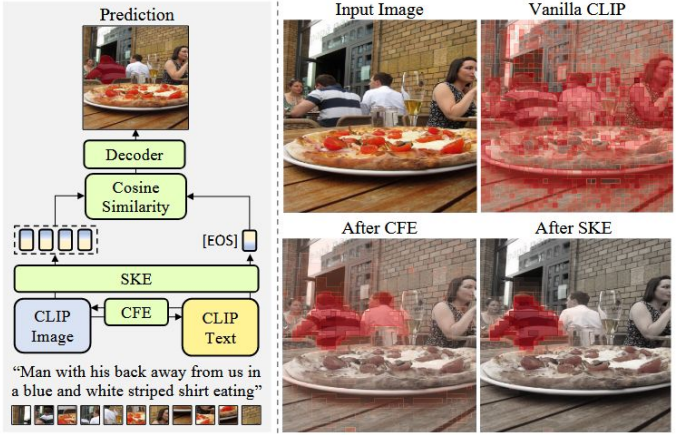


Figure 1: CLIP's image-text alignment produces preliminary patch-level groundings through cosine similarity between patch-level image and sentence-level text features. Building upon this alignment, we refine CLIP's groundings into accurate segmentations with three modules. Cross-modal Feature Extraction (CFE) modules enhance CLIP's unimodal image and text features by aligning them at candidate regions. Shared-space Knowledge Exploitation (SKE) modules leverage the rich alignment knowledge in CLIP's image-text shared-embedding space to discern the target referent. Lastly, a decoder transforms the patch-level grounding into a pixel-wise segmentation.

- **Overview**

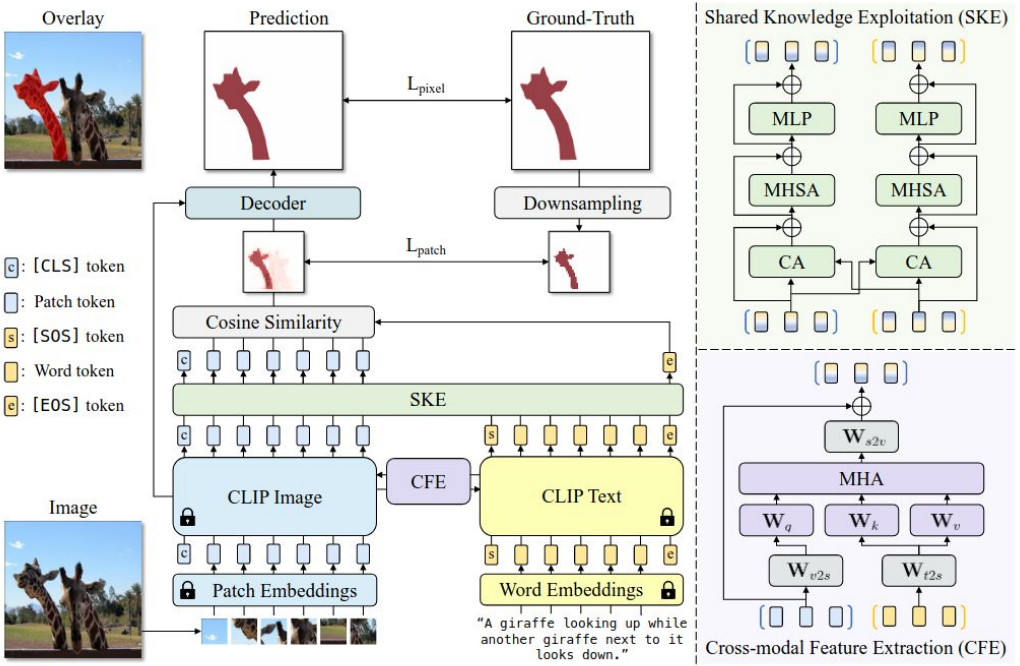


Figure 2: The overall pipeline of RISCLIP. We adopt frozen CLIP image and text encoders as backbones to exploit their aligned image and text features and adapt them to RIS with two modules, CFE and SKE. Firstly, the CFE modules between the encoders enable cross-modal commnuication between the two encoders to align their unimodal features at candidate regions. Secondly, the SKE modules on top of the encoders leverage the rich cross-modal alignment knowledge in CLIP's image-text shared embedding space to discern the target referent. Then a cosine similarity between the patch- and sentence-level features produces a patch-level grounding map. Lastly, a decoder refines the map into a pixel-level segmentation prediction.

