CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

模式识别国家重点实验室
National Lab of Pattern Recognition

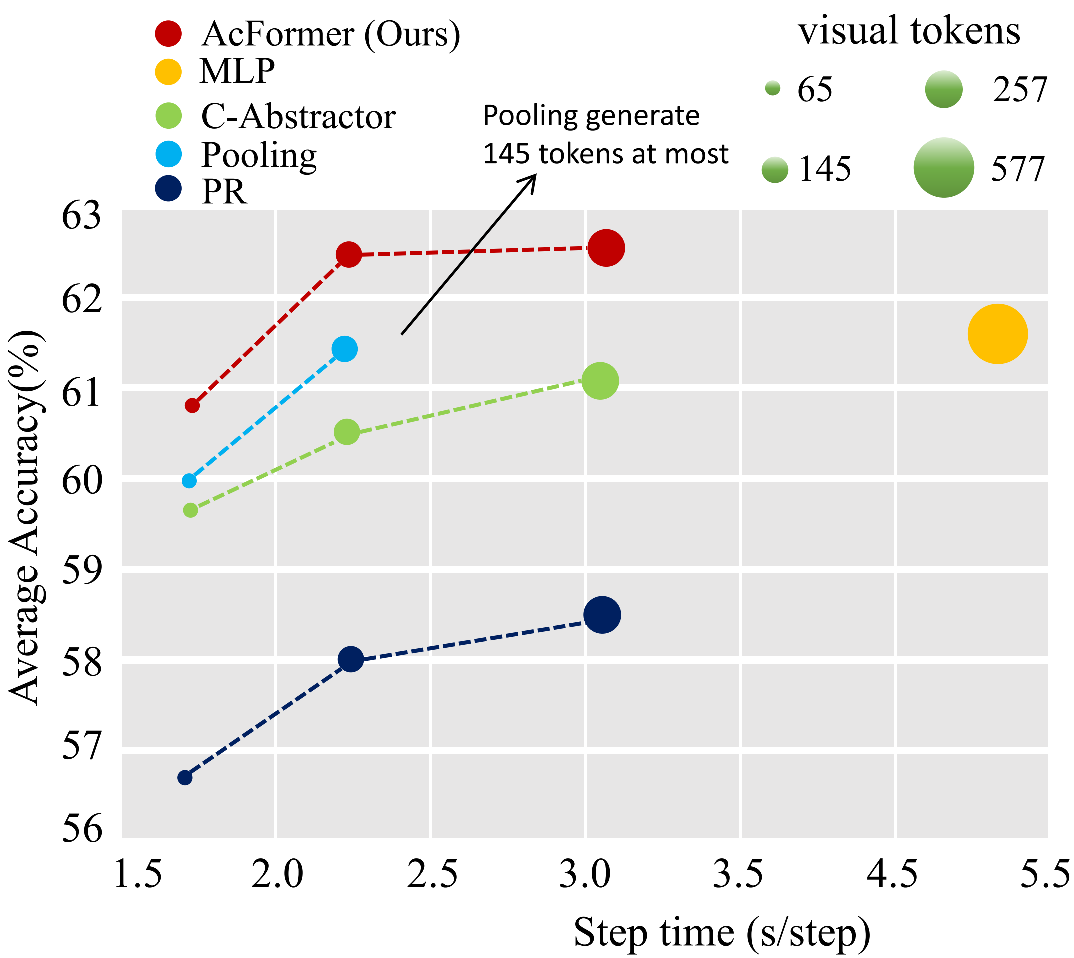
中国科学院自动化研究所
Institute of Automation Chinese Academy of Sciences

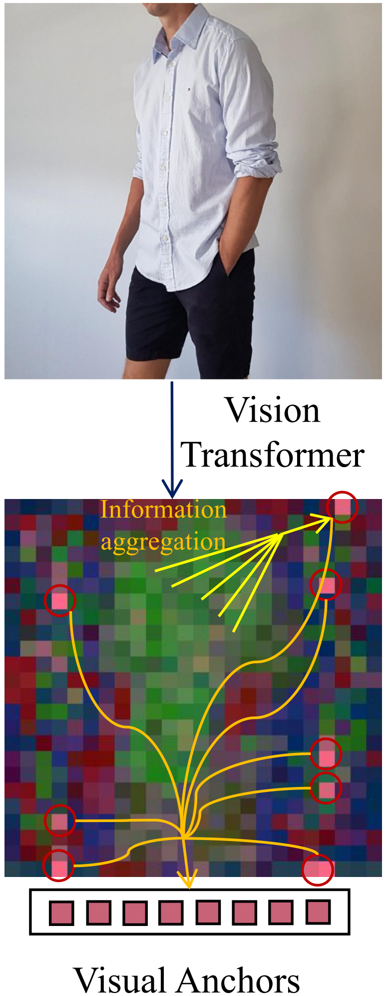
Visual Anchors Are Strong Information Aggregators For Multimodal Large Language Model

