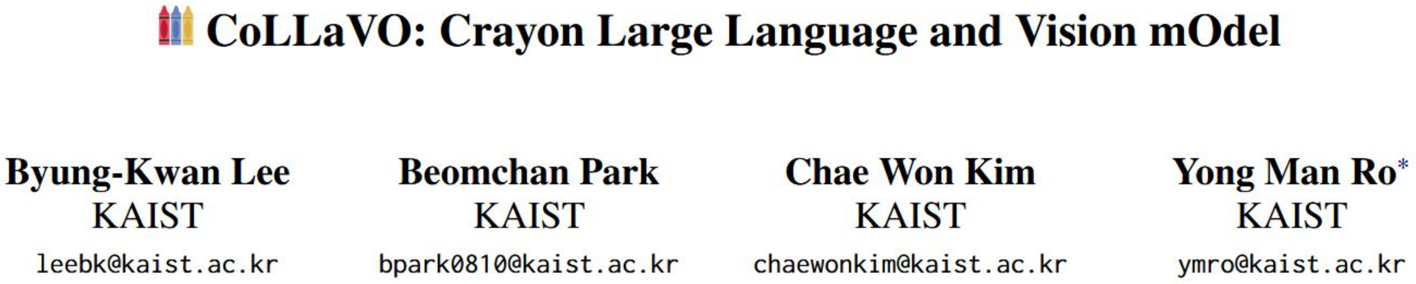
ACL 2024



CoLLaVO: Crayon Large Language and Vision mOdel

Byung-Kwan Lee
KAIST
leebk@kaist.ac.kr

Beomchan Park
KAIST
bpark0810@kaist.ac.kr

Chae Won Kim
KAIST
chaewonkim@kaist.ac.kr

Yong Man Ro*
KAIST
ymro@kaist.ac.kr

- Problem / objective

  VLM 이 object-level image understanding 잘했으면 좋겠어

- Contribution / Key idea
  - Crayon Large Language and Vision model (CoLLaVO) 제안
    - Crayon Prompt
    - Dual QLoRA

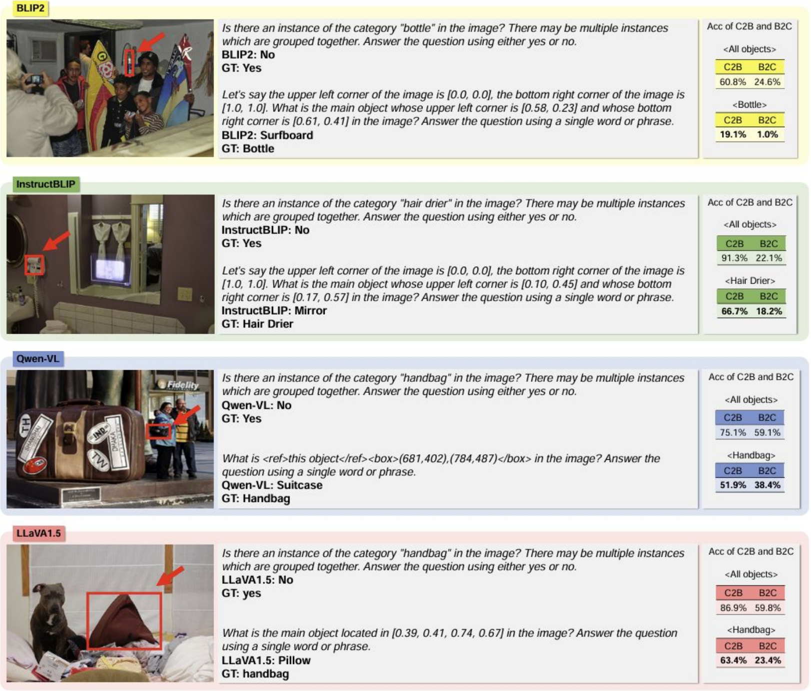
전유진

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

ACL 2024

[문제] VLM 들이 object-level image understanding 잘 못하더라

평가지표 : B2C, C2B 정확도



BLIP2

Is there an instance of the category "bottle" in the image? There may be multiple instances which are grouped together. Answer the question using either yes or no.
BLIP2: No
GT: Yes

