ECCV 2024

MoAI: Mixture of All Intelligence for Large Language and Vision Models

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro

School of Electrical Engineering
Korea Advanced Institute of Science and Technology (KAIST)
{leebk, bpark0810, chaewonkim, ymro}@kaist.ac.kr

- Problem / objective
  - 최신 LLVM 들의 Vision 정보 활용 부족
    (연구들이 보통 downstream task 에 맞춰 Instruction tuning 및 모델 Scaling 에만 집중)
- Contribution / Key idea
  - Vision 정보 적극 활용하여 새로운 LLVM 인 MoAI 제안

전유진

어떻게 적극 활용?

Out[MLLM] : 1) visual features, 2) language features

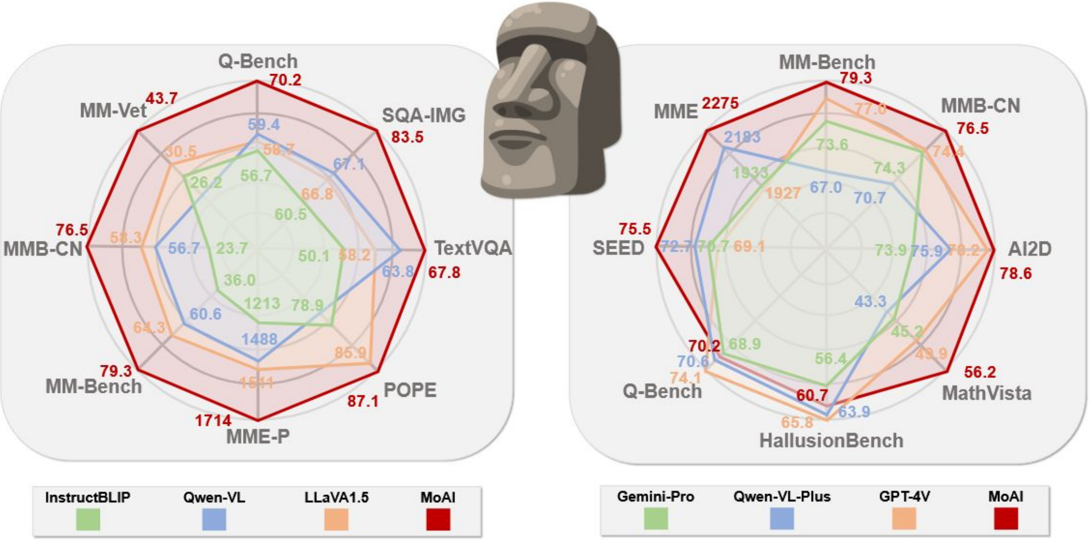
Out[외부 CV 모델들 -> MoAI-Compressor] : 3) auxiliary visual features

Out[MoAI-Mixer] : 6명의 전문가들이 위 3종류 feature 들을 잘 융합함으로서, 시각 인지 능력 크게 향상.

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

결과부터 보면

괴장히 좋아졌다!



(a) Open-source LLVMs

(b) Closed-source LLVMs

Fig. 1: Comparing the scores and accuracies of numerous VL benchmarks for various open-source and closed-source LLVMs with those for **MoAI**.

전유진

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

ECCV 2024

결과부터 보면

괴장히 좋아졌다!

