Dense Connector for MLLMs

Huanjin Yao[1,3*], **Wenhao Wu**[2*✉], **Taojiannan Yang**[4], **Yuxin Song**[3], **Mengxi Zhang**[3]
Haocheng Feng[3], **Yifan Sun**[3], **Zhiheng Li**[1], **Wanli Ouyang**[5], **Jingdong Wang**[3]
[1]Shenzhen International Graduate School, Tsinghua University [2]The University of Sydney
[3]Baidu Inc. [4]Amazon [5] The Chinese University of Hong Kong
[*] Equal Contribution ✉ Corresponding Author

郝飞洋

2025年1月5日

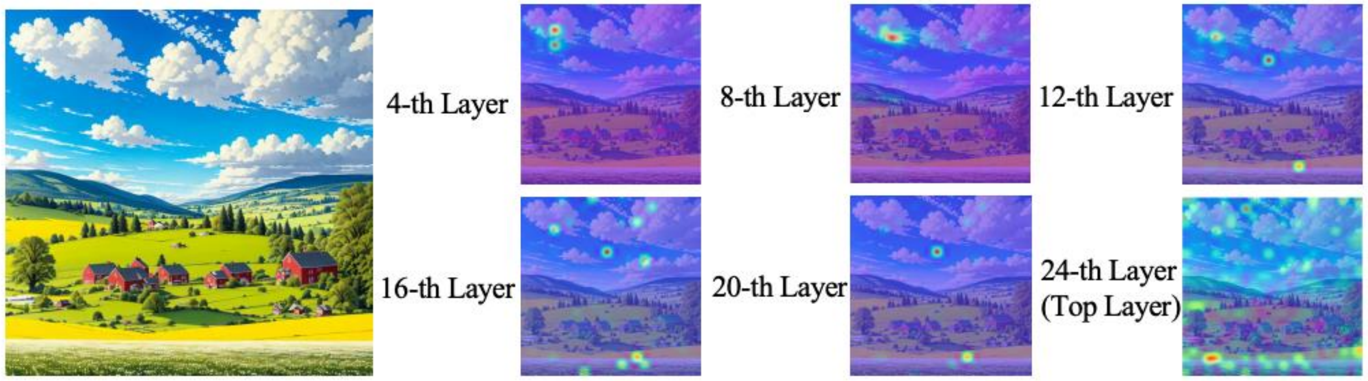
概述

- 多模态大语言模型的visual encoder潜力有待挖掘
- 使用visual encoder的多个层的信息
- 为了减轻计算压力，选择几层的信息进行融合/分组求平均值
- 效果很好:
  - 相同的token数性能更优
  - 性能相似token数更少

