2024 SAIL Seminar

EAGLE: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders

Min Shi et al.(NVIDIA), arXiv 2024

순천향대학교 미래융합기술학과

Senseable AI Lab

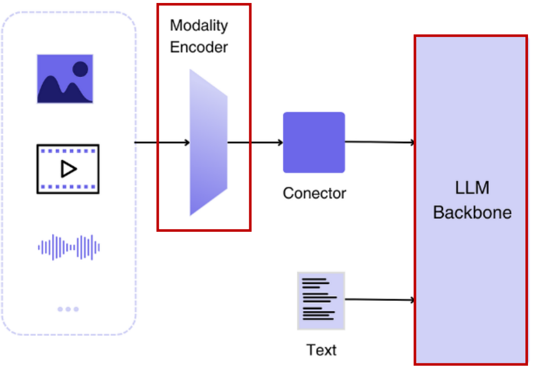
석사과정 김병훈

Introduction

Multi-modal Large Language Model(MLLM) Task

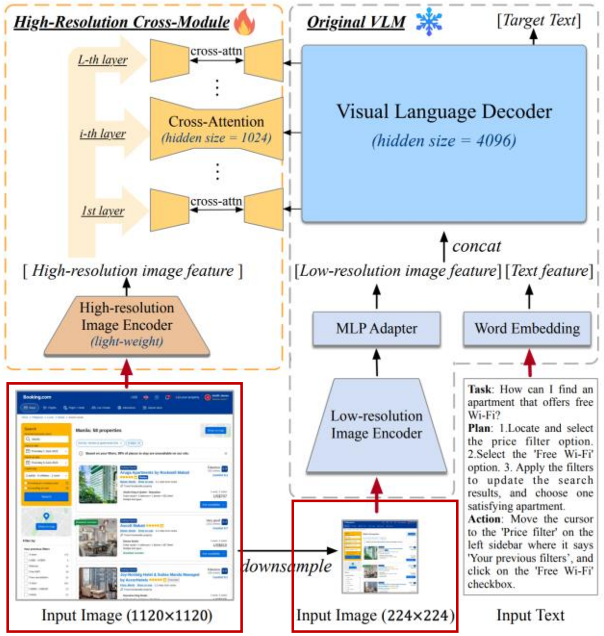
- LLM을 기반으로 다양한 모달리티 작업이 가능하도록 하는 작업 (≒LMM)
- 기본적 형태
  - Vision Encoder: 이미지를 Visual Token으로 변환(Pre-trained)
  - Text Encoder: Word Embedding
  - Other(Audio, Video etc.)
  - 위의 Feature를 Alignment → LLM에 입력 → 결과 출력

Multimodal Model Architecture



Mixture-of-Vision-Encoder

- 대부분의 LMM 모델은 사전 학습된 비전 인코더와 LLM 시퀀스 길이의 제한으로 낮은 입력 해상도를 가짐
  - OCR과 같은 해상도에 민감한 LMM Task는 이로 인해 제약을 받음
- 더 강력한 비전 인코더 설계 → LLM hallucination 완화, 해상도에 민감한 Task 개선
- 서로 다른 task와 입력 해상도로 사전 학습된 비전 인코더를 혼합 → 매우 효과적인 것으로 나타남