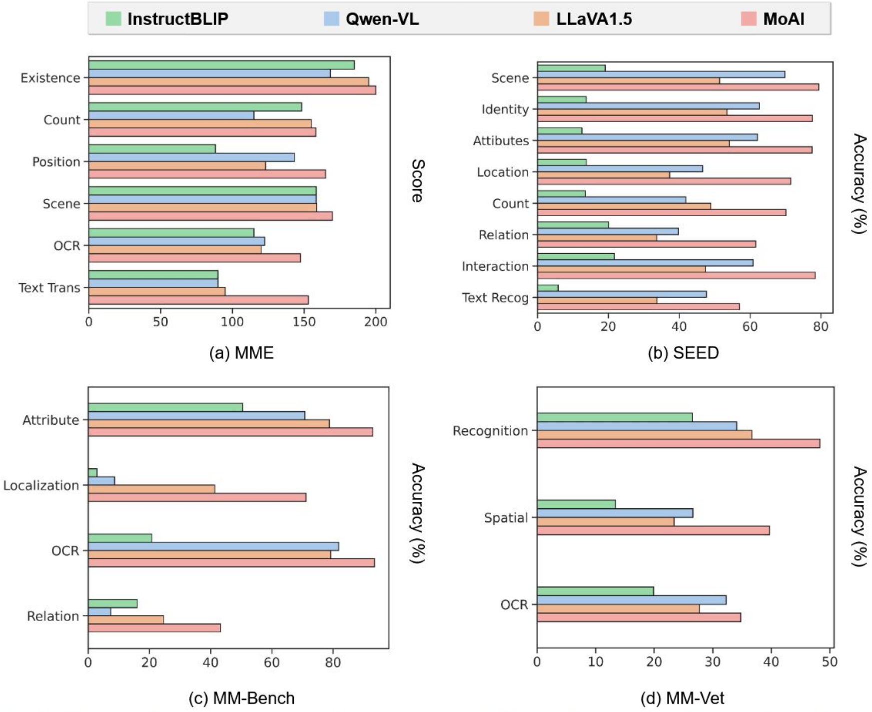


Fig. 2: Comparing the scores and accuracies of dimensions related to real-world scene understanding in MME [28], SEED [55], MM-Bench [66], and MM-Vet [91] for validating capabilities of various LLVMs such as InstructBLIP [19], Qwen-VL [4], and LLaVA1.5 [63].

전유진

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

Overview

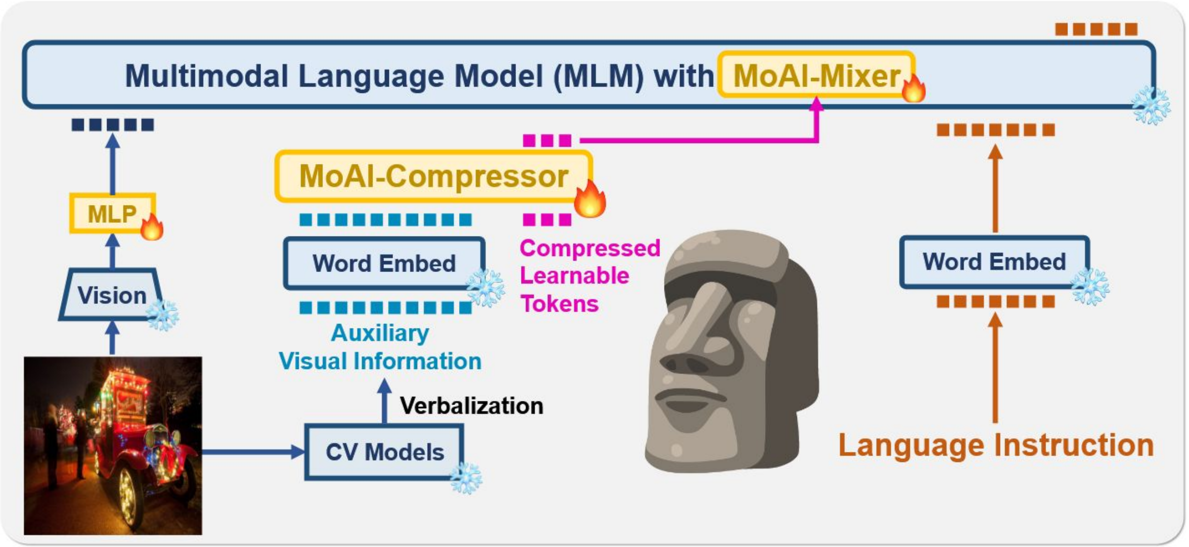


Fig. 3: Overview of **MoAI** architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