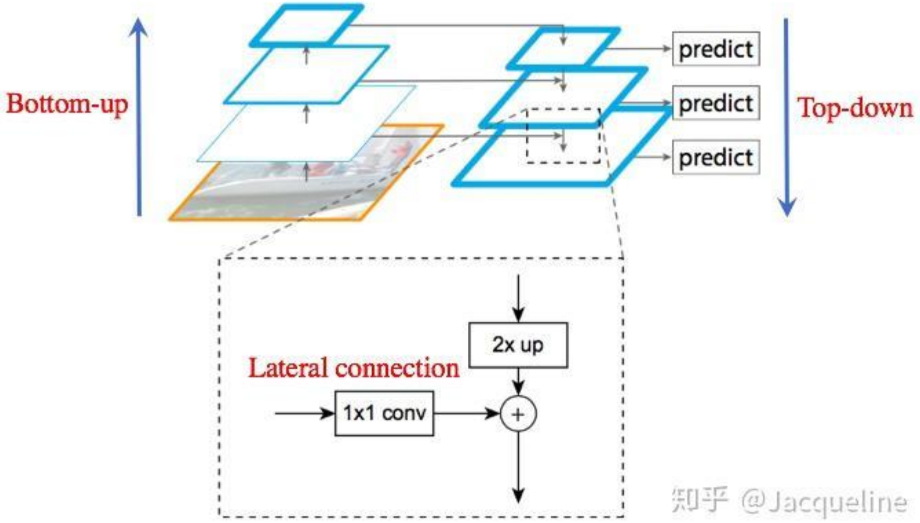
相关工作（其实顺便说了可扩展性）

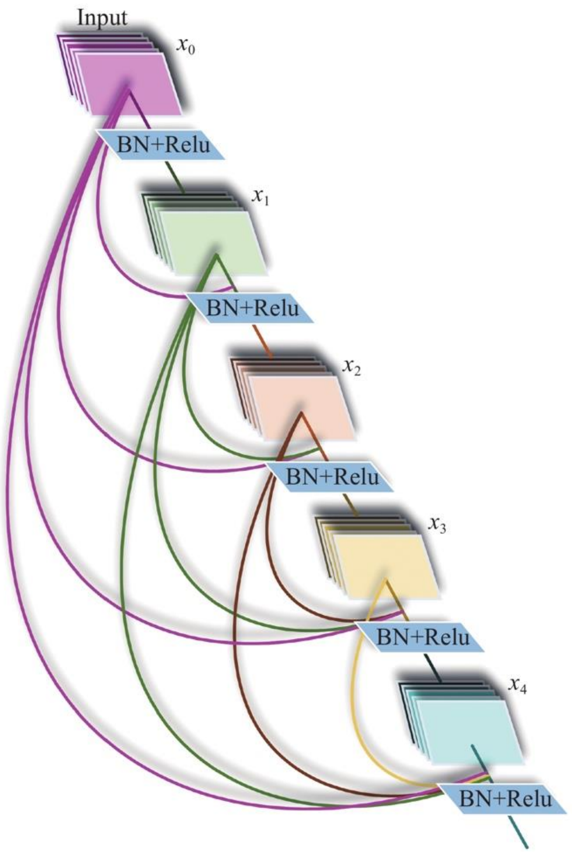
- 预训练的视觉模型：CLIP, SigLIP
- 大语言模型：2.7B到70B
- 多模态大语言模型：Q-former, linear projection, MLP, 视频……都只用了最后一层的表征。
- 这篇文章根据FreeVA的方法把图片模型不训练直接迁移到视频模型上去

预实验与历史经验



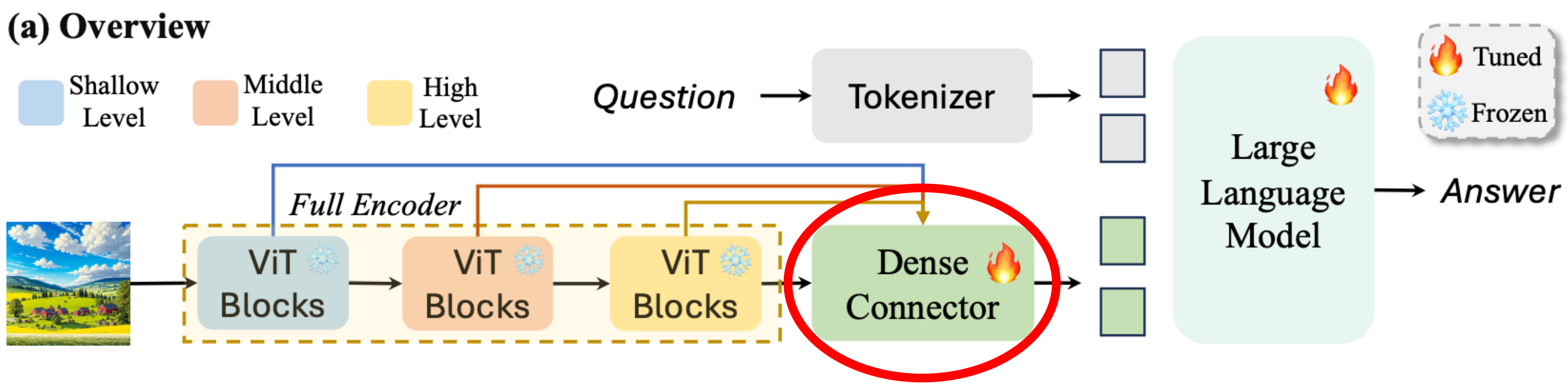
(a) Visualizing Attention Maps across ViT-L Layers.