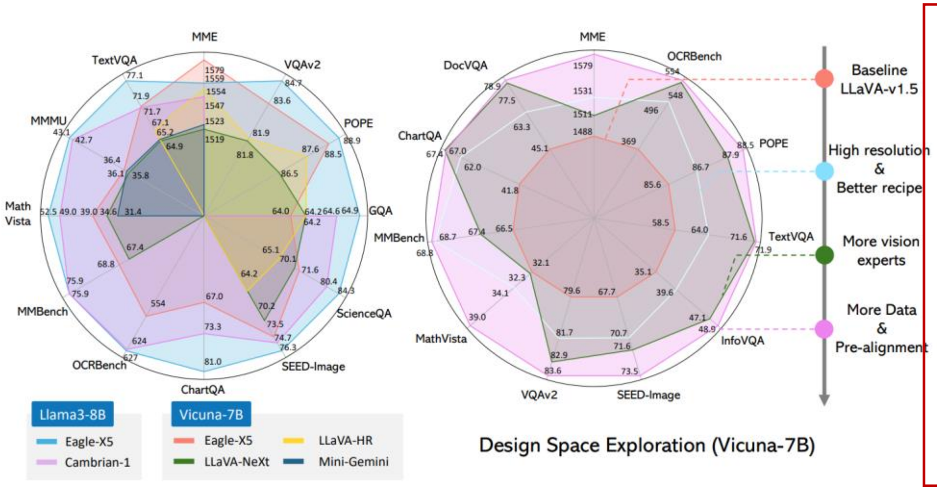
High-Resolution Cross-Module
L-th layer
cross-attn
i-th layer
Cross-Attention (hidden size = 1024)
1st layer
cross-attn
[High-resolution image feature]
High-resolution Image Encoder (light-weight)
Original VLM
[Target Text]
Visual Language Decoder (hidden size = 4096)
concat
[Low-resolution image feature] [Text feature]
MLP Adapter
Word Embedding
Low-resolution Image Encoder
downsample
Input Image (1120×1120)
Input Image (224×224)
Input Text

Task: How can I find an apartment that offers free Wi-Fi?
Plan: 1.Locate and select the price filter option. 2.Select the 'Free Wi-Fi' option. 3. Apply the filters to update the search results, and choose one satisfying apartment.
Action: Move the cursor to the 'Price filter' on the left sidebar where it says 'Your previous filters', and click on the 'Free Wi-Fi' checkbox.

- *그렇다면 어떤 비전 인코더 조합을 선택해야 할까?*
- *다양한 Vision Expert들을 어떻게 융합해야 할까?*
- *더 많은 비전 인코더에 따라 학습 전략을 어떻게 조정해야 할까?*

Introduction

Mixture-of-Vision-Encoder Design Space



Design Space Exploration (Vicuna-7B)

Eagle

Highlights

- ✓ Strong performance
- ✓ Thorough design space exploration
- ✓ Fully open-source (data + recipe)

Vision-Centric Design

- ✓ Mixture of vision encoders
- ✓ Simple & extendable fusion design
- ✓ Optimal encoder combination

Advanced Training Recipes

- ✓ High resolution & fusion adaptation
- ✓ Vision expert pre-alignment

Mixture-of-Vision-Encoder Design Space 탐색 방법

1) 다양한 비전 인코더 벤치마킹 및 High resolution 적용을 위한 레시피 탐색
2) 비전 인코더 융합 전략 간의 비교
3) 여러 비전 인코더의 최적 조합에 대한 점진적 식별
4) 개선된 vision expert pre-alignment 및 데이터 혼합

⇩

EAGLE

Exploring The Design Space for Multimodal LLMs with Mixture of Encoders

- 다양한 벤치마크에서 SOTA 성능 달성
- OCR, Text Recognition Task에 엄청난 이점

Summary of Results

1. LMM 학습 중에 비전 인코더를 고정 해제하는 것이 중요
2. Channel concatenation이 간단하면서도 경쟁력 있는 융합 전략
3. 추가 vision expert를 통합하면 일관된 성능 개선이 발생
4. pre-alignment 단계 제안 → 성능 향상
   - vision expert 학습되기 전, 고정된 LLM 활용하여 개별적으로 fine-tuning

Design space exploration

Base Setup

- 기반 아키텍쳐: LLaVA
- 구성 요소
    - ① Vision Encoder; ② LLM; ③Projector;

