NEURAL INFORMATION PROCESSING SYSTEMS

Hierarchical Visual Feature Aggregation for OCR-Free Document Understanding

Jaeyoo Park[1], Jin Young Choi[1], Jeonghyung Park[2], Bohyung Han[1]

ComputerVisionLab
Seoul National University

SAMSUNG SDS

Motivation

- Existing document understanding models heavily rely on off-the-shelf OCR engines, that struggle with complex text styles and require costly extra processing steps

- MLLMs' emergent strong OCR ability shows their potential for document understanding tasks by reducing reliance on external OCR engines

- Current OCR-free models based on MLLMs still struggle to capture diverse visual scales and font sizes, often missing local details in documents, since they are limited to single scale visual inputs

Contribution

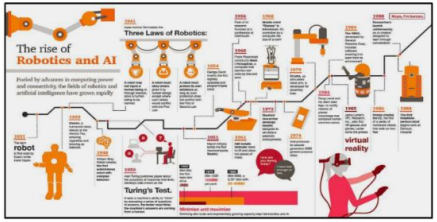
- Present a novel framework for **OCR-free document understanding**, built on a pretrained MLLMs, which integrates **multi-scale visual features** to handle varying font sizes in document images.

- Introduce the **Hierarchical Visual Feature Aggregation (HVFA)** module, which employs cross-attentive pooling to **effectively balance information preservation and computational efficiency**, addressing the escalating costs associated with detailed visual inputs.

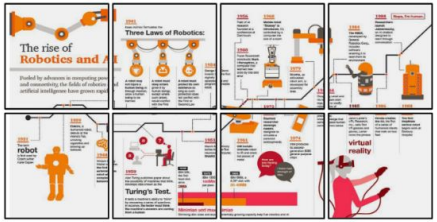
- Employ a novel instruction tuning task, which aims to **predict the relative positions of input text**, to enhance the model's comprehensive text reading capability.