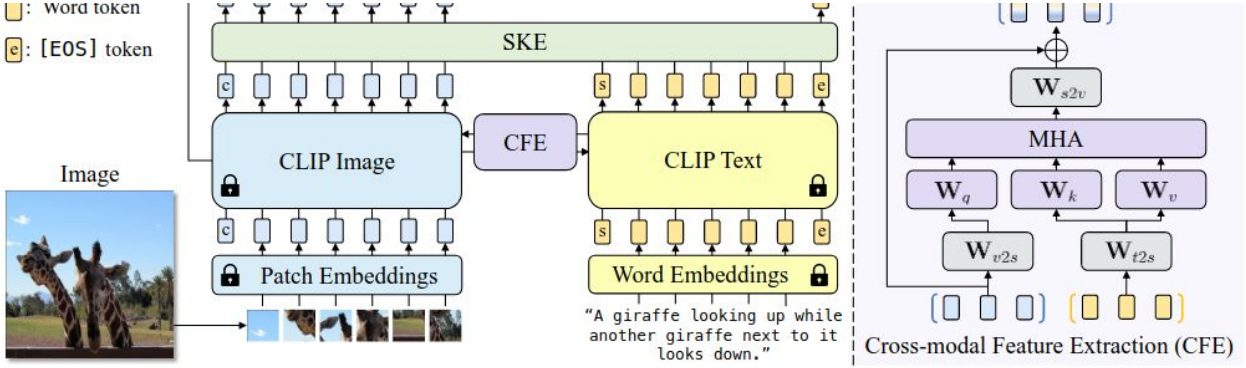
- **CLIP for RIS**
  - **Feature extraction**
    CLIP의 각 인코더의 아웃풋인, 패치 임베딩들과 문장 임베딩(EOS 토큰) 간에 코사인 유사도를 통해 patch-level features 획득
  - **Patch-level grounding map**
    MaskCLIP [1]에 기반해, 트랜스포머 내 각 블록에서 MHSA 연산한 임베딩 대신에 value 토큰을 사용
    이렇게 획득한 초기 patch-level grounding map의 mIoU: 23.86 (데이터셋: RefCOCOg-UMD test set)
  - **Adapters**
    구조: down-projection linear layer -> non-linear activation -> up-projection linear layer
    위치: 트랜스포머 내 각 블록의 MHSA 모듈과 MLP 모듈에 residually attached.
    어댑터 붙이니 mIoU: 48.29

- **Cross-modal Feature Extraction (CFE)**
  - **과정**
    이미지/텍스트 인코더 내 트랜스포머 각 블록의 feature들끼리 cross-attention
    1. Query: image feature, Key, value: text feature
    2. Query: text feature, Key, value: image feature
  - **효과**
    CLIP의 unimodal feature들에 cross-modal 정보 담음으로서, target patch-level features와 sentence-level features의 더 나은 정렬 도모, i.e., image feature and text feature containing cross-modal information
  - **결과**
    mIoU: 60.58