Haogeng Liu

Motivation

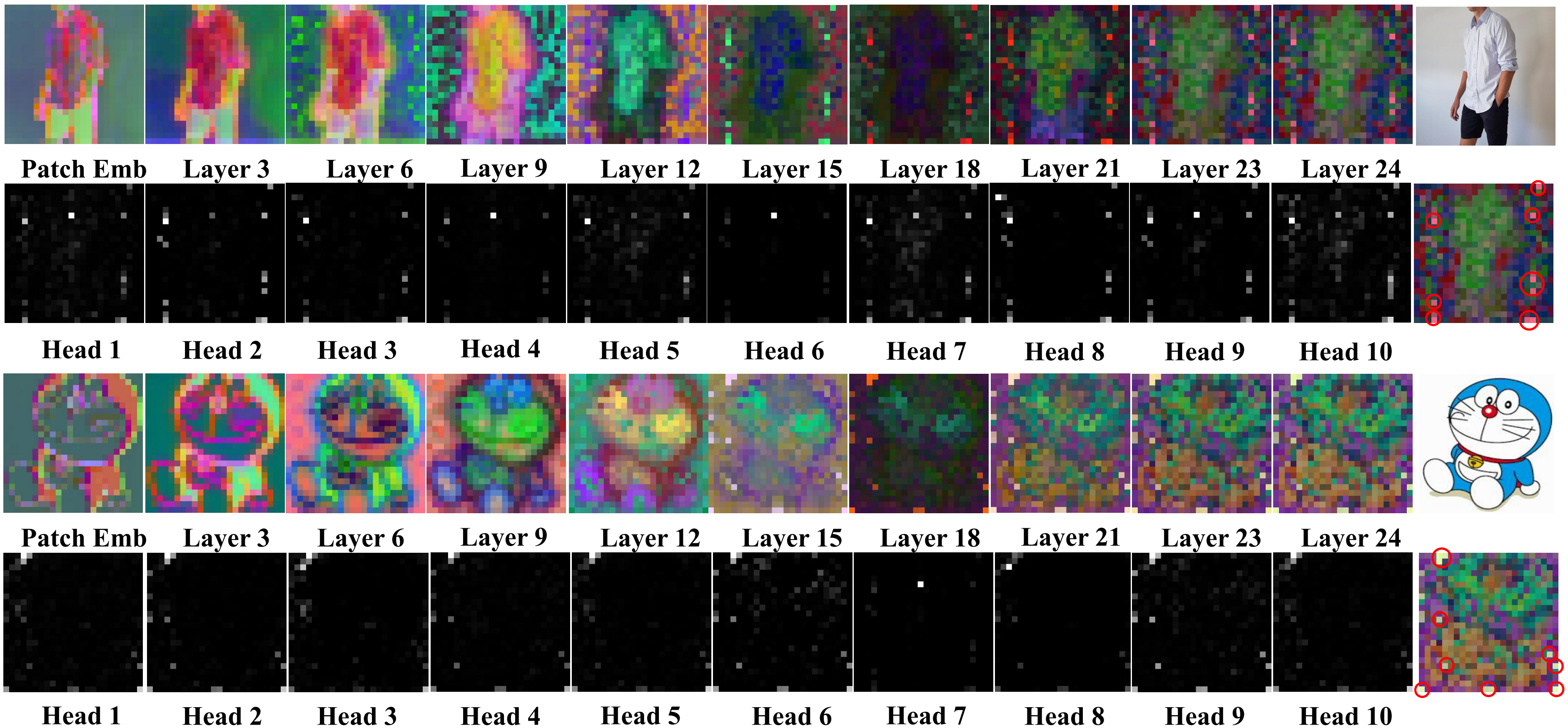
Using visual representations from a pretrained encoder, we analyzed features with PCA, revealing "Visual Anchor" phenomenon. From this, we developed Anchor Former as an optimized visual-language connector, reducing inference latency and enhancing model generalization for high-performance Vision-Language tasks.