Overall Framework

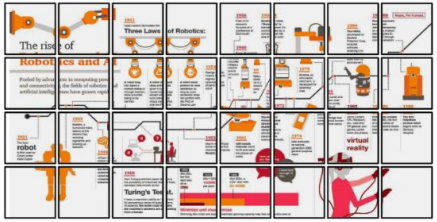
Global View



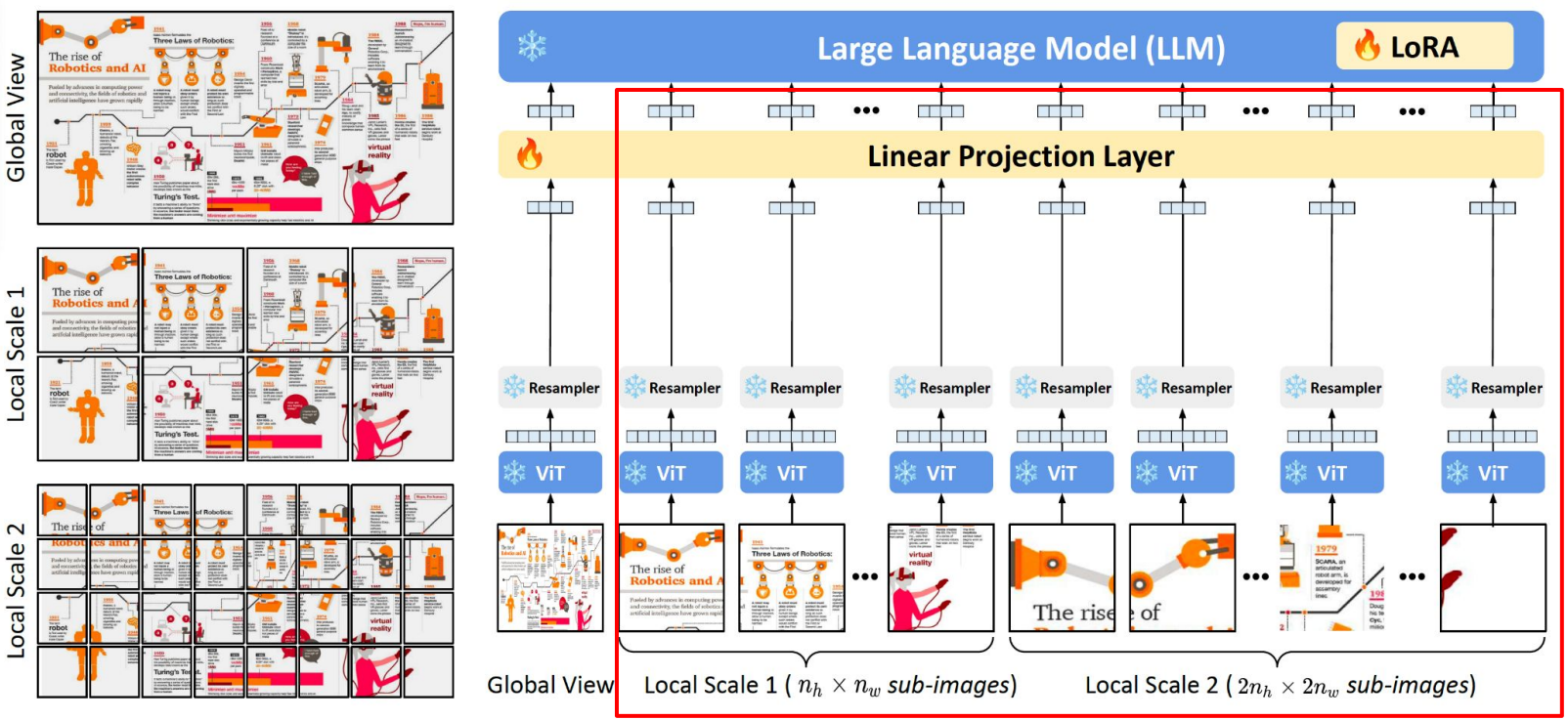
Local Scale 1



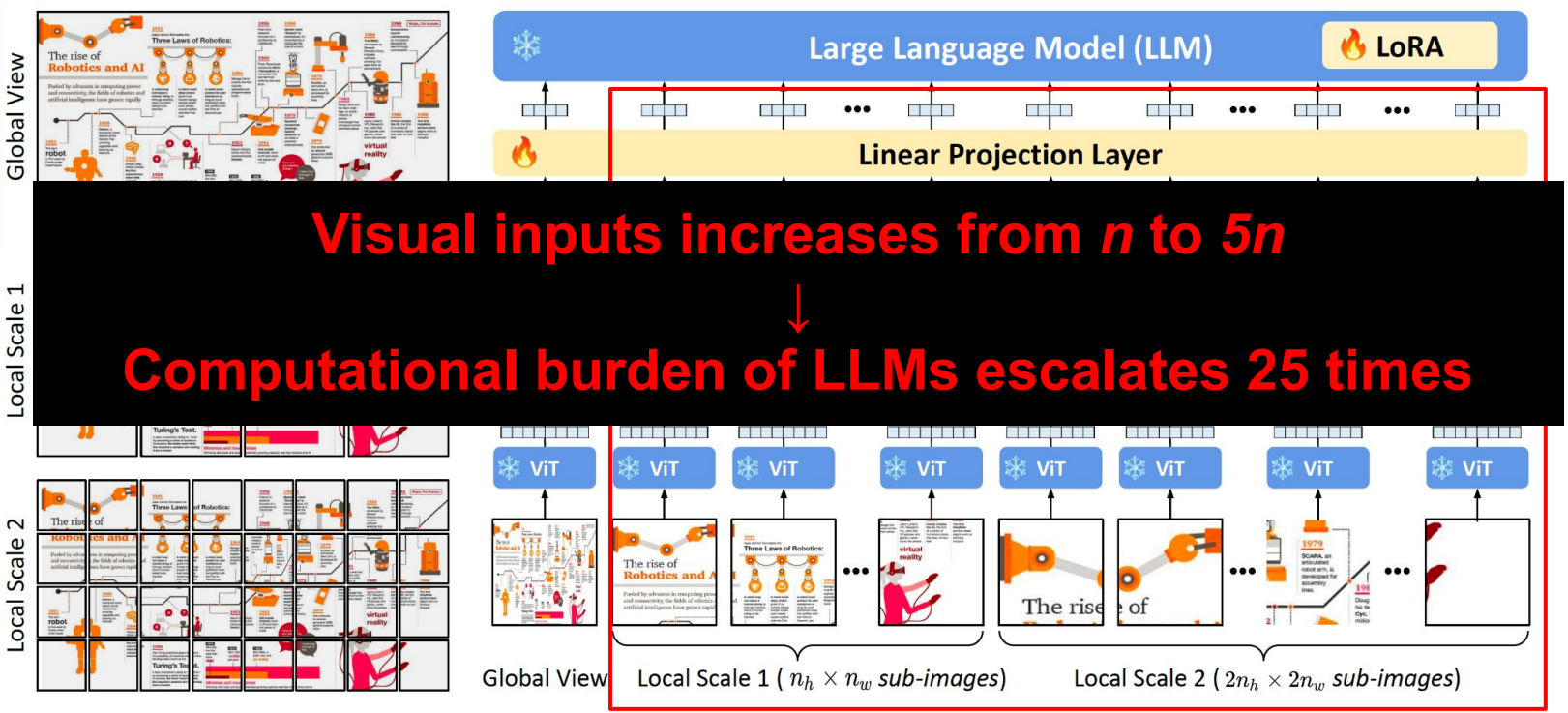