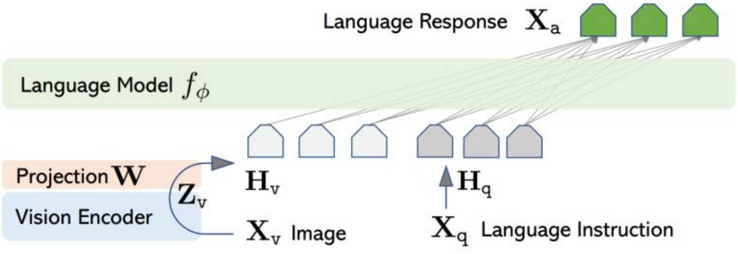
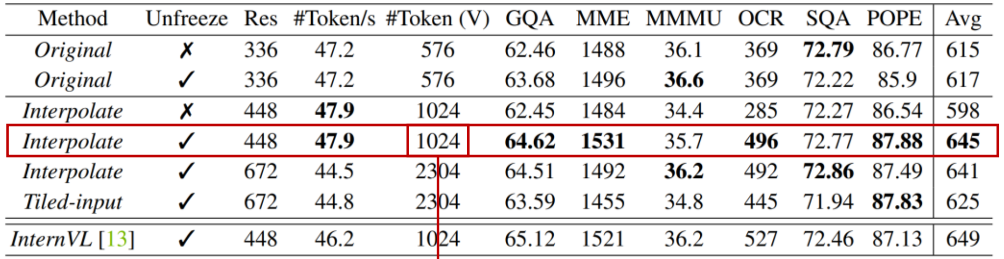


Figure 1: LLaVA network architecture.

High Resolution Recipe

- 기존 LMM은 사전 학습된 CLIP 해상도 (ex. 224×224, 336×336)를 입력 해상도로 사용
- Low → High resolution 전략 (Model Freeze/Unfreeze)
    - 입력 이미지를 타일로 나누어 별도로 인코딩
    - 입력 해상도를 직접 확대 → ViT 모델의 위치 임베딩을 보간

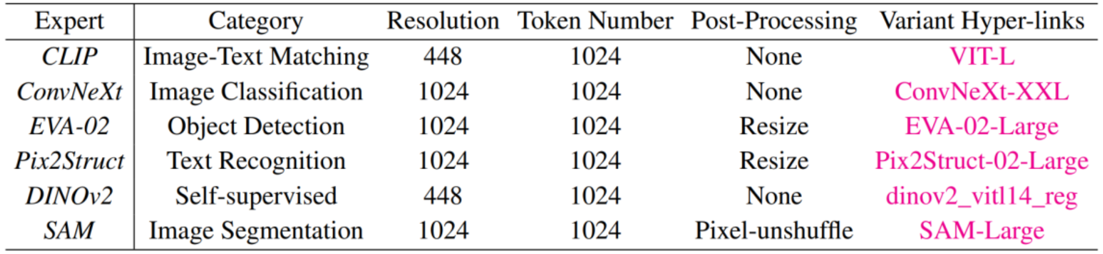


Method	Unfreeze	Res	#Token/s	#Token (V)	GQA	MME	MMMU	OCR	SQA	POPE	Avg
Original	✗	336	47.2	576	62.46	1488	36.1	369	**72.79**	86.77	615
Original	✓	336	47.2	576	63.68	1496	**36.6**	369	72.22	85.9	617
Interpolate	✗	448	**47.9**	1024	62.45	1484	34.4	285	72.27	86.54	598
Interpolate	✓	448	**47.9**	1024	**64.62**	**1531**	35.7	**496**	72.77	**87.88**	**645**
Interpolate	✓	672	44.5	2304	64.51	1492	**36.2**	492	**72.86**	87.49	641
Tiled-input	✓	672	44.8	2304	63.59	1455	34.8	445	71.94	**87.83**	625
InternVL [13]	✓	448	46.2	1024	65.12	1521	36.2	527	72.46	87.13	649