- ViT各层的不同attention
- Densenet和FPN对各层的feature进行融合

模型设计


(a) Overview
Shallow Level
Middle Level
High Level
Question
Tokenizer
Tuned
Frozen
Large Language Model
Answer
Full Encoder
ViT Blocks
ViT Blocks
ViT Blocks
Dense Connector

模型设计

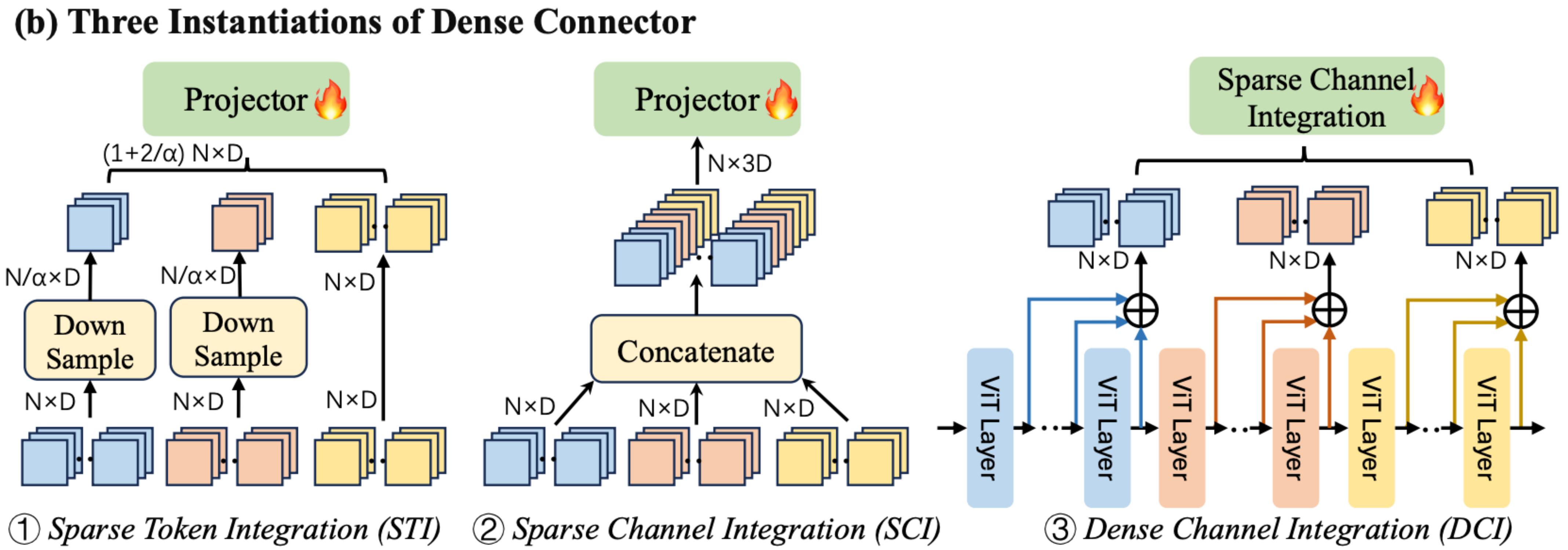


Figure 2: Dense Connector in MLLM: Overview and Three Instantiations. N is the number of tokens, D is the feature dimension, and α is the downsampling ratio.

Efficient Dense Connector for Visual Token Optimization

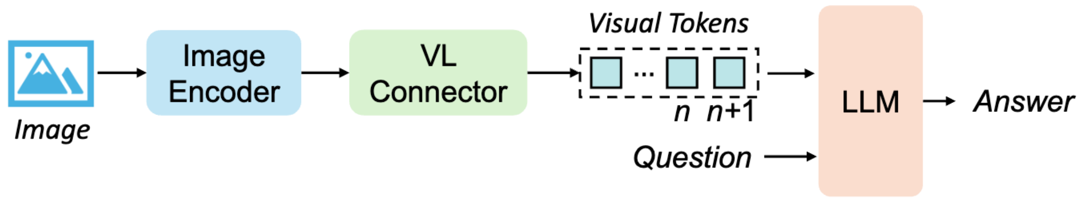
- 当通过上面的方法获得e_v后，可以使用一个二维差值函数来下采样这些visual token。