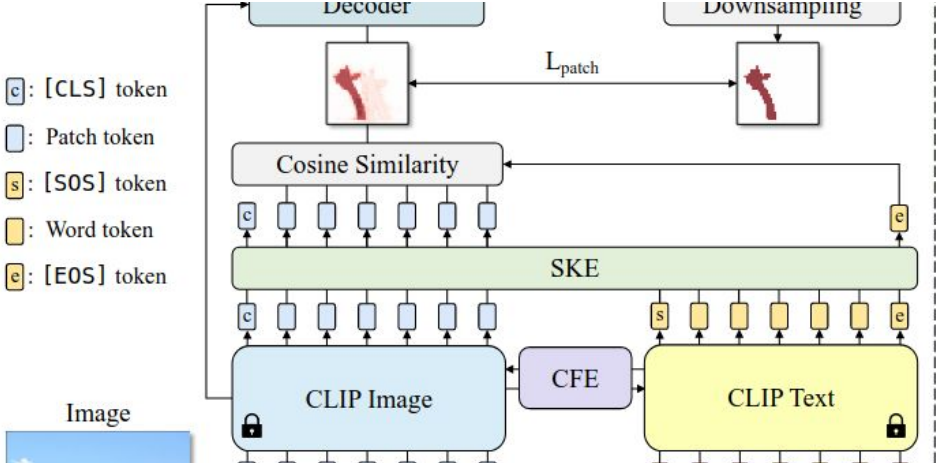
$$\mathbf{t}_k^s = \mathbf{W}_{t2s}\mathbf{t}_k, \quad \mathbf{v}_l^s = \mathbf{W}_{v2s}\mathbf{v}_l, \tag{1}$$

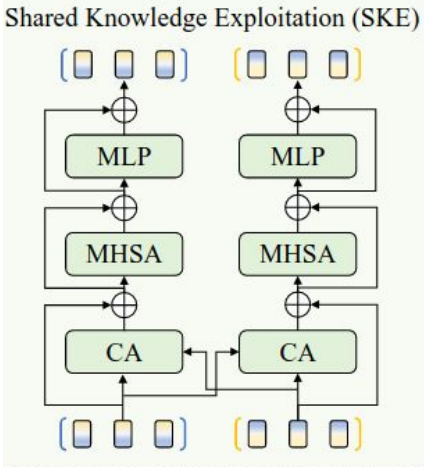
$$\mathbf{t}_k^m = \mathrm{MHCA}(\mathbf{t}_k^s, \mathbf{v}_l^s, \mathbf{v}_l^s) \tag{2}$$

$$\mathbf{t}_k^{m'} = \mathbf{W}_{s2t}\mathbf{t}_k^m. \tag{3}$$



- **Shared-space Knowledge Exploitation (SKE)**
  - **문제**
    복잡한 텍스트(e.g., 여러 객체들 간 복잡한 관계가 기술된 텍스트)에 해당하는 타겟 놓쳐
  - **해결**
    CLIP의 잘 정렬된 image-text shared embedding space를 활용하겠다!
    SKE 모듈을 통해, image/text feature들이 CLIP의 shared embedding space 내에서 inter-/intra- modal conditioning함.
  - **효과**
    1. Intra-modal interaction: model grasp intricate relationships b/w objects within in each modality.
    2. Inter-modal interaction: cross-modal alignment and grounding.
  - **결과**
    mIoU: 62.64