Let's say the upper left corner of the image is [0.0, 0.0], the bottom right corner of the image is [1.0, 1.0]. What is the main object whose upper left corner is [0.58, 0.23] and whose bottom right corner is [0.61, 0.41] in the image? Answer the question using a single word or phrase.
BLIP2: Surfboard
GT: Bottle

Acc of C2B and B2C

	C2B	B2C
<All objects>	60.8%	24.6%
<Bottle>	**19.1%**	**1.0%**

InstructBLIP

Is there an instance of the category "hair drier" in the image? There may be multiple instances which are grouped together. Answer the question using either yes or no.
InstructBLIP: No
GT: Yes

Let's say the upper left corner of the image is [0.0, 0.0], the bottom right corner of the image is [1.0, 1.0]. What is the main object whose upper left corner is [0.10, 0.45] and whose bottom right corner is [0.17, 0.57] in the image? Answer the question using a single word or phrase.
InstructBLIP: Mirror
GT: Hair Drier

Acc of C2B and B2C

	C2B	B2C
<All objects>	91.3%	22.1%
<Hair Drier>	**66.7%**	**18.2%**

Qwen-VL

Is there an instance of the category "handbag" in the image? There may be multiple instances which are grouped together. Answer the question using either yes or no.
Qwen-VL: No
GT: Yes

What is <ref>this object</ref><box>(681,402),(784,487)</box> in the image? Answer the question using a single word or phrase.
Qwen-VL: Suitcase
GT: Handbag

Acc of C2B and B2C

	C2B	B2C
<All objects>	75.1%	59.1%
<Handbag>	**51.9%**	**38.4%**

LLaVA1.5

Is there an instance of the category "handbag" in the image? There may be multiple instances which are grouped together. Answer the question using either yes or no.
LLaVA1.5: No
GT: yes

What is the main object located in [0.39, 0.41, 0.74, 0.67] in the image? Answer the question using a single word or phrase.
LLaVA1.5: Pillow
GT: handbag

Acc of C2B and B2C

	C2B	B2C
<All objects>	86.9%	59.8%
<Handbag>	**63.4%**	**23.4%**

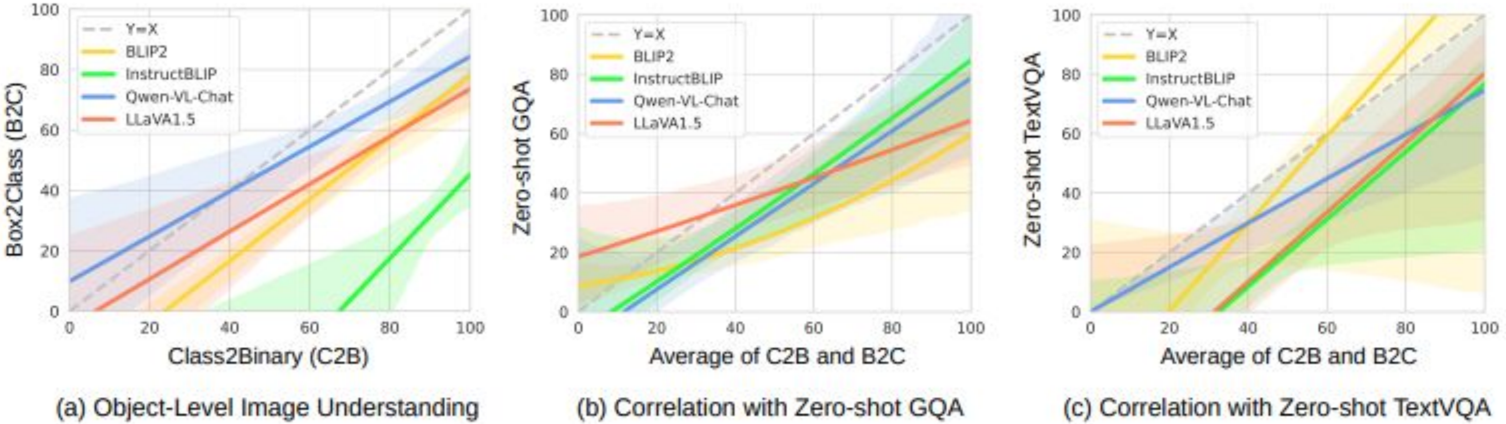
Figure 2: Asking four baselines (BLIP2, InstructBLIP, Qwen-VL, and LLaVA1.5) two types of questions, Class2Binary (C2B) and Box2Class (B2C), and measuring their accuracies on each object category.