- 上面是说同样的token长度达到了更好的性能，下面又说相同的性能用了更少的token，就多了一个下采样步骤。

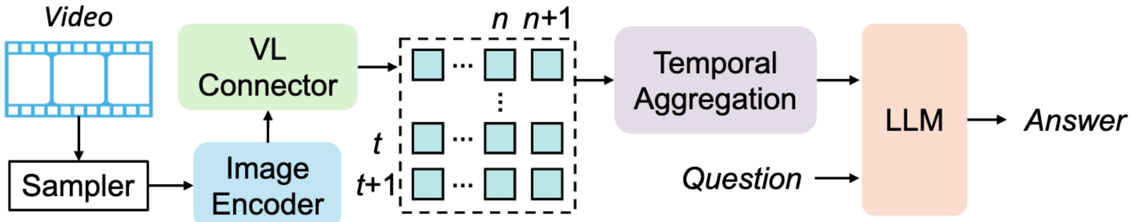
- 推理速度提升3倍

Training-Free Extension from Image to Video Conversational Models

- 借鉴了FreeVA的方法（也是这个作者的论文）
- 1. 均匀采样T帧，每一帧过一遍视觉编码器，得到 $\{e_{v1},\ldots,e_{vT}\}$
- 2. 然后把上面的序列过FreeVA的pipeline喂给LLM即可



(a) Inference workflow of image MLLMs (*e.g.*, BLIP2 [10], LLaVA [3]). An input image is first processed by the *Image Encoder* (*e.g.*, ViT-L [6]) to extract visual features, which are then converted into language embeddings by the *Vision-Language (VL) Connector* (*e.g.*, Q-former [10], projection [3]). Finally, the *LLM* (*e.g.*, Vicuna [9]) interprets these visual tokens to answer questions. Here, n represents the index of patch tokens.



(b) FreeVA: A training-free pipeline for video question answering using existing image MLLMs. Here, t indicates the index of the sampled frames. *Too simple? That's enough!*

模型细节

- Visual encoders: CLIP-ViT-L-336px 和 SigLIP-ViT-SO
- LLMs: 2.7B到70B的一系列模型Phi-2-2.7B, Vicuna-7B&13B, Hermes-2-Yi-34B, Llama3-8B&70B-Instruct
- Dense connector: 24-layer CLIP-ViT-L-336px - 8, 16, 最后一层
- STI: alpha = 8
- DCI: 两组
- 数据集： LLaVA-1.5 pre-training dataset, Mini-Gemini

训练细节

- 预训练
  - Visual encoder和LLM用原有参数（冻结）
  - Dense connector随机初始化（训练）
  - 1 epoch
  - Batch size=256; lr=1e-3
- 指令微调
  - 仍然冻结visual encoder
  - 改变LLM和Dense connector的参数
  - Batch size=128; lr=2e-5
  - 参数小的LLM全量微调，参数大的用lora(rank=128, alpha=256)

消融实验-SCI, STI, DCI

Table 1: Ablations on Visual Layer Selection in Dense Connector. Here, we explore three instantiations (*STI*, *SCI*, and *DCI*) of our Dense Connector integrated with the baseline (*i.e.*, LLaVA-1.5 [16]), which utilizes a 24-layer CLIP-ViT-L-336px.

Model	Layer Index	**GQA**	**VQA**^{v2}	**SQA**^{I}	**VQA**^{T}	**POPE**	**MMB**	**MMV**	**LBW**
Baseline	24	62.0	78.5	66.8	58.2	85.9	64.3	31.1	65.4
+ *STI*	8,16,24	63.3	79.1	68.0	58.0	85.8	67.2	30.9	65.5
+ *STI*	8,16,20,24	63.0	79.1	68.0	58.8	85.9	67.6	30.8	65.7
+ *SCI*	8,16,24	63.7	79.2	68.9	58.2	86.1	66.2	32.2	66.0
+ *SCI*	16,24	63.0	79.0	67.6	58.2	86.0	65.6	31.7	65.6
+ *SCI*	8,16,20,24	63.6	79.2	67.0	58.1	86.0	65.8	31.9	66.0
+ *DCI*	(1-8),(9-16),(17-24)	63.6	79.3	67.8	58.6	86.3	66.5	32.6	66.0
+ *DCI*	(1-12),(13-24)	**63.8**$^{1.8\uparrow}$	**79.5**$^{1.0\uparrow}$	**69.5**$^{2.7\uparrow}$	**59.2**$^{1.0\uparrow}$	**86.6**$^{0.7\uparrow}$	**66.8**$^{2.5\uparrow}$	**32.7**$^{1.6\uparrow}$	**66.1**$^{0.7\uparrow}$

