

Multi-Modal Large Language Models are Effective Vision Learners

Li Sun^{1*} Chaitanya Ahuja² Peng Chen² Matt D’Zmura²
Kayhan Batmanghelich¹ Philip Bontrager²
¹Boston University ²Meta

- Problem / objective
 - Visual representation learning
- Contribution / Key idea
 - LLM-augmented visual representation learning (**LMVR**)

Overview

기존의 Vision Encoder 를 통해 이미지 feature 추출하는 방법 대신, 우리는 LLM 사용해서 더 나은 이미지 feature 추출했다.

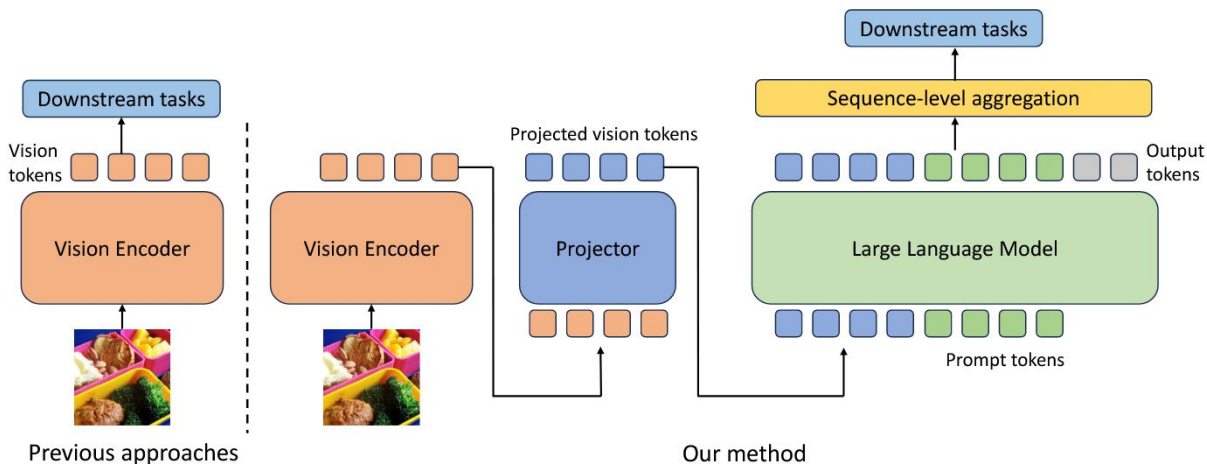


Figure 2. Illustration of our approach. Previous methods only use a vision encoder for representation learning. Our method is comprised of four steps. First, a vision encoder extracts features from images. Then, the extracted features are projected into the space of word embedding. Next, the LLM takes the projected visual features and a text prompt as input and generates a response. Finally, we perform sequence-level aggregation of features of hidden layers to obtain image-level representation.