전유진

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

ACL 2024

[사실] 근데 VLM 들의 object-level image understanding 이 zero-shot 성능과 관련 있는 거 알어?



(a) Object-Level Image Understanding (b) Correlation with Zero-shot GQA (c) Correlation with Zero-shot TextVQA

Figure 3: Plotting the regressed relationships between (a) C2B and B2C for each object category, (b) the average of C2B & B2C and zero-shot GQA (Hudson and Manning, 2019) performance for each object category, (c) the average of C2B & B2C and zero-shot TextVQA (Singh et al., 2019) performance for each object category to visualize their correlations. The light-colored areas indicate the vertical span with the probability of confidence interval 0.95.

전유진

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

ACL 2024

[결론] 그래서 우리는 VLM 이 object-level image understanding 잘하게 해서 zero-shot 성능 올릴거야.

- Crayon Prompt
  - Visual prompt for object-level image understanding
  - 구체적으로는, Panoptic colormap 의 semantic, instance 정보를 담은 학습가능한 쿼리 벡터
  - MLM 의 백본 내 모든 어텐션 모듈 층에서 이미지 피쳐에 Crayon Prompt 를 통합시킴.
- Dual QLoRA
  - 하나는 Crayon Instructions 에 대해 학습
  - 다른 하나는 Visual instruction tuning 데이터셋 에 대해 학습

Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in neural information processing systems* 36 (2023): 10088-10115.

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

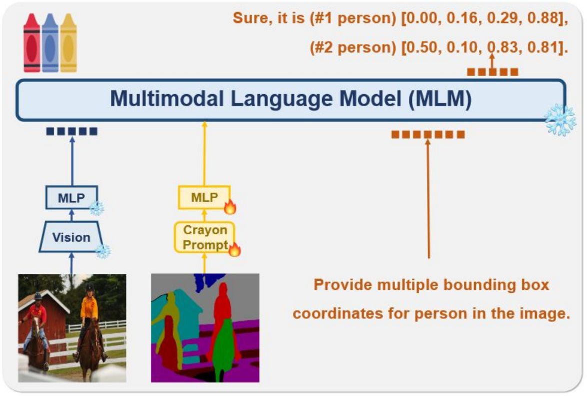
Model Architecture and Prompt Protocol

- Model Architecture
  - Vision encoder : CLIP
  - Crayon prompt
  - backbone MLM : InternLM-7B
  - MLP connectors : 2 fully-connected MLPs w/ GELU activation function
- Prompt Protocol
  - ‘<image>’ : a special token for image embedding features
  - ‘<stop>’ : a stop token for text generation
  - ‘User: {}’ : a question template
  - ‘Assistant: {}’ : an answer template

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

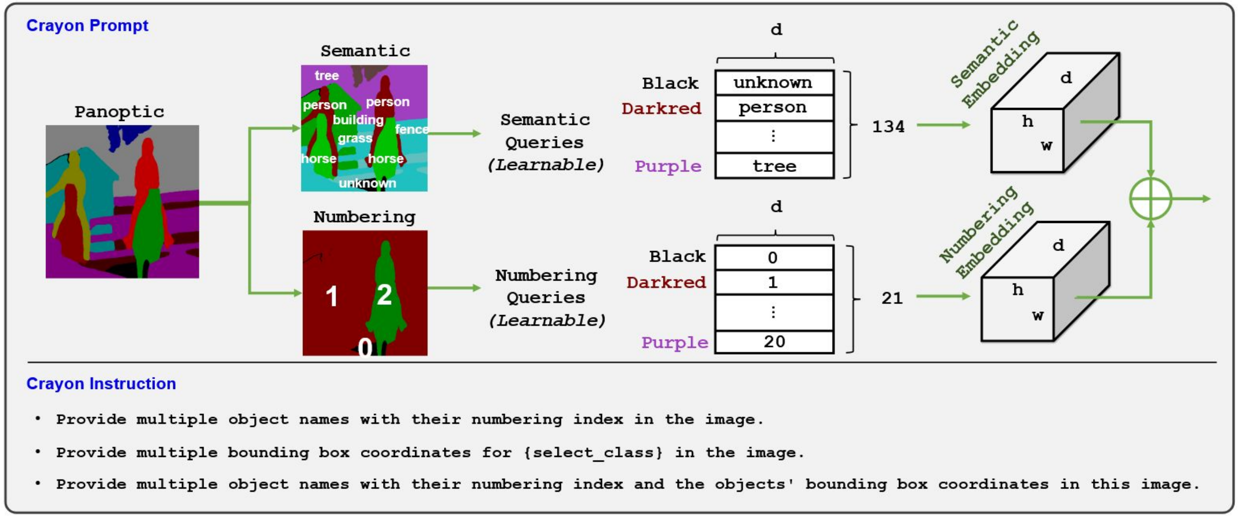
ACL 2024

Crayon Prompt Tuning (CPT)