消融实验-scalablity

Table 2: Exploring the Compatibility and Scalability of Dense Connector (DC). Scaling results on visual encoder (VE), resolution (Res.), pre-training (PT) / instruction tuning (IT) data, and LLM are provided. "0.5M+0.6M" denotes the training data from LLaVA-1.5 [16], while "1.2M+1.5M" denotes the data from Mini-Gemini [18]. * indicates results evaluated using official model.

Method	VE	Res.	PT+IT	LLM	**GQA**	**SQA**^{I}	**VQA**^{T}	**MMB**	**MMV**	**MMMU**^{v}	**Math**
				Scaling to more powerful visual encoder							
LLaVA [16]	CLIP-L	336	0.5M+0.6M	Vicuna-7B	62.0	66.8	58.2	64.3	31.1	35.3*	24.9*
LLaVA [16]	CLIP-L	336	0.5M+0.6M	Vicuna-13B	63.3	71.6	61.3	67.6	36.1	36.4	27.6
DC (*w/* LLaVA)	CLIP-L	336	0.5M+0.6M	Vicuna-7B	63.8	69.5	59.2	66.8	32.7	34.8	26.9
DC (*w/* LLaVA)	SigLIP-SO	384	0.5M+0.6M	Vicuna-7B	64.2	70.5	62.6	68.4	35.4	**36.7**	25.5
DC (*w/* LLaVA)	SigLIP-SO	384	0.5M+0.6M	Vicuna-13B	**65.4**	**73.0**	**64.7**	**71.4**	**41.6**	34.3	**29.6**
				Scaling to larger-scale training data							
DC (*w/* LLaVA)	SigLIP-SO	384	1.2M+1.5M	Vicuna-7B	63.8	72.9	64.6	71.7	45.0	35.8	33.1
DC (*w/* LLaVA)	SigLIP-SO	384	1.2M+1.5M	Vicuna-13B	**64.6**	**77.1**	**65.0**	**74.4**	**47.7**	**37.2**	**36.5**
				Scaling to high resolution with a dual visual encoder							
MGM [18]	CLIP-L +ConvX-L	336 +768	1.2M+1.5M	Vicuna-7B	62.6*	70.4*	65.2	69.3	40.8	36.1	31.4
MGM [18]	CLIP-L +ConvX-L	336 +768	1.2M+1.5M	Vicuna-13B	63.4*	72.6*	65.9	68.5	46.0	38.1	37.0
DC (*w/* MGM)	CLIP-L +ConvX-L	336 +768	1.2M+1.5M	Vicuna-7B	63.3	70.7	66.0	70.7	42.2	36.8	32.5
DC (*w/* MGM)	CLIP-L +ConvX-L	336 +768	1.2M+1.5M	Vicuna-13B	**64.2**	**74.9**	**66.7**	**70.7**	**49.8**	**39.3**	**38.1**
				Scaling to dynamic high resolution							
LLaVA-NeXT [16]	CLIP-L	AnyRes	0.5M+0.6M	Vicuna-7B	64.0	69.5	64.5	66.5	33.1	35.4	25.7
DC (*w/* LLaVA)	CLIP-L	AnyRes	0.5M+0.6M	Vicuna-7B	64.6	**70.5**	65.6	**67.4**	33.7	**37.6**	26.2
DC (*w/* LLaVA)	SigLIP-SO	AnyRes	0.5M+0.6M	Vicuna-7B	**64.8**	69.3	**66.5**	67.2	**34.8**	36.3	**27.0**

和其他方法的对比

Table 3: Comparison of Efficient Dense Connector with Other Efficient Methods. * indicates results evaluated using official model.