Local Scale 2



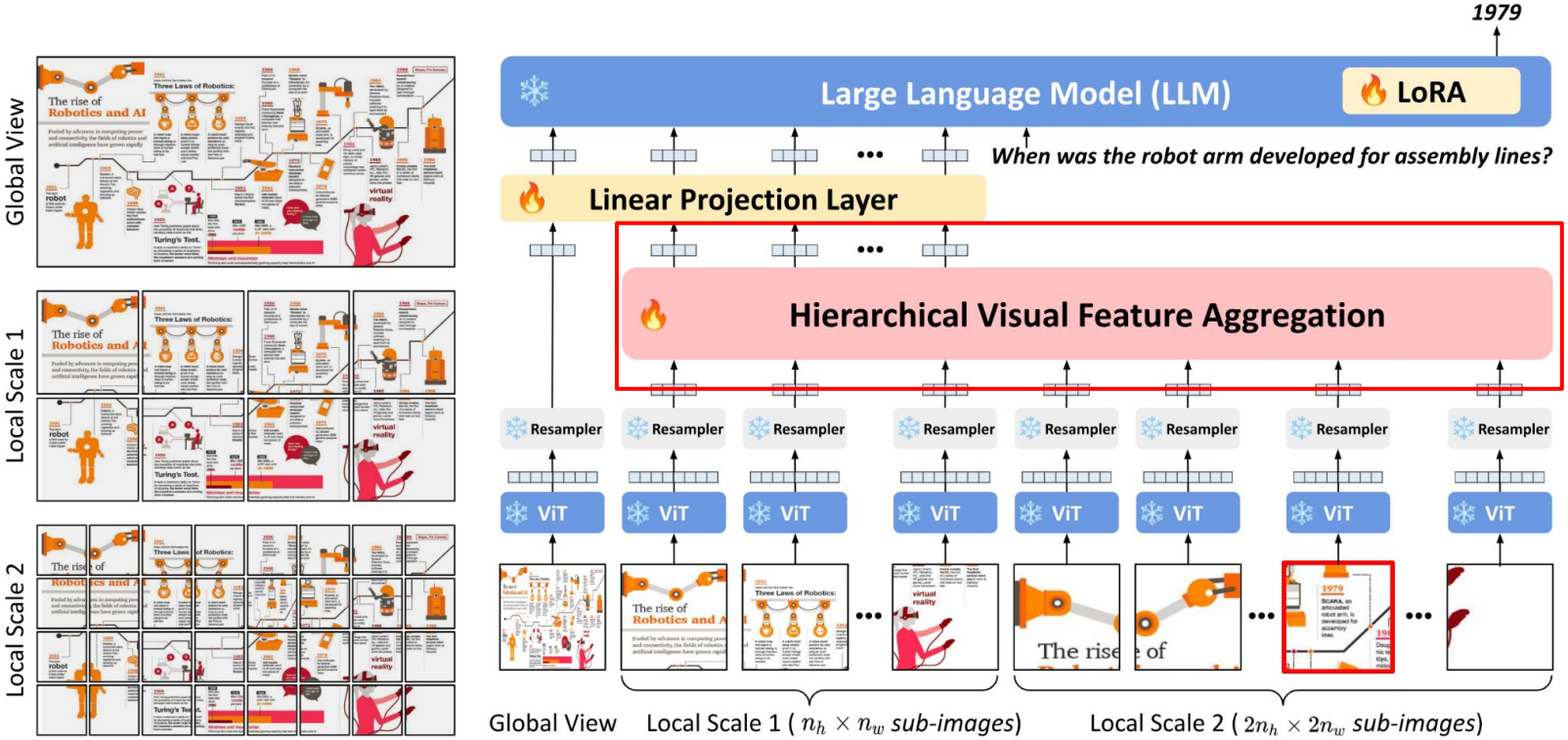
Overall Framework



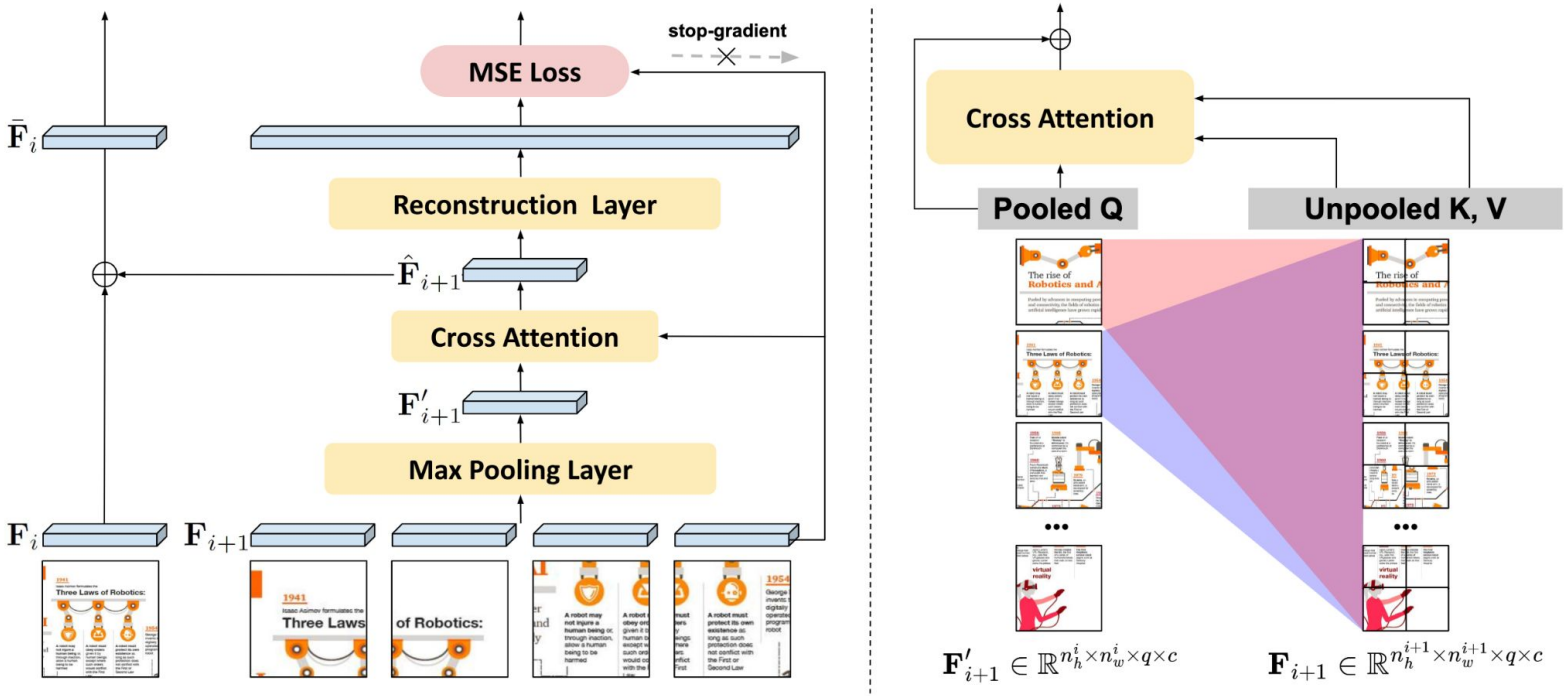
Overall Framework