Method

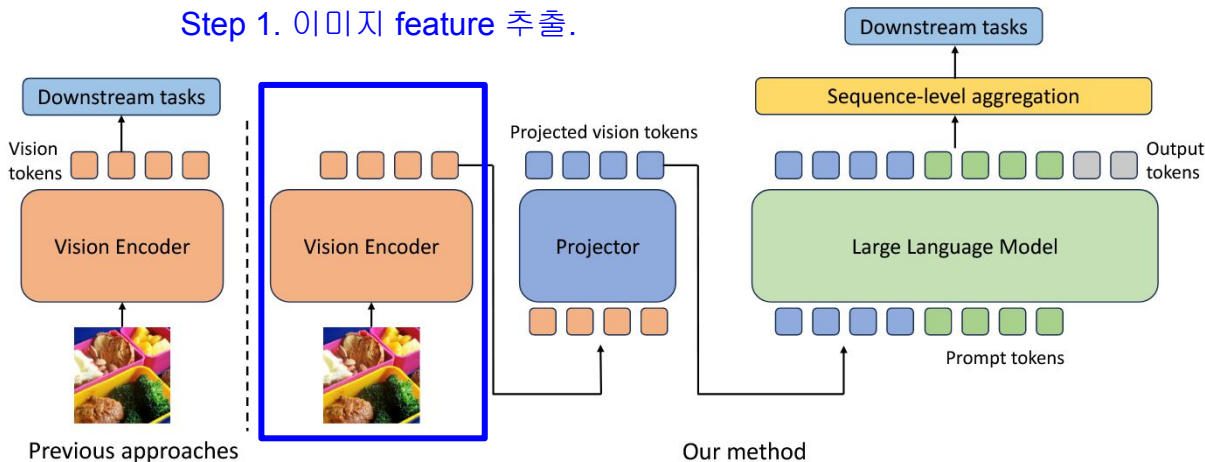


Figure 2. Illustration of our approach. Previous methods only use a vision encoder for representation learning. Our method is comprised of four steps. First, a vision encoder extracts features from images. Then, the extracted features are projected into the space of word embedding. Next, the LLM takes the projected visual features and a text prompt as input and generates a response. Finally, we perform sequence-level aggregation of features of hidden layers to obtain image-level representation.

Method

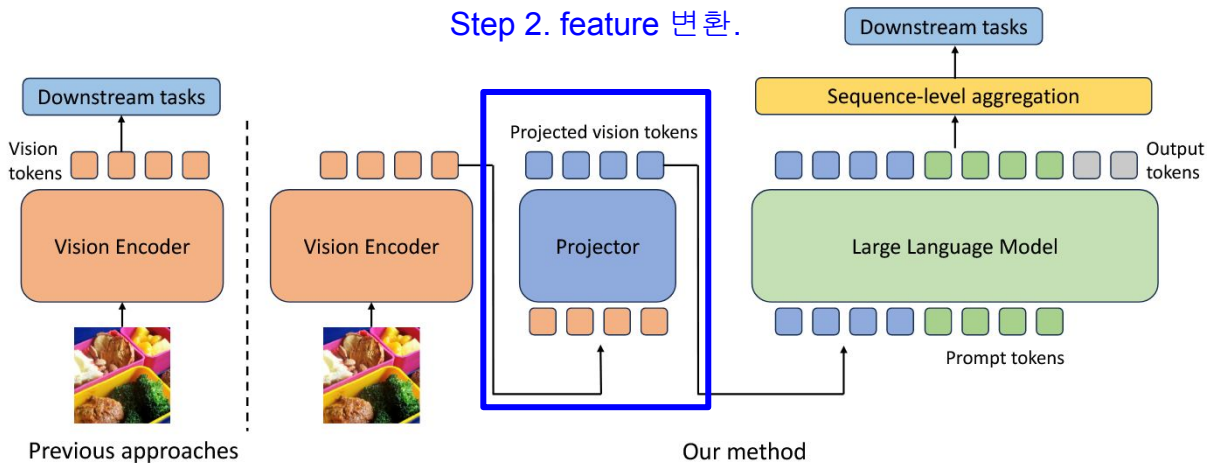


Figure 2. Illustration of our approach. Previous methods only use a vision encoder for representation learning. Our method is comprised of four steps. First, a vision encoder extracts features from images. Then, the extracted features are projected into the space of word embedding. Next, the LLM takes the projected visual features and a text prompt as input and generates a response. Finally, we perform sequence-level aggregation of features of hidden layers to obtain image-level representation.