전유진

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

Architecture

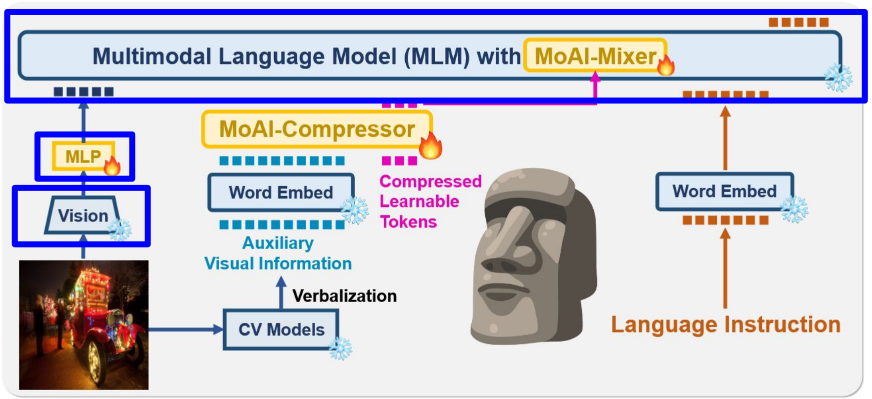


Fig. 3: Overview of **MoAI** architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

- Vision encoder: CLIP-L/14
- Multimodal language model: InternLM2-7B
- MLP: 2 linear layers with GELU 활성 함수

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vi
Nature Switzerland, 2024.

ger
024

Verbalization

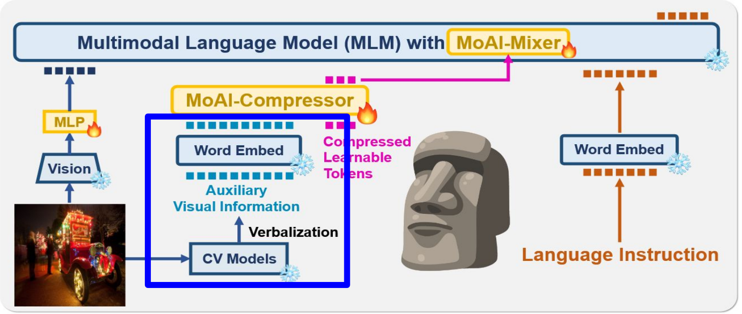


Fig. 3: Overview of MoAI architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

- 외부 Vision 모델 4개 사용하여,
  - Panoptic Segmentation (PS)
  - Open-World Object Detection (OWOD)
  - Scene Graph Generation (SGG)
  - Optical Character Recognition (OCR)
- 4개의 보조 visual tokens 생성.

$$A_{\text{PS}},\ A_{\text{OWOD}},\ A_{\text{SGG}},\ \text{and}\ A_{\text{OCR}}$$