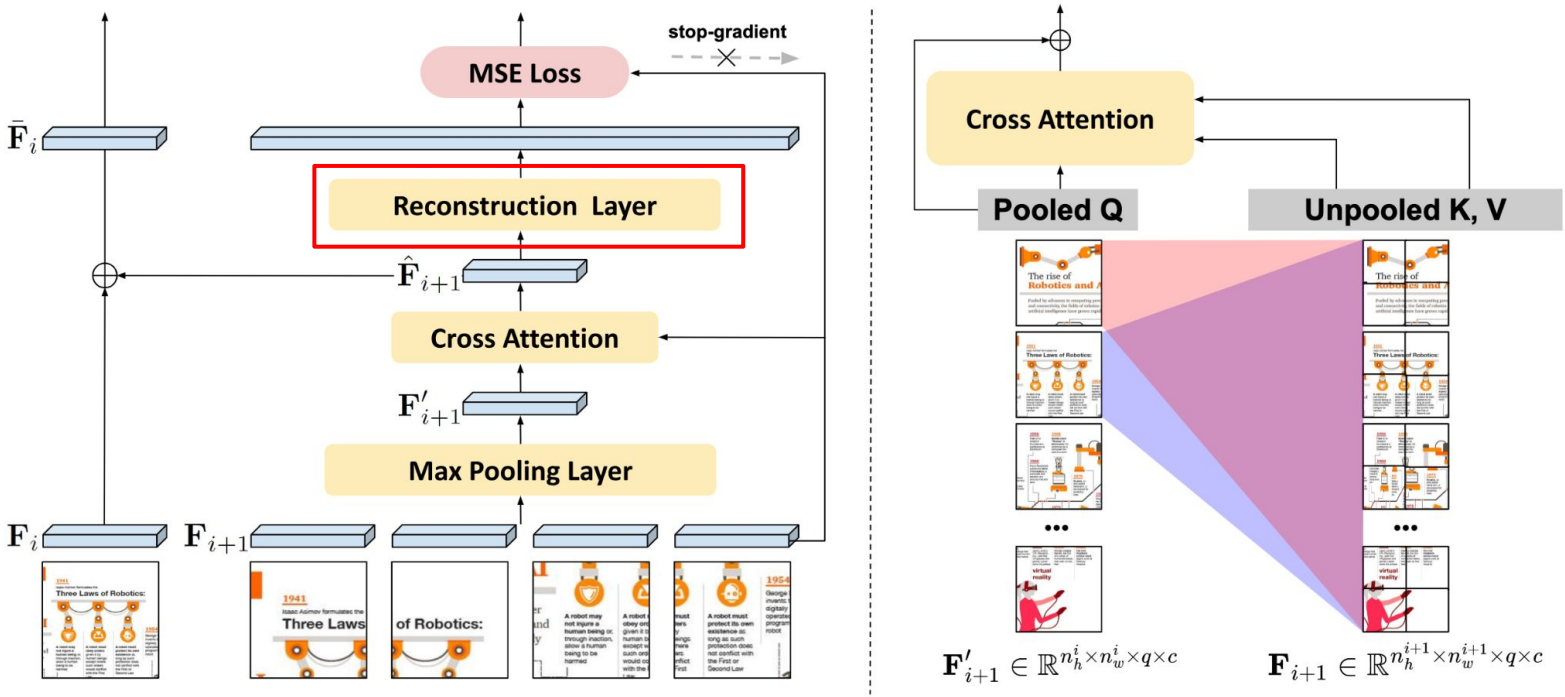
Overall Framework



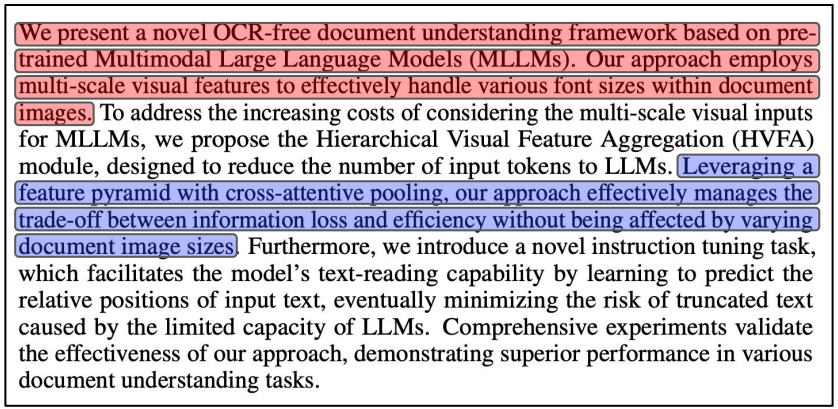
Hierarchical Visual Feature Aggregation



Hierarchical Visual Feature Aggregation



Relative Text-Position Prediction Task



Reading Partial Text (RPT)

What's in the **first 20%** of the image text?