Method

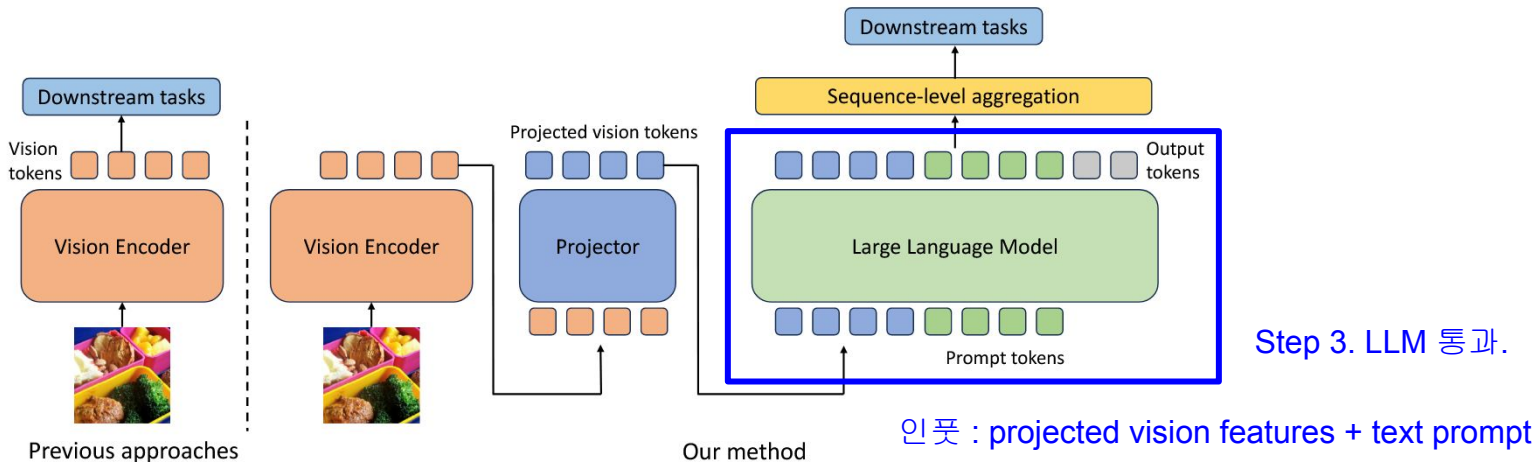


Figure 2. Illustration of our approach. Previous methods only use a vision encoder for representation learning. Our method is comprised of four steps. First, a vision encoder extracts features from images. Then, the extracted features are projected into the space of word embedding. Next, the LLM takes the projected visual features and a text prompt as input and generates a response. Finally, we perform sequence-level aggregation of features of hidden layers to obtain image-level representation.

Method

Step 4. token-level representation 을 image-level 로 통합.

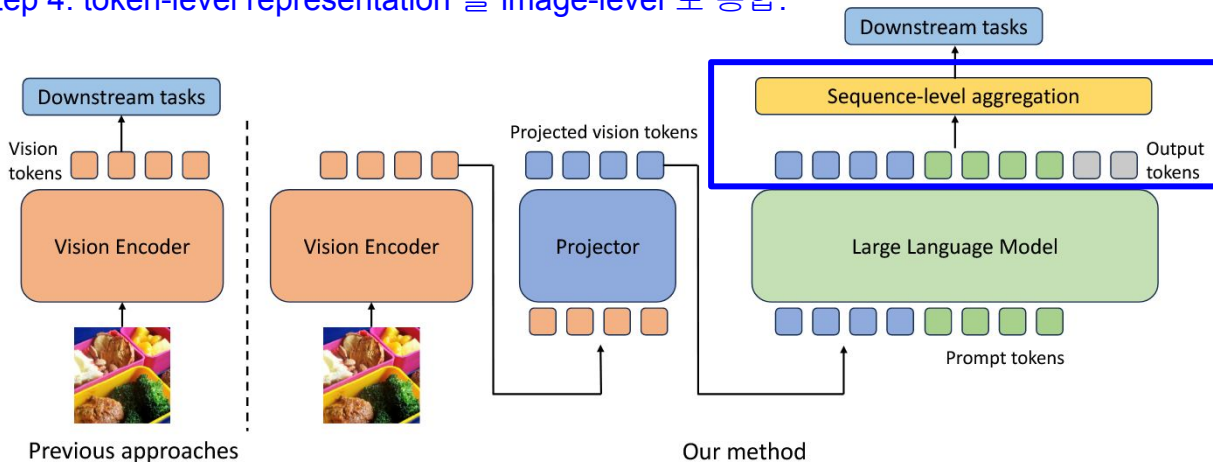


Figure 2. Illustration of our approach. Previous methods only use a vision encoder for representation learning. Our method is comprised of four steps. First, a vision encoder extracts features from images. Then, the extracted features are projected into the space of word embedding. Next, the LLM takes the projected visual features and a text prompt as input and generates a response. Finally, we perform sequence-level aggregation of features of hidden layers to obtain image-level representation.

