LLaVA-SP: Enhancing Visual Representation with Visual Spatial Tokens for MLLMs

Haoran Lou[1], Chunxiao Fan[1,†], Ziyan Liu[1], Yuexin Wu[1], Xinliang Wang[2]

[1]Beijing University of Posts and Telecommunications , [2]Beihang University

Code link：https://github.com/CnFaker/LLaVA-SP

Motivation and Contribution

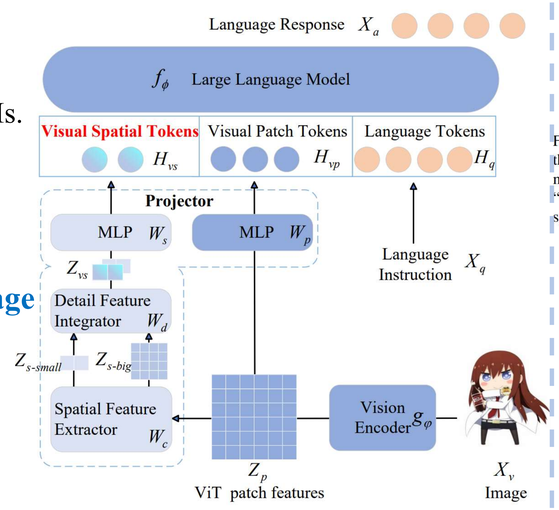
Motivation

- Current MLLMs use CLIP-ViT, which captures global but not local patch relationships.
- Recent improvements like dynamic resolution and multi-encoder fusion boost features but increase visual tokens.
- **Question: Can we enhance visual representation without adding many visual tokens?**

Contribution

1. Visual spatial tokens enhance visual representation for MLLMs.

A small set of visual spatial tokens can enhance visual-language understanding.



2. Two model variants handle diverse tasks.

3. Performance improvements on various multimodal benchmarks.

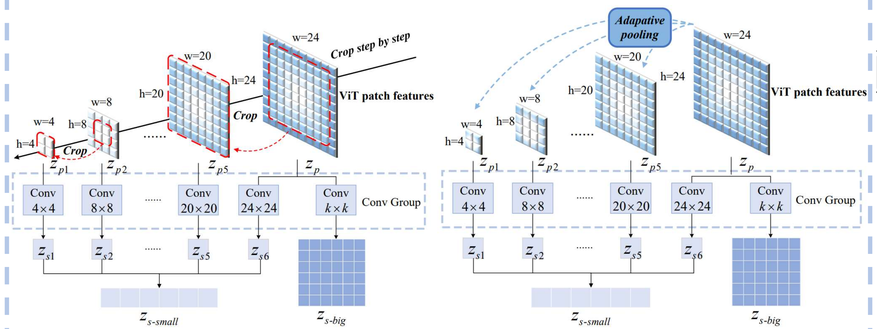
Method

LLaVA-SP

Adds **only six spatial visual tokens** to enhance visual representations.

Spatial Feature Extractor (SFE)

Uses convolutional kernels to extract visual spatial tokens from ViT patch features, capturing local spatial ordering.



(a) The SFE module in LLaVA-SP-Cropping model.

(b) The SFE module in LLaVA-SP-Pooling model.

Figure 3. **SFE Structure.** (a) illustrates the process of obtaining precise multi-scale features using the cropping operation, simulating the arrangement of visual spatial features as "from central region to global", emphasizing details in image regions. (b) demonstrates the method of obtaining abstract feature maps at multi-scale using adaptive pooling, simulating the arrangement of visual spatial features "from abstract to specific", emphasizing the global semantics of the image. We use a group of convolutional kernels to extract visual spatial features $Z_{s\text{-}small}$, and $Z_{s\text{-}big}$ is used to feature fusion in DFI.

Detail Feature Integrator (DFI)

DFI addresses SFE's trade-off by injecting fine-grained features, capturing details without increasing visual tokens.

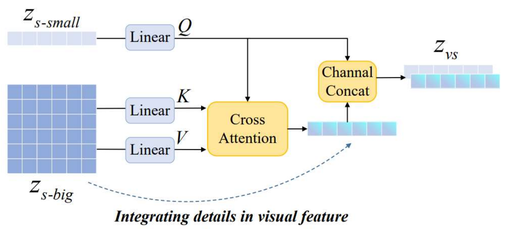


Figure 4. **DFI architecture.** Integrating $Z_{s\text{-}big}$ details and injecting them into $Z_{s\text{-}small}$.