Predicting Text Position (PTP)

Specify the relative position within the image where **"Leveraging a feature pyramid … image sizes"** is found.

MLLM

We present a novel … font sizes within document images.

42%-61%

Experiments

- Equipped with two MLLM backbones
  - BLIP-2-OPT-2.7B
  - mPLUG-Owl-7B
- 10 document understanding benchmarks

Method	DocVQA	InfoVQA	DeepForm	KLC	WTQ	TabFact	ChartQA	VisualMRC	TextVQA	TextCaps
Document-specific Pretraining										
Dessurt [16]	63.2	–	–	–	–	–	–	–	–	–
Donut [17]	67.5	11.6	61.6	30.0	18.8	54.6	41.8	93.91	43.5	74.4
Pix2Struct$_{Base}$ [18]	72.1	38.2	–	–	–	–	56.0	–	–	88.0
Pix2Struct$_{Large}$ [18]	76.6	40.0	–	–	–	–	58.6	–	–	95.5
MLLM-based Instruction Tuning										
Qwen-VL [28]	65.1	35.4	4.1	15.9	21.6	–	**65.7**	–	63.8	–
Monkey [32]	66.5	36.1	40.6	32.8	25.3	–	65.1	–	**67.6**	–
BLIP-2-OPT-2.7B + UReader [19]	38.7	22.9	5.6	18.3	17.4	58.5	37.1	214.3	43.2	126.3
BLIP-2-OPT-2.7B + Ours	51.4	29.6	14.6	23.8	21.2	59.9	50.4	**228.7**	57.3	**135.2**
mPLUG-Owl-7B + UReader [19]	65.4	42.2	49.5	32.8	29.4	67.6	59.3	221.7	57.6	118.4
mPLUG-Owl-7B + Ours	**72.7**	**45.9**	**53.0**	**36.7**	**34.5**	**68.2**	63.3	226.4	59.2	123.1

Experiments - Ablation Study

- Effectiveness of each component

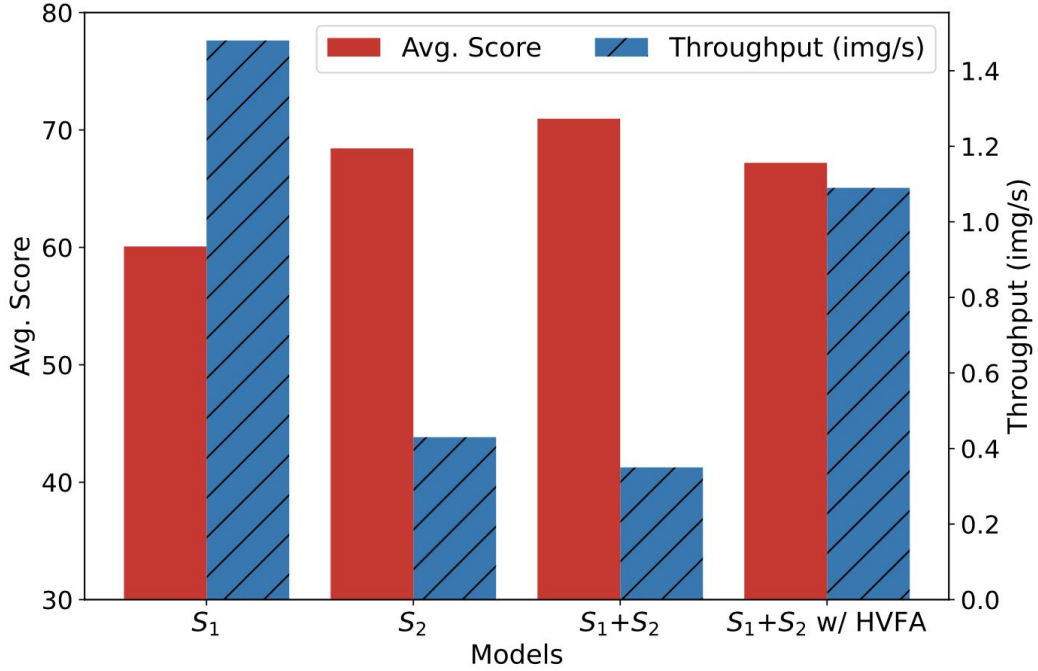
MS + HVFA	Recon	RTPP	DocVQA	InfoVQA	DeepForm	KLC	WTQ	TabFact	ChartQA	VisualMRC	TextVQA	TextCaps
			35.17	20.34	4.14	18.50	15.36	53.37	31.76	201.39	43.20	126.34
√			45.22	25.87	8.23	19.82	18.14	58.28	44.16	221.38	48.56	129.98
		√	40.18	25.19	6.49	18.97	17.80	56.85	38.88	219.74	47.54	128.39
√	√		47.36	27.37	10.92	19.17	18.24	58.14	46.70	222.56	52.65	130.86
√		√	50.46	28.48	13.18	**23.82**	20.06	59.83	49.48	226.65	56.04	134.06
√	√	√	**51.38**	**29.60**	**14.58**	23.78	**21.15**	**59.87**	**50.41**	**228.65**	**57.32**	**135.24**