Sequence-Level Aggregation

1. feature 추출할 토큰 : single token vs **aggregate many tokens**
2. aggregation 방법 : **average pooling** vs max pooling
3. feature 추출 위치 : **last hidden layer** vs previous layers
4. feature 추출 시점 : before generation vs **after generation**

Training

- 사용한 모델

Vision Encoder : CLIP (ViT-L)

projector : MLP

LLM : Vicuna-7B

- 학습 방법

LLaVA 의 학습 방식 그대로 따름. (두 단계 : 1. feature alignment -> 2. visual instruction tuning)

1단계. feature alignment

- VE, LLM 고정시키고, projector 만 1 에포크 학습.
- CC3M 데이터셋의 595K image-text pairs 사용

2단계. visual instruction tuning

- VE 고정시키고, projector와 LLM을 3 에포크 파인튜닝.
- LLaVA 로부터 instruction tuning data 사용

Experiments - Benchmark extracted visual features

- 평가 방식 : linear probe evaluation
- 데이터셋 : VOC07, COCO

Table 1. Benchmark the design choices in feature extraction with LLM on VOC07 dataset.

Method	mAP (%)
Only use embedding from last [END] token	90.5
Max pooling of embedding of all generated tokens	92.2
Average embedding of all generated tokens	93.5
Average embedding from all output tokens (second-to-last layer)	93.7
Average embedding from all output tokens	93.8

Table 2. Evaluation of component-wise comparison. We use the mean average precision (mAP) as the evaluation metric.

	VOC	COCO
Vision encoder (CLIP)	91.3%	79.5%
w/Adaptor	91.0%	78.5%
w/LLM	93.8%	84.0%

Experiments - [Ours 자량1] LMVR Learns Object-Level Concepts

LLM generated responses : *The image displays four plastic containers filled with various types of food, placed on a table. The containers hold a variety of items, including **meat**, **vegetables**, **fruit**, and **bread**. Some of the specific items include broccoli, which can be seen in multiple containers, and oranges, which are present in one of the containers.*

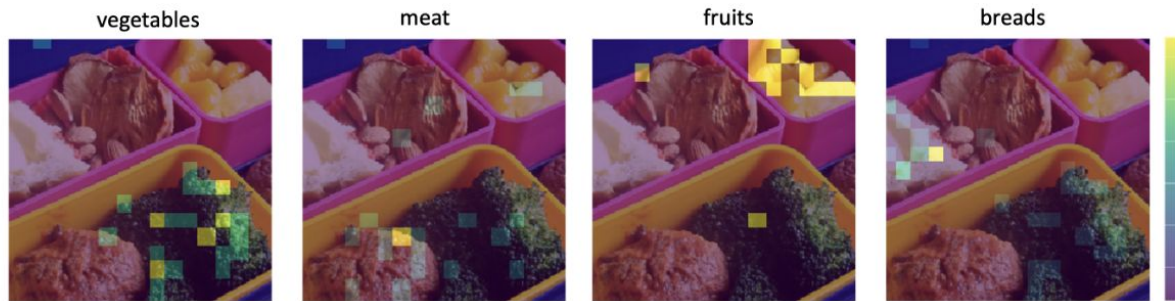


Figure 4. Results of token-level visualization. We overlay a sample image with attention maps between selected text tokens in generated response and vision tokens. The full generated response can be found in Section. 4.2. We discover that the attention map accurately highlights the area that matches each query word, showing that our model grasps multi-modal concepts at the object level.

Experiments - [Ours 자랑] LMVR Learns More Generalizable Visual Representation

- VisDA 2017 데이터셋으로 실험

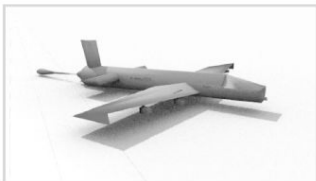
12 object categories

train set : synthetic images from 3D CAD models

test set : real images from YouTube BBox datasets

Table 3. Evaluation results on VisDA test set

Method	Accuracy
Vision encoder (CLIP)	60.9%
LMVR (Ours)	79.1%



LLM description:

The image features a **large grey airplane** with a pointy nose, sitting on top of a white paper plate. The airplane occupies a major portion of the scene, while the paper plate appears to be partially transparent and covers almost the entire base of the image. The plane seems to be a model, possibly a toy, due to its size and prominence in the image. There are no other objects or people visible in this scene.