Experiment

General visual-language understanding

Method	LLM	Res.	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	POPE	MME^{P}	MMB	SEED^{I}	LLaVA^{W}	MM-Vet
BLIP-2 [26]	Vicuna-13B	224	41.0	41.0	19.6	61.0	42.5	85.3	1293.8	–	46.4	38.1	22.4
InstructBLIP [12]	Vicuna-7B	224	–	49.2	34.5	60.5	50.1	–	–	36.0	53.4	60.9	26.2
InstructBLIP [12]	Vicuna-13B	224	–	49.5	33.4	63.1	50.7	78.9	1212.8	–	–	58.2	25.6
Shikra [5]	Vicuna-13B	224	77.4	–	–	–	–	–	–	58.8	–	–	–
Qwen-VL [2]	Qwen-7B	448	78.8	59.3	35.2	67.1	**63.8**	–	–	38.2	56.3	–	–
Qwen-VL-Chat [2]	Qwen-7B	448	78.2	57.5	38.9	68.2	61.5	–	1487.5	60.6	58.2	–	–
DeCo [58]	Vicuna-7B	336	74.0	54.1	49.7	–	56.2	85.9	1373.4	60.6	62.8	–	–
LLaVA-1.5† [32]	Vicuna-7B	336	78.5	62.0	50.0	66.8	58.2	85.9	**1510.7**	64.3	66.2	63.4	30.5
LLaVA-1.5* [32]	Vicuna-7B	336	78.4	61.9	45.7	67.6	56.2	85.8	1477.4	64.5	67.0	64.2	32.1
LLaVA-SP-Cropping	Vicuna-7B	336	79.2	62.4	50.1	**69.7**	58.7	86.4	1473.8	**65.8**	**67.6**	66.7	32.2
LLaVA-SP-Pooling	Vicuna-7B	336	79.1	**62.5**	**51.6**	69.0	58.3	**86.5**	1475.9	65.7	67.5	**68.3**	**33.4**

Visual spatial understanding

Method	MME POS	MMB SR	MMB OL	MMB PR	SEED SR	SEED IL	Avg^{N}
LLaVA-1.5†	128.3	20.0	44.4	25.0	51.1	59.9	44.1
Honeybee [3]	116.7	15.6	42.0	54.2	43.5	54.4	44.7
DeCo [58]	116.7	24.4	48.1	41.7	46.6	58.5	46.3
LLaVA-SP-Pooling	**138.3**	15.6	45.7	37.5	49.0	61.4	46.4
LLaVA-SP-Cropping	126.7	**24.4**	**50.6**	29.2	**49.8**	61.7	**46.5**

Inference speed

Method	LLM	Vision Encoder	N	Tokens / s
Qwen-VL [2]	7B	CLIP-ViT-G	256	13.01
Qwen2-VL [52]	7B	ViT-B	dynamic	12.23
LLaVA-1.5 [32]	7B	CLIP-ViT-L	576	20.76
LLaVA-SP-Cropping	7B	CLIP-ViT-L	582	20.51
LLaVA-SP-Pooling	7B	CLIP-ViT-L	582	20.28

LLaVA-SP with stonger vision encoder and novel MLLM

Method	Vision Encoder	LLM	Res.	GQA	SQA^{I}	VQA^{T}	POPE	MME^{P}	MMB	SEED^{I}	MM-Vet	Avg^{N}
LLaVA-1.5	SigLIP-L/16	Vicuna-7B	384	61.3	66.4	57.6	85.1	1450.0	65.2	67.9	32.2	63.5
LLaVA-SP-Cropping	SigLIP-L/16	Vicuna-7B	384	62.4	**68.9**	**59.9**	**85.7**	1509.2	**65.6**	68.0	31.9	64.7
LLaVA-SP-Pooling	SigLIP-L/16	Vicuna-7B	384	**62.9**	68.6	59.6	**85.7**	**1514.8**	65.5	**68.3**	**33.3**	**65.0**
InternVL-2.0 [9]	InternViT-300M	Qwen2-0.5B	448	56.8	56.7	41.2	84.6	1064.0	52.1	55.5	20.4	52.4
InternVL-2.0-SP-Cropping	InternViT-300M	Qwen2-0.5B	448	**58.4**	57.9	41.4	**85.2**	1187.5	**53.3**	**56.8**	22.7	54.4
InternVL-2.0-SP-Pooling	InternViT-300M	Qwen2-0.5B	448	**58.4**	**58.2**	**41.8**	85.0	**1223.4**	52.2	56.6	**24.0**	**54.7**