- Ablation study on text reading tasks

RFT [19]	RPT	PTP	DocVQA	InfoVQA	DeepForm	KLC	WTQ	TabFact	ChartQA	VisualMRC	TextVQA	TextCaps
			47.36	27.37	10.92	19.17	18.24	56.14	43.88	222.56	52.65	130.86
√			49.23	28.16	12.58	21.29	19.42	58.46	46.05	224.39	55.78	133.55
√		√	50.62	29.49	12.55	22.88	20.51	59.28	48.41	226.55	57.19	134.44
	√		50.46	28.44	**14.88**	22.55	20.88	59.12	48.28	226.56	56.91	134.92
	√	√	**51.38**	**29.60**	14.58	**23.78**	**21.15**	**59.87**	**50.41**	**228.65**	**57.32**	**135.14**

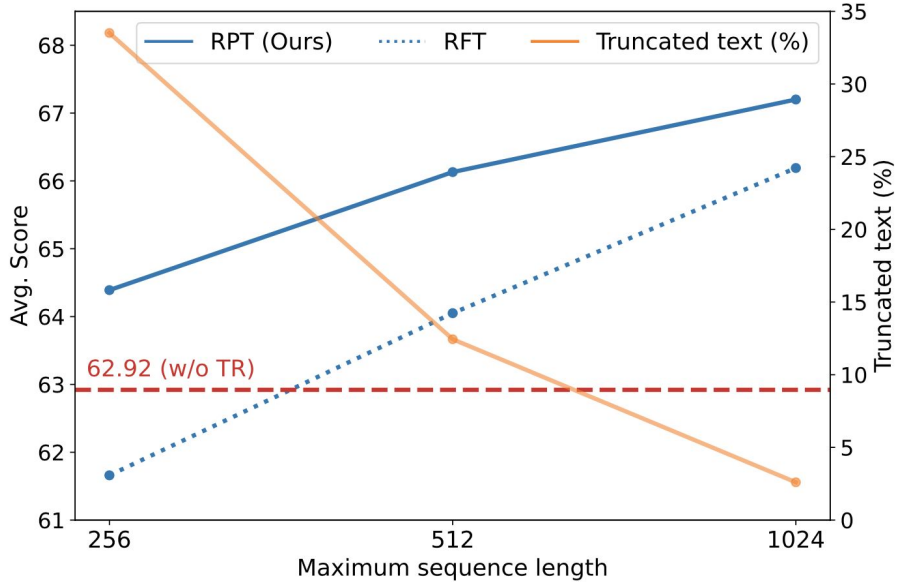
Complexity Analysis on Visual Input Scale

- HVFA efficiently integrates finer-scale features into a coarser scale, achieving a balance between performance gains and computational demands



Analysis on Text Reading Task

- RPT ensures reliable data quality while incorporating stochasticity and positional information, contributing to its effectiveness.
- The design of RPT is helpful for mitigating truncation issues while RFT suffers from truncation of text reading data.



Thank You!!

Please Visit Our Poster!

Primary Contact : bellos1203@snu.ac.kr