Method	Res.	#Token	PT+IT	LLM	**GQA**	**VQA**^{v2}	**SQA**^{I}	**VQA**^{T}	**MMB**	**MMV**	**Math**
LLaVA [16]	336	576	0.5M+0.6M	Vicuna-7B	62.0	78.5	66.8	**58.2**	64.3	31.1	24.9*
Qwen-VL-Chat [24]	448	256	1.4B+50M	Qwen-7B	57.5	68.2	61.5	-	-	-	-
TokenPacker [56]	336	144	0.5M+0.6M	Vicuna-7B	61.9	77.9	-	-	65.1	33.0	-
Dense Connector	336	144	0.5M+0.6M	Vicuna-7B	**62.8**	**79.4**	**68.8**	58.1	**67.6**	**34.4**	**25.8**

Table 4: Comparisons with State-of-the-Arts. * indicates the dataset have been used for training, and † indicates the dataset is not publicly accessible. "PT," "IT," and "Res." denote pre-training data, instruction fine-tuning data, and image resolution, respectively.

Method	PT+IT	Res.	LLM	SQA^I	MMB	MME^p	MM-Vet	$MMMU^v$	Math	$LLaVA^W$	GQA
MobileVLM V2 [57]	1.2M+3.6M	336	ML-2.7B	70.0	63.2	1441	–	–	–	–	61.1
TinyLLaVA [72]	0.5M+0.6M	384	Phi2-2.7B	69.9	–	–	32.1	–	–	67.9	61.3
mPLUG-Owl2 [73]	348M+1.2M	448	Llama2-7B	68.7	64.5	1450	36.2	32.7	22.2	–	56.1
Qwen-VL-Chat† [24]	1.4B+50M	448	Qwen-7B	68.2	60.6	1488	–	–	–	–	57.5*
LLaVA-v1.5 [16]	0.5M+0.6M	336	Vicuna-13B	71.6	67.7	1531	36.1	36.4	27.6	72.5	63.3
ShareGPT4V [17]	1.2M+0.7M	336	Vicuna-13B	71.2	68.5	1619	43.1	–	–	79.9	64.8
MobileVLM V2 [57]	1.2M+3.6M	336	Vicuna-7B	74.8	70.8	1559	–	–	–	–	64.6
LLaMA-VID [74]	0.8M+0.7M	336	Vicuna-7B	70.0	66.6	1542	–	–	–	–	65.0*
SPHINX-Plus [75]	16M	448	Llama2-13B	74.2	71.0	1458	47.9	–	36.8	71.7	–
LLaVA-LLaMA3 [76]	0.5M+0.6M	336	Llama3-8B	73.3	68.9	1506	–	36.8	–	–	63.5
CuMo [77]	0.5M+0.6M	336	Mistral-7B	71.7	69.6	1429	34.3	–	–	68.8	63.2
MM1 [77]	3B+1.4M	1344	MM1-7B	72.6	79.0	1529	42.1	37.0	35.9	81.5	–
VILA [78]	50M+1M	336	Llama-2-13B	73.7	70.3	1570	38.8	–	–	73.0	63.3*
Mini-Gemini [18]	1.2M+1.5M	336+768	Vicuna-13B	72.6	68.5	1565	46.0	38.1	37.0	87.7	63.4
LLaVA-NeXT [25]	0.5M+0.7M	336_{AnyRes}	Vicuna-13B	73.6	70.0	1575	48.4	36.2	35.3	87.3	65.4
Scaling to a wider range of parameter sizes (2B → 70B) for LLMs											
Dense Connector	0.5M+0.6M	384	Phi2-2.7B	70.3	70.5	1487	33.8	36.6	28.2	65.1	61.5
Dense Connector	0.5M+0.6M	384	Vicuna-7B	70.5	68.4	1523	35.4	36.7	25.5	67.4	64.4
Dense Connector	0.5M+0.6M	384	Vicuna-13B	73.0	71.4	1569	41.6	34.3	29.6	73.6	**65.4**
Dense Connector	0.5M+0.6M	384	Llama3-8B	75.2	74.4	1558	34.6	40.4	28.6	68.8	65.1
Dense Connector	0.5M+0.6M	384	$Yi\text{-}34B_{LoRA}$	80.5	77.7	1588	41.0	47.1	33.5	75.1	63.9
Dense Connector	0.5M+0.6M	384	$Llama3\text{-}70B_{LoRA}$	**82.4**	79.4	1622	46.1	47.0	32.9	74.5	64.0
Dense Connector	1.2M+1.5M	384	Vicuna-13B	77.1	74.4	1579	47.8	37.2	36.5	88.9	64.6
Dense Connector	1.2M+1.5M	384_{AnyRes}	Vicuna-7B	72.0	69.2	1535	44.4	36.4	32.7	88.8	63.9
Dense Connector	1.2M+1.5M	384_{AnyRes}	Vicuna-13B	75.2	72.3	1573	47.0	36.8	35.5	93.2	64.3
Dense Connector	1.2M+1.5M	384_{AnyRes}	Yi-34B	78.0	**81.2**	**1696**	**59.2**	**51.8**	**40.0**	**97.7**	**66.6**