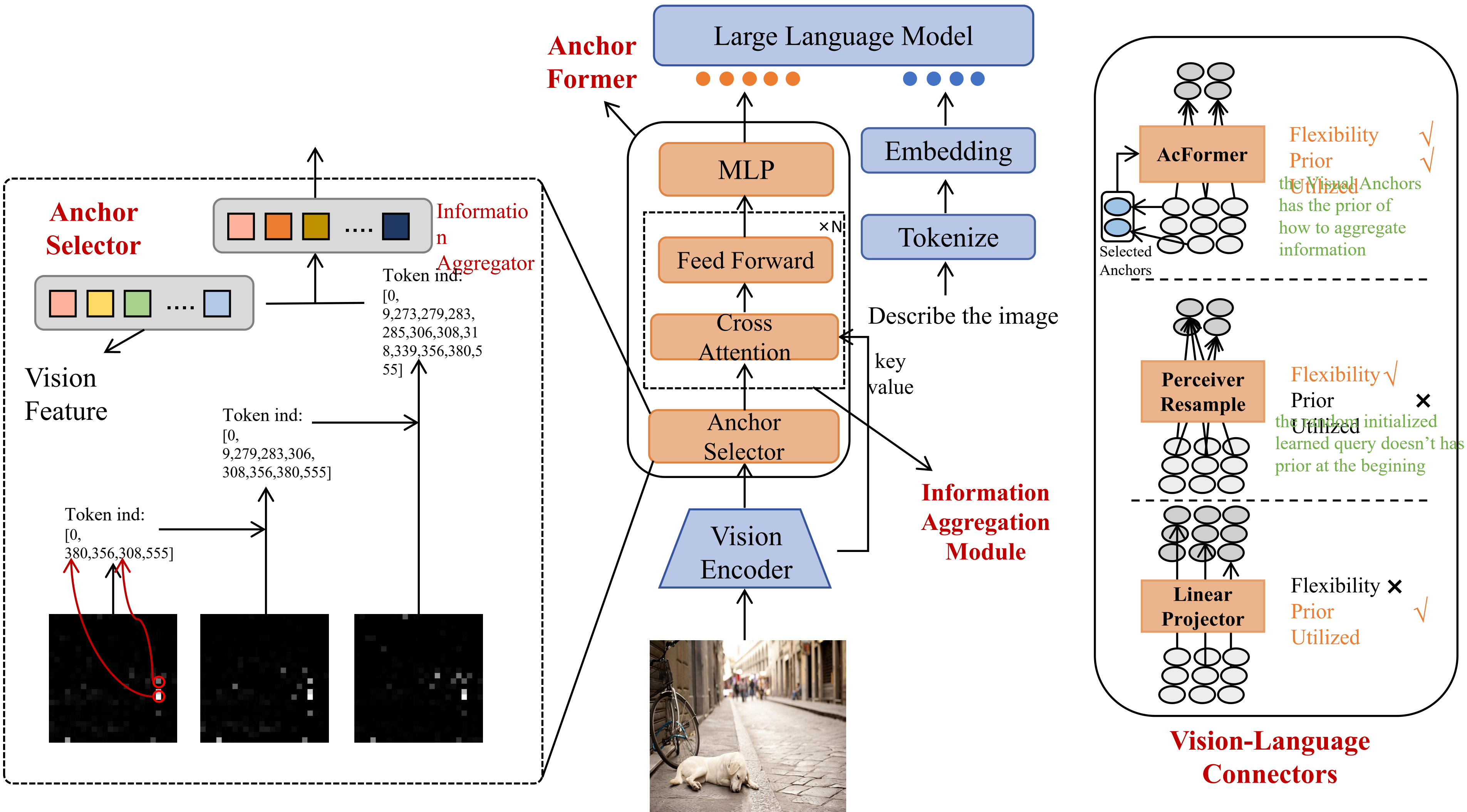
Method

Visual anchoring refers to specific background tokens that, as visual signals transform within the ViT, become high-norm tensors. These tokens often receive greater attention from the classification head and the CLS token, highlighting their significance in visual perception tasks.



Method

Based on these observations, we propose that visual anchors function as local information extractors during visual feature transformations, transmitting localized signals to the global representation through multiple visual anchors. Accordingly, we introduce the Anchor Former architecture, as illustrated below.



Method

We propose to use the top-k method to select the visual anchors with the CLS token's attention map.