Sure, it is (#1 person) [0.00, 0.16, 0.29, 0.88],
(#2 person) [0.50, 0.10, 0.83, 0.81].
Multimodal Language Model (MLM)
MLP
Vision
MLP
Crayon Prompt
Provide multiple bounding box coordinates for person in the image.

(a) Crayon Prompt Tuning (CPT)

Figure 4: Overview of two-step training for CoLLaVO. fire symbols represent the modules to learn.



- Provide multiple object names with their numbering index in the image.
- Provide multiple bounding box coordinates for {select_class} in the image.
- Provide multiple object names with their numbering index and the objects' bounding box coordinates in this image.

Figure 5: Describing how the Crayon Prompt is generated from a panoptic color map with learnable semantic queries and numbering queries. In addition, crayon instruction examples are given, which are used to conduct CPT and CIT. Note that, '{}' denotes the place where we adaptively input information.

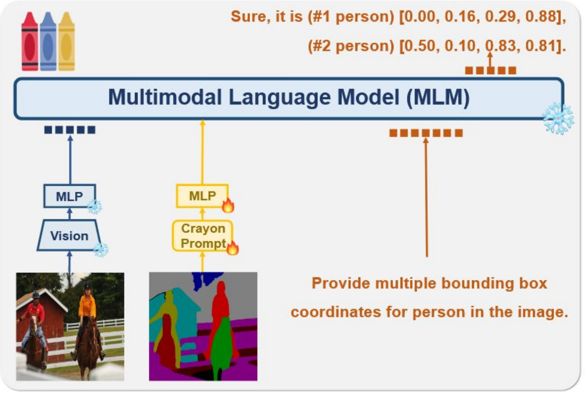
- 과정 : Image -> Panoptic semantic/numbering color map -> Semantic/Numbering queries -> Crayon prompt
- Learnable queries
  - 133+1(unk)semantic queries
  - 20+1('0' for unk)numbering queries (한 이미지 내 동일한 객체 최대 20개까지 존재한다는 가정)

전유진

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

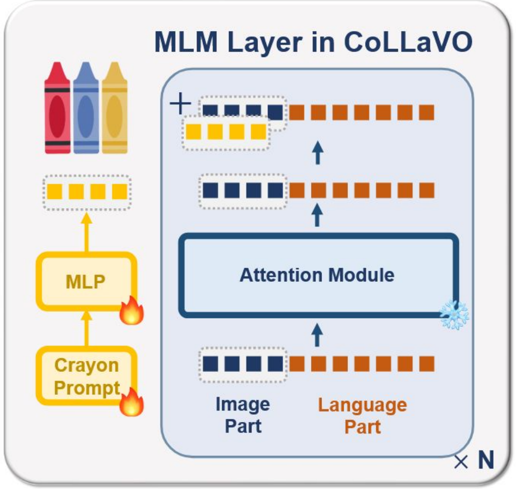
ACL 2024

Crayon Prompt Tuning (CPT)


Sure, it is (#1 person) [0.00, 0.16, 0.29, 0.88],
(#2 person) [0.50, 0.10, 0.83, 0.81].
Multimodal Language Model (MLM)
MLP
Vision
MLP
Crayon Prompt
Provide multiple bounding box coordinates for person in the image.

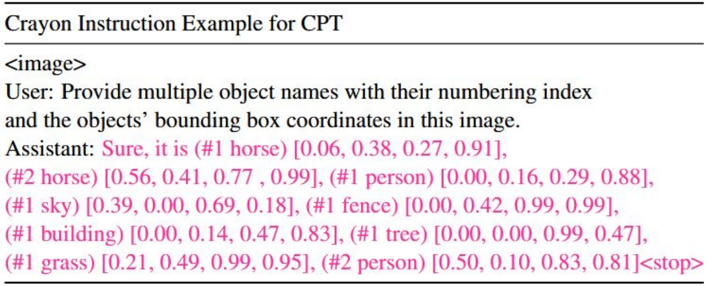
(a) Crayon Prompt Tuning (CPT)

Figure 4: Overview of two-step training for CoLLaVO. fire symbols represent the modules to learn.