Table 5: Comparisons with Leading Methods on Zero-shot Video QA Benchmarks. Following FreeVA [54], we specify the GPT-3.5 versions used for evaluation to ensure fairness in performance comparison across different versions. "MAR" denotes the `GPT-3.5-Turbo-0301`, "JUN" denotes the `GPT-3.5-Turbo-0613`, and "JAN" denotes the latest `GPT-3.5-Turbo-0125`.

Method	LLM Size	GPT-3.5 Version	Train Free	MSVD-QA		MSRVTT-QA		ActivityNet-QA		Video-ChatGPT benchmark				
				Acc	Score	Acc	Score	Acc	Score	CI	DO	CU	TU	CO
FrozenBiLM [79]	0.9B	MAR	✗	33.8	–	16.7	–	25.9	–	–	–	–	–	–
Video-LLaMA[52]	7B	MAR	✗	51.6	2.5	29.6	1.8	12.4	1.1	1.96	2.18	2.16	1.82	1.79
LLaMA-Adapter [80]	7B	MAR	✗	–	–	–	–	–	–	2.03	2.32	2.30	1.98	2.15
VideoChat [81]	7B	MAR	✗	56.3	2.8	45.0	2.5	26.5	2.2	2.23	2.50	2.53	1.94	2.24
Video-ChatGPT [51]	7B	MAR	✗	64.9	3.3	49.3	2.8	35.2	2.7	2.50	2.57	2.69	2.16	2.20
VaQuitA [82]	7B	MAR	✗	74.6	3.7	68.6	3.3	48.8	3.3	–	–	–	–	–
LLaVA+FreeVA [54]	7B	MAR	✓	81.5	4.0	72.9	3.5	58.3	3.5	2.88	2.52	3.25	2.32	3.07
BT-Adapter [83]	7B	JUN	✗	67.5	3.7	57.0	3.2	45.7	3.2	2.68	2.69	3.27	2.34	2.46
Video-LLaVA [53]	7B	JUN	✗	70.7	3.9	59.2	3.5	45.3	3.3	–	–	–	–	–
LLaMA-VID [74]	13B	JUN	✗	70.0	3.7	58.9	3.3	47.5	3.3	3.07	3.05	3.60	2.58	2.63
LLaVA+FreeVA [54]	13B	JUN	✓	71.8	3.8	59.2	3.3	54.5	3.5	2.90	2.52	3.26	2.32	3.07
LLaVA+FreeVA [54]	13B	JAN	✓	74.4	4.1	61.1	3.6	51.6	3.5	2.88	2.52	3.25	2.34	3.05
DC+FreeVA	7B	JAN	✓	75.0	4.1	58.4	3.5	52.2	3.5	2.80	2.51	3.17	2.22	3.05
DC+FreeVA	13B	JAN	✓	75.1	4.1	60.8	3.5	52.6	3.5	2.85	2.53	3.23	2.29	2.96
DC+FreeVA	34B	JAN	✓	77.4	4.2	62.1	3.6	55.8	3.6	3.00	2.53	3.25	2.65	2.92

Limitation?

- 模型在融合时没有引入新的参数
- 作者没有找到一种引入新参数的好方法
- 期待以后可以有更好的办法来连接视觉编码器和语言模型