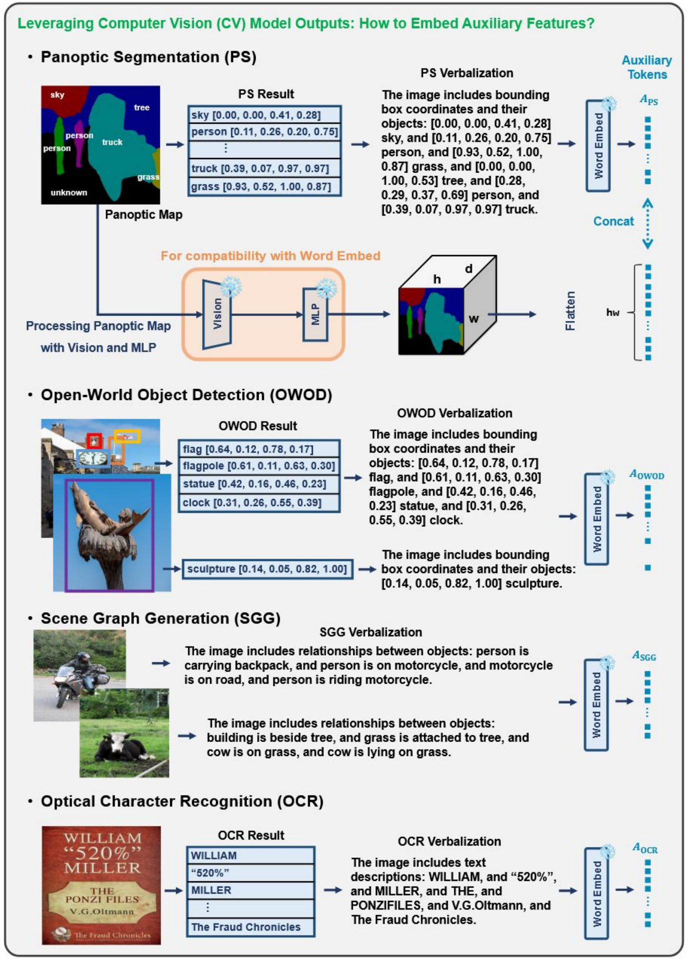


Fig. 4: Verbalization process of MoAI for external CV models: panoptic segmentation (PS), open-world object detection (OWOD), scene graph generation (SGG), and optical character recognition (OCR). Note that, 'd' denotes channel dimension of MLM, thus auxiliary tokens have equal channel dimension.

유진

MoAI-Compressor

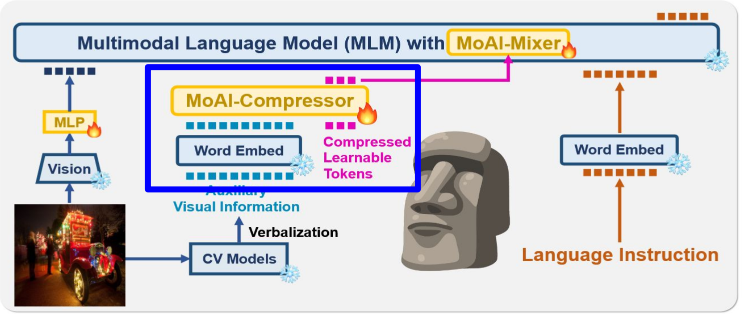


Fig. 3: Overview of MoAI architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

- 구조 : Perceiver Resampler
- 인풋 : 앞서 생성된 4개의 보조 토큰들 concat 한것 + 고정된 길이의 학습가능 토큰
- 아웃풋 : 똑같이 고정된 길이의 토큰

$$\boxed{A} = \text{MoAI-Compressor}(\boxed{[A_{\text{PS}}, A_{\text{OWOD}}, A_{\text{SGG}}, A_{\text{OCR}}]}, \boxed{A_{\text{input}}}). \quad (1)$$

$A \in \mathbb{R}^{d\times 64}$ 가변길이 $A \in \mathbb{R}^{d\times 64}$

전유진

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." *Advances in neural information processing systems* 35 (2022): 23716-23736.

[Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.]
ECCV 2024

MoAI-Mixer

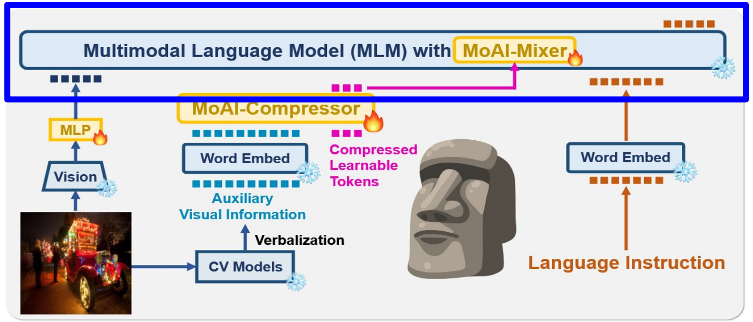


Fig. 3: Overview of MoAI architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