LLM description:

The image shows a **red fighter jet** flying in a blue sky, with a few wispy clouds scattered across the sky. The jet appears to be at full throttle, and it is heading straight up into the sky.

Figure 5. Example of LLM generated responses on VisDA dataset. The results show that the LLM-generated description can identify the key objects in the images, as marked in red, even when the images belong to different domains.

Experiments - [Ours 자랑3] LMVR Learns More Robust Visual Representation

- ImageNet-C 데이터셋으로 실험

Table 4. Evaluation results on ImageNet-C dataset. Our proposed LMVR is more robust to different types of corruption compared to using an image encoder alone.

Accuracy(%)	Weather				Noise			Blur				Digital				
Methods	Brightness	Fog	Frost	Snow	Gaussian	Impulse	Shot	Defocus	Glass	Motion	Zoom	Contrast	Elastic	JPEG	Pixelate	Average
Vision encoder (CLIP)	79	67	56	65	62	57	60	49	36	60	49	67	69	69	68	61
LMVR (Ours)	81	71	57	67	64	61	62	62	45	68	54	75	71	74	75	66

Table 5. Evaluation results on ImageNet-C dataset, averaged by perturbation category

Accuracy(%)	Weather	Noise	Blur	Digital
Vision encoder (CLIP)	66.9	59.7	48.6	68.0
LMVR (Ours)	69.0	62.3	57.1	73.7

Experiments - [Ours 자량3] LMVR Learns More Robust Visual Representation




Clean		<p>LLM description:</p> <p>In the image, there is a young man kneeling down while holding a large fish in his hands. The fish is lying on its side, and the boy appears to be proud of his catch. The surroundings include a grassy area with green grass and trees in the background, suggesting that the boy might have been fishing in a nearby body of water.</p>
Gaussian Blur		<p>LLM description:</p> <p>In the image, there is a man standing in a grassy area holding a large fish in his hands. The fish appears to be a bass, and it is quite a sizeable catch. The man is posing with the fish, likely displaying his skills as a fisherman or simply enjoying the moment.</p>
Gaussian Noise		<p>LLM description:</p> <p>In the image, there is a young man kneeling down next to a large fish. The fish appears to be bass-like with its mouth wide open. The young man is posing with the fish, showcasing its size and catching a memorable moment. The scene likely takes place near a body of water where the fish was caught.</p>

Figure 6. Example of LLM generated responses under image corruption. The images are from the ImageNet-C dataset. We found even under a moderate amount of noise, the generated response from LLM is robust and can still capture the gist of the image, as highlighted in red. To preserve privacy, the person's face is masked.

Experiments - LMVR Boosts the Comprehension of Objects, Instead of Low-Level Textures

- DeepFashion-Multimodal 데이터셋으로 실험
- LMVR 이 더 generalizable 하고 robust 한 visual features 를 생성하는 이유 : LLM 이 low-level features 보다 high-level features 에 더 집중해서

Table 6. Evaluation results on DeepFashion-Multimodal dataset

Accuracy (%)	Object	Fabric	Texture
Vision encoder (CLIP)	89.1	73.8	71.3
LMVR	90.0	73.9	71.4
LMVR (fine-tuned)	91.2	73.9	71.2

Experiments - Generalizability to Other Architectures

Benchmark extracted visual features

Table 7. Evaluation of our method with BLIP-2 on VOC07 and COCO datasets

mAP	VOC	COCO
Vision encoder (EVA)	92.2%	82.3%
LMVR (Ours)	93.7%	84.5%

- 사용한 모델
Vision Encoder : EVA
projector : Q-former
LLM : FlanT5-XXL
- 학습 방법
BLIP-2 의 학습 방식 그대로 따름.

Experiments - [Ours 자량2] LMVR Learns More Generalizable Visual Representation

Table 8. Evaluation results on VisDA test set with BLIP-2

Method	Accuracy
Vision encoder (EVA)	72.1%
LMVR (Ours)	81.0%