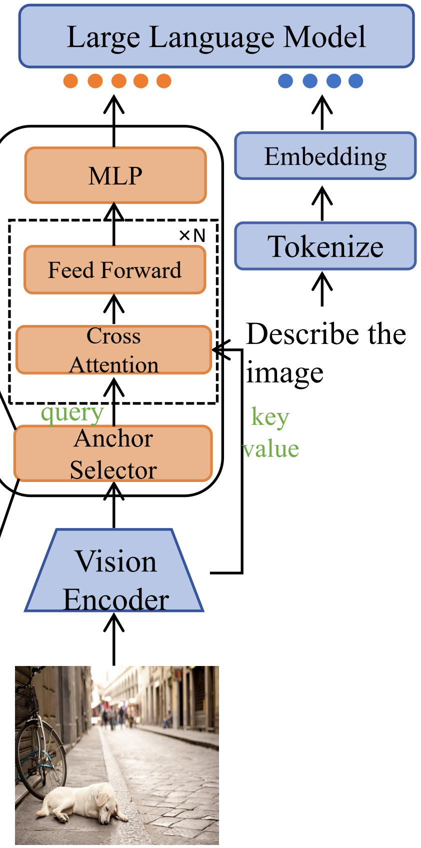
Algorithm: Anchor Selection

Input:
Visual Feature Map **V** (B, N, D)
Visual Attention Map **A** (B, H, N, N)
Token Number T

Caculation:

```
selected_anchor = None
Per_head_num = int((T-1) / H) #[CLS] is choosen by default
for i in range(B):
  max_indice = [0]
  for j in range(H):
    tmp_attn = A[i, j, 0, 1:]
    ind_sorted = Argsort(tmp_attn) + 1
    tmp_res = set(ind_sorted[-Per_head_num:] + max_indice)
    count = 1
    while len(tmp_res) < ((j+1) × per_head_num + 1) :
      tmp_res = set(ind_sorted[-Per_head_num-count:] + max_indice)
      count = count + 1
    max_indices = sorted(list(tmp_res))
    if selected_anchor is not None:
      selected_anchor = torch.cat((selected_anchor, V[[i], max_indices, :]), dim=0)
    else:
      selected_anchor = V[[i], max_indices, :]
```

Return:
selected_anchor



Experiment

We evaluated our method across various benchmarks, with results demonstrating exceptional performance in both computational efficiency and accuracy.

Table 1: Results on benchmark designed for MLLMs. V-T Num means the visual tokens number. V-T Num influences the computation cost that the bigger the V-T Num the heavier the computation cost is. Speed here means the relative pre-training speed with respect to LLaVA-1.5.

Model	LLM	Connector	V-T Num	Res	POPE	MME	MMB	MM-Vet	Speed (↑)
Approaches using 7B Large Language Models									
MiniGPT-4 [53]	Vicuna-7B	Resampler	32	224	72.2	726.0	24.3	22.1	-
mPLUG-Owl2[46]	LLaMA2-7B	Resampler	32	224	-	1243.4	49.4	-	-
InstructBLIP[11]	LLaMA2-7B	Q-Former	32	224	78.9	-	36.0	26.2	-
LLaVA (v1) [34]	LLaMA-7B	Linear	257	224	67.7	717.5	38.7	-	-
LLaMA-AdapterV2 [16]	LLaMA2-7B	LLaMA-Adapter	257	224	-	1221.6	41.0	31.4	-
Shikra [6]	Vicuna-7B	Linear	257	224	-	-	58.8	-	-
Qwen-VL[4]	Qwen-7B	Resampler	256	448	-	-	38.2	-	-
Qwen-VL-Chat[4]	Qwen-7B	Resampler	256	448	-	1845.3	60.6	-	-
LLaVA-1.5 [33]	Vicuna-7B	Linear	577	336	85.9	1794.6	64.3	**30.5**	1.00×
Ours	Vicuna-7B	AcFormer	145	336	**86.4**	**1846.1**	**68.4**	30.3	**2.23**×
Approaches using 13B Large Language Models									
MiniGPT-4 [53]	Vicuna-13B	Resampler	32	224	-	1158.7	-	24.4	-
InstructBLIP[11]	Vicuna-13B	Q-Former	32	224	78.9	1504.6	-	25.6	-
BLIP-2[28]	Vicuna-13B	Q-Former	32	224	85.3	-	-	22.4	-
LLaVA-1.5 [33]	Vicuna-13B	Linear	577	336	85.9	1826.7	67.7	**35.4**	1.00×
Ours	Vicuna-13B	AcFormer	145	336	**86.1**	**1870.0**	**69.2**	34.1	**2.30**×

Experiment

Table 2: Results on General VQA tasks. V-T Num means the visual tokens number. V-T Num influences the computation cost that the bigger the V-T Num the heavier the computation cost is. Speed here means the relative pre-training speed with respect to LLaVA-1.5.