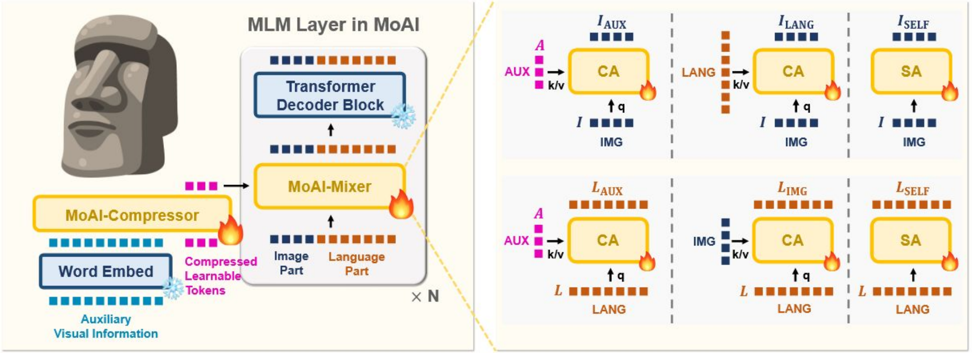


Fig. 5: Illustrating MoAI-Mixer in MLM Layer of MoAI. In MoAI-Mixer, there are six expert modules to harmonize auxiliary features A and two original features (*i.e.*, visual I and language L features).

- MoAI-Mixer 는 MLM 에 임베딩되어있음.
- 인풋 : (MoAI-Compressor 의 아웃풋인) 보조 토큰들, Visual features, Language features

$$A \in \mathbb{R}^{d\times 64} \qquad I^{(l)} \in \mathbb{R}^{d\times N_I} \quad L^{(l)} \in \mathbb{R}^{d\times N_L}$$

- l-th MLM layer with MoAI-Mixer :

$$[\hat{I}^{(l)}, \hat{L}^{(l)}] = \text{MoAI-Mixer}^{(l)}(A, I^{(l)}, L^{(l)}),$$
$$[I^{(l+1)}, L^{(l+1)}] = \text{TransDec}^{(l)}(\hat{I}^{(l)}, \hat{L}^{(l)}), \qquad (2)$$

전유진

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

ECCV 2024

MoAI-Mixer

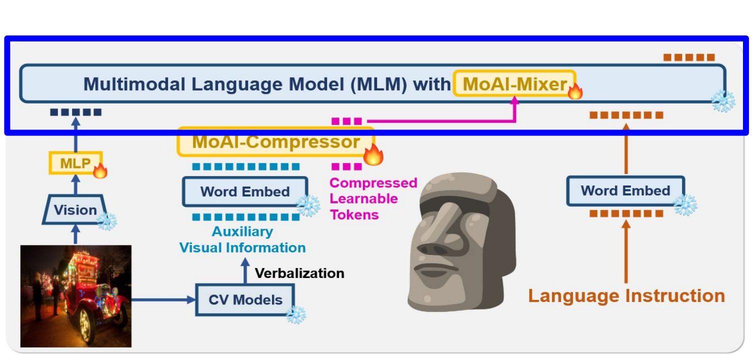


Fig. 3: Overview of MoAI architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

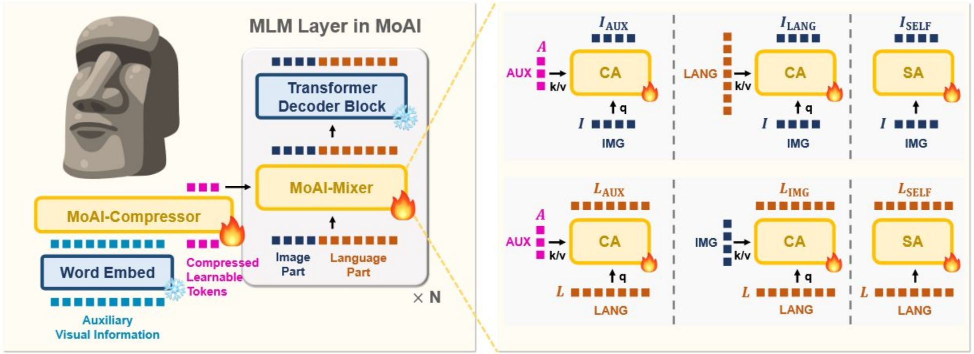


Fig. 5: Illustrating MoAI-Mixer in MLM Layer of MoAI. In MoAI-Mixer, there are six expert modules to harmonize auxiliary features A and two original features (*i.e.*, visual I and language L features).