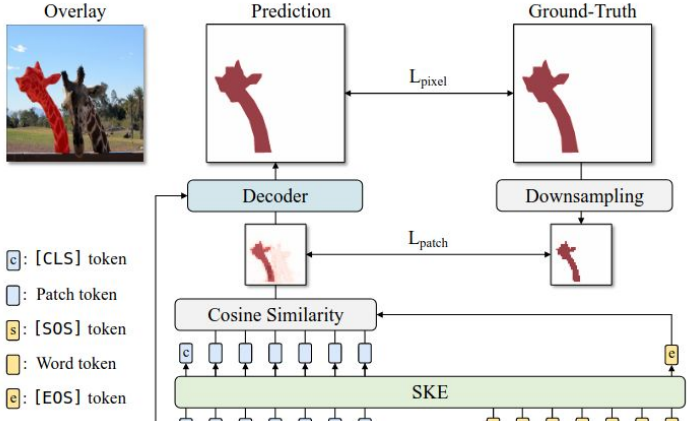


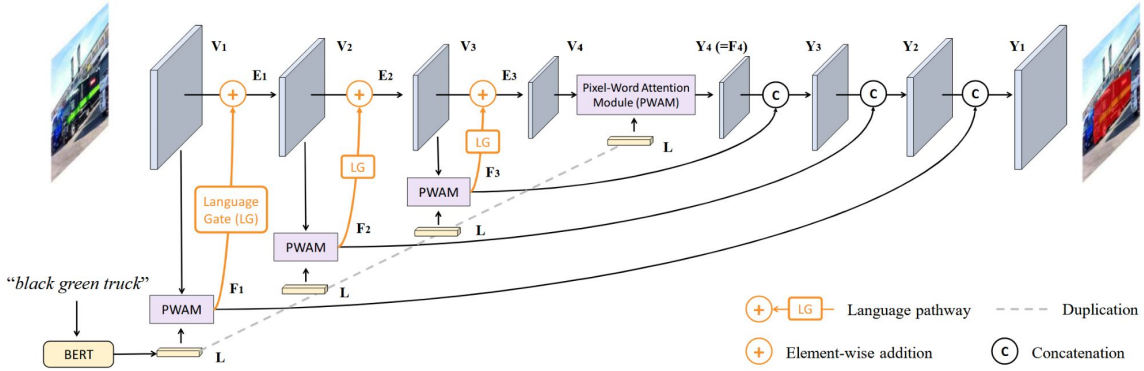


- **Decoder**
  - ❑ 구조
    - LAVT [1]의 디코더 구조 사용
      - 4 layers로 구성
      - 각 layer 구조: {3x3 conv -> RELU -> BN} x2 -> {Upsampling} x2
      - CLIP의 1~4 layer의 intermediate features 사용

$$\mathbf{d}_4 = \mathbf{D}_4([\mathbf{v}_4; \mathbf{map}_{\text{patch}}])$$

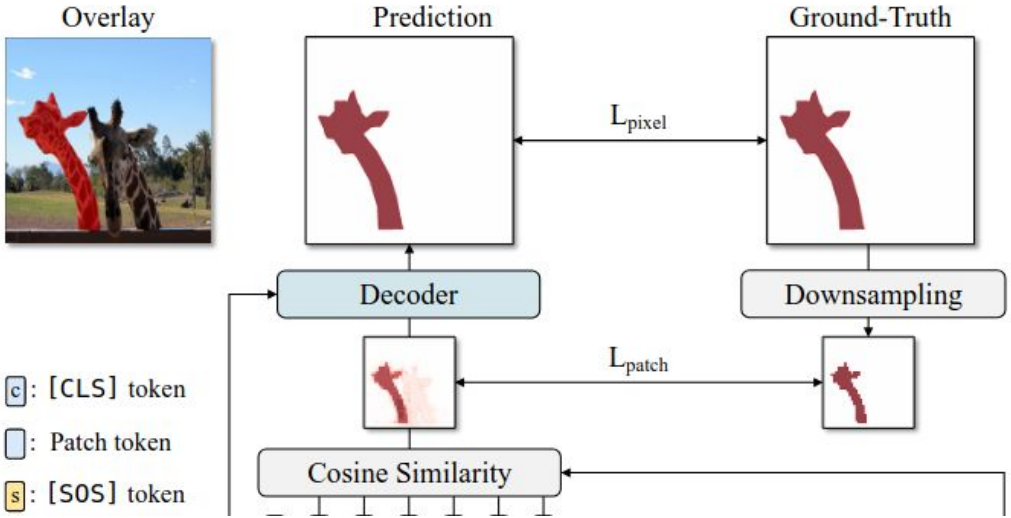
$$\mathbf{d}_i = \mathbf{D}_i([\mathbf{d}_{i+1}; u(\mathbf{v}_i)]), i = 1, 2, 3,$$





[1] YANG, Zhao, et al. Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 18155-18165.

- **Loss functions**
  - **Two-stage training**
    - Stage1: Patch-level training
      - Train target: Adapters, CFE, SKE modules
      - Loss b/w patch-level prediction map and downsampled ground-truth mask
    - Stage2: Pixel-level training
      - Train target: Decoder
      - Loss b/w pixel-level prediction map and ground-truth mask
    - Loss: Combination of DICE/F-1 loss and focal loss for both training stages



- **Results**

Figure 3: Visualization of RISCLIP-B predictions on RefCOCOg-UMD (Nagaraja et al., 2016) test set. Row a) shows RISCLIP's understanding of various instances, row b) RISCLIP's detection of partial, blurry instances and differentiate similar objects, row c) RISCLIP's discernment of the target instance among resembling instances described by lengthy texts.