(a) Crayon Prompt Operation in CoLLaVO

Figure 6: Illuminating (a) how the Crayon Prompt is injected i
of (b), (c) Dual QLoRA for the object-level image understan
(VL-CIT) to efficiently coexist without catastrophic forgetting



Crayon Instruction Example for CPT

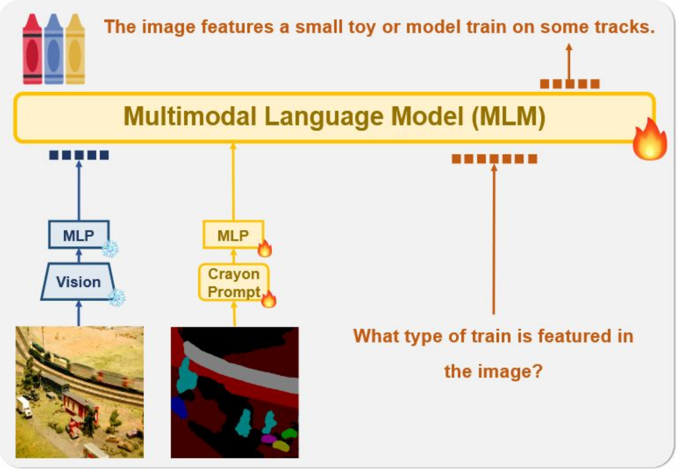
<image>
User: Provide multiple object names with their numbering index and the objects' bounding box coordinates in this image.
Assistant: Sure, it is (#1 horse) [0.06, 0.38, 0.27, 0.91], (#2 horse) [0.56, 0.41, 0.77 , 0.99], (#1 person) [0.00, 0.16, 0.29, 0.88], (#1 sky) [0.39, 0.00, 0.69, 0.18], (#1 fence) [0.00, 0.42, 0.99, 0.99], (#1 building) [0.00, 0.14, 0.47, 0.83], (#1 tree) [0.00, 0.00, 0.99, 0.47], (#1 grass) [0.21, 0.49, 0.99, 0.95], (#2 person) [0.50, 0.10, 0.83, 0.81]<stop>

- 학습 목적 : object-level image understanding, 데이터셋 : MS-COCO 2017 dataset
- Crayon prompt 는 MLP 거쳐서 attention layer 내의 모든 image features 에 더해짐
- Crayon instructions 사용해서 Crayon Prompt Tuning 함. (Crayon Prompt 를 backbone MLM 과 정렬 위해)전유진

Lee, Byung-Kwan, et al. "Collavo: Crayon large language and vision model." *arXiv preprint arXiv:2402.11248* (2024).

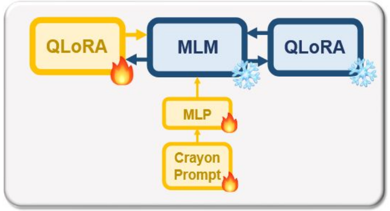
ACL 2024

Crayon Prompt-based Instruction Tuning (CIT)

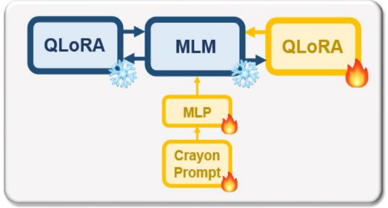


The image features a small toy or model train on some tracks.
Multimodal Language Model (MLM)
MLP
Vision
MLP
Crayon Prompt
What type of train is featured in the image?

(b) Crayon prompt-based Instruction Tuning (CIT)



(b) Dual QLoRA for Image-CIT



(c) Dual QLoRA for VL-CIT

- 학습 목적 : complex question answering, 데이터셋 : crayon instructions, visual instruction tuning datasets
- Dual QLoRA
  - Image-CIT : 첫번째 QLoRA 모듈만 학습, 학습 목표 : object-level image understanding
  - VL-CIT : 두번째 QLoRA 모듈만 학습, 학습 목표 : zero-shot VL performance