- MoAI-Mixer 에는 6개의 전문가 모듈 존재
  - 구조 : Cross-attention, Self-attention
  - 인풋 : I / L (Visual features / Language features)
  - 아웃풋 : IAUX, ILANG, and ISELF / LAUX, LIMG, and LSELF

$$I^{(l)}_{\{\text{AUX or LANG}\}} = \text{CA}^{(l)}(q = I^{(l)}, k = \{A \text{ or } L^{(l)}\}, v = k),$$
$$L^{(l)}_{\{\text{AUX or IMG}\}} = \text{CA}^{(l)}(q = L^{(l)}, k = \{A \text{ or } I^{(l)}\}, v = k). \quad (3)$$

전유진

Lee, Byung-Kwan, et al. "Moai: Mixture of all intelligence for large language and vision models." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
ECCV 2024

MoAI-Mixer

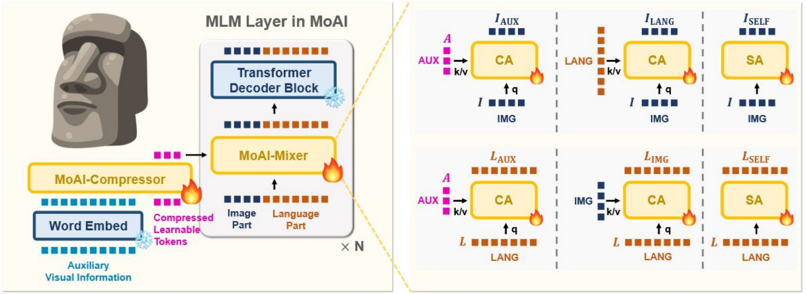
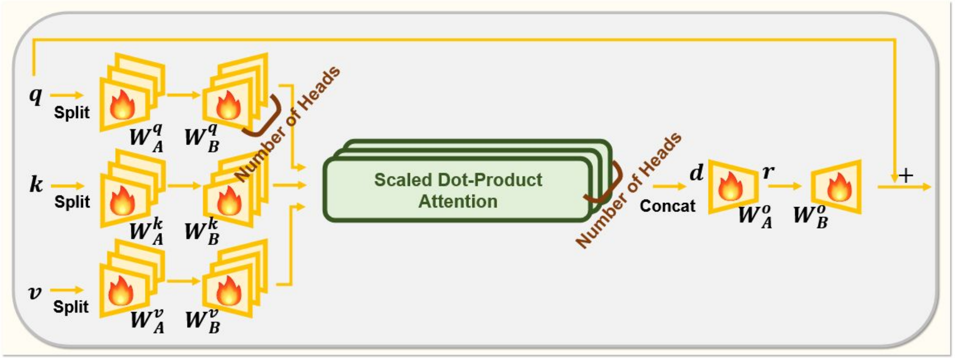


Fig. 5: Illustrating MoAI-Mixer in MLM Layer of **MoAI**. In MoAI-Mixer, there are six expert modules to harmonize auxiliary features A and two original features (*i.e.*, visual I and language L features).



(a) CA/SA with Low Rank Adaptation (LoRA) for Expert Modules

- MoAI-Mixer 에는 6개의 전문가 모듈 존재
  - 구조 : Cross-attention, Self-attention
    - Projection matrix W 를 LoRA 기반 decompose 해서 연산량 줄임.
    - Residual addition 통해 트랜스포머 디코더 최적화 과정 안정화시킴.