Model	LLM	Connector	V-T Num	Res	TextVQA	GQA	VQAv2	VisWiz	SQA_{img}	Speed (↑)
Approaches using 7B Large Language Models										
InstructBLIP[11]	LLaMA2-7B	Q-Former	32	224	-	49.2	-	34.5	60.5	-
Shikra [6]	Vicuna-7B	Linear	257	224	-	-	77.4	-	-	-
IDEFICS-9B [25]	LLaMA-7B	Cross Attn	257	224	-	38.4	50.9	35.5	-	-
Qwen-VL[4]	Qwen-7B	Resampler	256	448	-	59.3	**78.8**	35.2	67.1	-
Qwen-VL-Chat[4]	Qwen-7B	Resampler	256	448	-	57.5	78.2	38.9	68.2	-
LLaVA-1.5 [33]	Vicuna-7B	Linear	577	336	58.2	**62.0**	78.5	50.0	66.8	1.00×
Ours	Vicuna-7B	AcFormer	257	336	**58.2**	61.2	78.4	**52.8**	**69.4**	**1.65**×
Approaches using 13B Large Language Models										
InstructBLIP[11]	Vicuna-13B	Q-Former	32	224	-	49.5	-	33.4	63.1	-
BLIP-2[28]	Vicuna-13B	Q-Former	32	224	-	41.0	41.0	19.5	61.0	-
LLaVA-1.5 [33]	Vicuna-13B	Linear	577	336	61.2	**63.3**	**80.0**	53.6	71.6	1.00×
Ours	Vicuna-13B	AcFormer	257	336	**61.3**	63.0	79.8	**53.7**	**71.8**	**1.69**×

Table 4: Ablation studies on whether to directly use the selected tokens as input.

Model	LLM	Connector	V-T Num.	TextVQA	GQA	MMB	MME
LLaVA-1.5	Vicuna-7B	Top-P	145	56.3	60.8	68.2	1798.8
	Vicuna-7B	E-ViT	146	57.1	61.0	68.3	1808.4
	Vicuna-7B	AcFormer	145	**58.0**	**61.3**	**68.4**	**1846.1**

Experiment

We also provide ablation studies with other token reduction method.

Table 3: Ablation studies. "Pooling" denotes direct pooling of visual token. "Pooling-PR" employs the pooled tokens as queries for the Perceiver Resampler. "Random-PR" means the Perceiver Resampler using randomly selected tokens from the vision feature map as query. "PR" refers to the Perceiver Resampler using learnable queries. "AcFormer" represents our proposed Anchor Former. The configuration of the C-Abstractor follows Honeybee [5]. V-T Num means the visual tokens number.

Model	LLM	Connector	V-T Num.	TextVQA	GQA	MMB	MME
LLaVA-1.5	Vicuna-7B	Pooling	65	53.4	59.8	66.8	1734.0
	Vicuna-7B	Pooling-PR	65	53.9	**60.0**	66.8	1728.9
	Vicuna-7B	Random-PR	65	53.9	59.1	66.9	1728.7
	Vicuna-7B	PR	65	51.0	56.1	63.2	1702.8
	Vicuna-7B	C-Abstractor	65	52.8	59.0	67.0	1743.3
	Vicuna-7B	AcFormer	65	**56.1**	59.2	**67.3**	**1744.2**
LLaVA-1.5	Vicuna-7B	Pooling	145	55.1	60.9	68.0	1791.4
	Vicuna-7B	Pooling-PR	145	54.7	60.9	68.0	1759.1
	Vicuna-7B	Random-PR	145	54.6	59.7	67.0	1772.7
	Vicuna-7B	PR	145	52.1	56.4	65.4	1720.8
	Vicuna-7B	C-Abstractor	145	53.4	60.2	67.8	1775.4
	Vicuna-7B	AcFormer	145	**58.0**	**61.3**	**68.4**	**1846.1**
LLaVA-1.5	Vicuna-7B	PR	257	52.3	56.8	65.7	1735.9
	Vicuna-7B	C-Abstractor	257	53.7	60.8	68.3	1790.0
	Vicuna-7B	AcFormer	257	**58.2**	**61.2**	**68.3**	**1848.8**
LLaVA-1.5	Vicuna-13B	PR	145	53.4	56.9	64.7	1749.3
	Vicuna-13B	C-Abstractor	145	58.5	62.1	68.8	1823.6
	Vicuna-13B	AcFormer	145	**60.7**	**62.8**	**69.2**	**1869.3**

Thanks