First Training Step - 즉, 전문가 친구들 독립적으로 학습

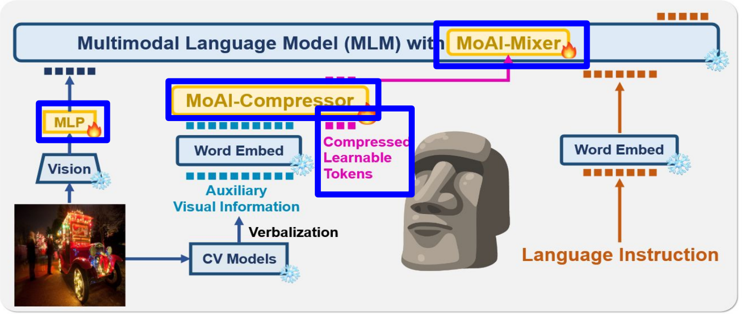


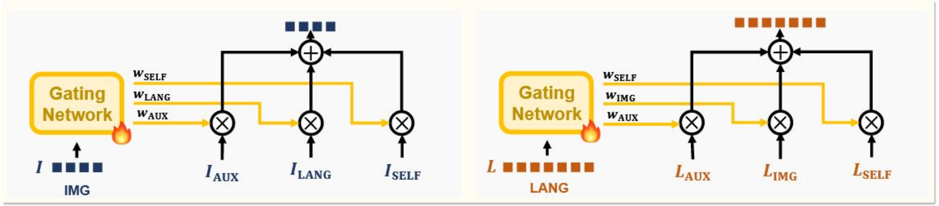
Fig. 3: Overview of **MoAI** architecture. Compressed learnable tokens, the parameters of MoAI-Compressor and MoAI-Mixer are learned. 'Vision' represents vision encoder to embed visual features and ice/fire symbols represent the modules to freeze or learn. Note that, 'Word Embed' represents the word embedding dictionary of MLM.

- Visual instruction tuning 데이터셋으로 MLP connector, Ainput, MoAI-Compressor, and MoAI-Mixer 4개 학습.
- MoAI-Mixer 내 6명의 전문가의 아웃풋인, 각각 3개의 visual/language features 중 각각 1개 랜덤 샘플링해서 학습.

$$\hat{I}^{(l)} = \mathrm{Sample}(I^{(l)}_{\mathrm{AUX}}, I^{(l)}_{\mathrm{LANG}}, I^{(l)}_{\mathrm{SELF}}), \quad \hat{L}^{(l)} = \mathrm{Sample}(L^{(l)}_{\mathrm{AUX}}, L^{(l)}_{\mathrm{IMG}}, L^{(l)}_{\mathrm{SELF}}). \quad (4)$$

$$\mathrm{TransDec}_l(\hat{I}^{(l)}, \hat{L}^{(l)})$$

Second Training Step - 이번엔, 학습된 전문가들 조합할 수 있도록 Gating Networks 학습



(b) Gating Networks for MoAI-Mixer

- 2개의 gating networks 존재 (Visual features 용도의 single layer, Language features 용도의 single layer)

$$W_{\text{Gating}_I} \text{ and } W_{\text{Gating}_L} \in \mathbb{R}^{d\times 3}$$

- 원래의 feature 와 MoAI-Mixer 가 생성한 3개의 features 간 weight 구해서 적용

$$\begin{aligned}
&[w_{\text{AUX}}, w_{\text{LANG}}, w_{\text{SELF}}] \leftarrow \text{Softmax}({I^{(l)}}^{\mathsf{T}} W_{\text{Gating}_I}, \text{dim}{=}1),\\
&\hat{I}^{(l)} = w_{\text{AUX}} \odot I^{(l)}_{\text{AUX}} + w_{\text{LANG}} \odot I^{(l)}_{\text{LANG}} + w_{\text{SELF}} \odot I^{(l)}_{\text{SELF}}\\
&[w_{\text{AUX}}, w_{\text{IMG}}, w_{\text{SELF}}] \leftarrow \text{Softmax}({L^{(l)}}^{\mathsf{T}} W_{\text{Gating}_L}, \text{dim}{=}1),\\
&\hat{L}^{(l)} = w_{\text{AUX}} \odot L^{(l)}_{\text{AUX}} + w_{\text{IMG}} \odot L^{(l)}_{\text{IMG}} + w_{\text{SELF}} \odot L^{(l)}_{\text{SELF}},
\end{aligned} \tag{5}$$

Experiments

- External CV models
  - panoptic segmentation : Mask2Former w/ Swin-B/4
  - open-world object detection : OWLv2 w/ CLIP-B/16
  - scene graph generation : panoptic SGG w/ ResNet-50
  - ocr : PaddleOCRv2
  - -> 위의 파라미터 다 합치면 332 M 개로 괜춘.