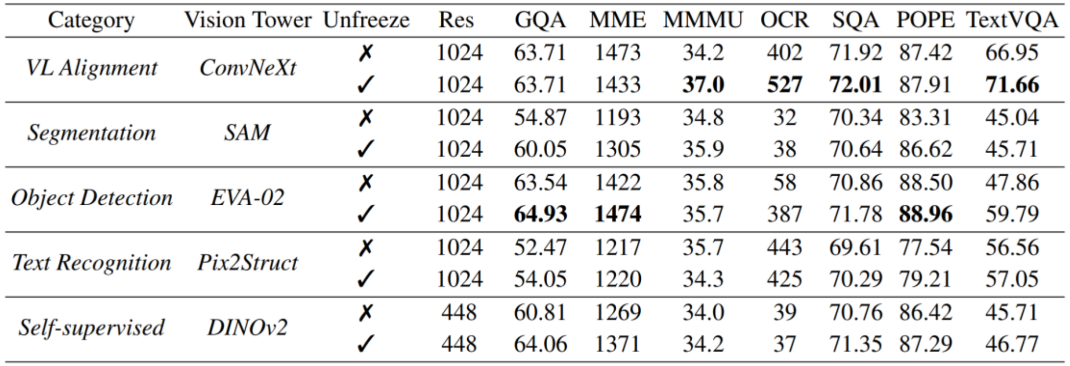
1024개의 Token을 출력하도록 설정 고정

Vision Experts

bilinear interpolation
pixel shuffle



Expert	Category	Resolution	Token Number	Post-Processing	Variant Hyper-links
CLIP	Image-Text Matching	448	1024	None	VIT-L
ConvNeXt	Image Classification	1024	1024	None	ConvNeXt-XXL
EVA-02	Object Detection	1024	1024	Resize	EVA-02-Large
Pix2Struct	Text Recognition	1024	1024	Resize	Pix2Struct-02-Large
DINOv2	Self-supervised	448	1024	None	dinov2_vitl14_reg
SAM	Image Segmentation	1024	1024	Pixel-unshuffle	SAM-Large

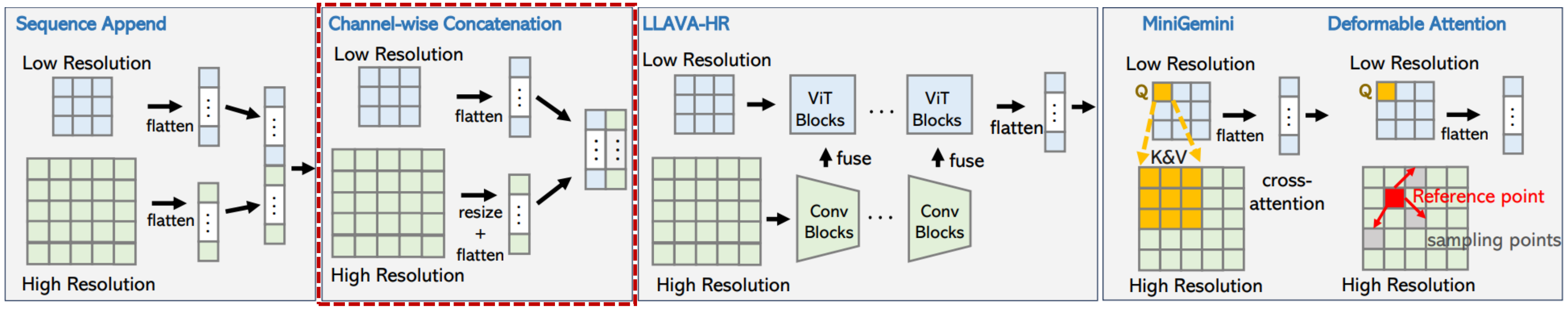


Category	Vision Tower	Unfreeze	Res	GQA	MME	MMMU	OCR	SQA	POPE	TextVQA
VL Alignment	*ConvNeXt*	✗	1024	63.71	1473	34.2	402	71.92	87.42	66.95
		✓	1024	63.71	1433	**37.0**	**527**	**72.01**	87.91	**71.66**
Segmentation	*SAM*	✗	1024	54.87	1193	34.8	32	70.34	83.31	45.04
		✓	1024	60.05	1305	35.9	38	70.64	86.62	45.71
Object Detection	*EVA-02*	✗	1024	63.54	1422	35.8	58	70.86	88.50	47.86
		✓	1024	**64.93**	**1474**	35.7	387	71.78	**88.96**	59.79
Text Recognition	*Pix2Struct*	✗	1024	52.47	1217	35.7	443	69.61	77.54	56.56
		✓	1024	54.05	1220	34.3	425	70.29	79.21	57.05
Self-supervised	*DINOv2*	✗	448	60.81	1269	34.0	39	70.76	86.42	45.71
		✓	448	64.06	1371	34.2	37	71.35	87.29	46.77

Design space exploration

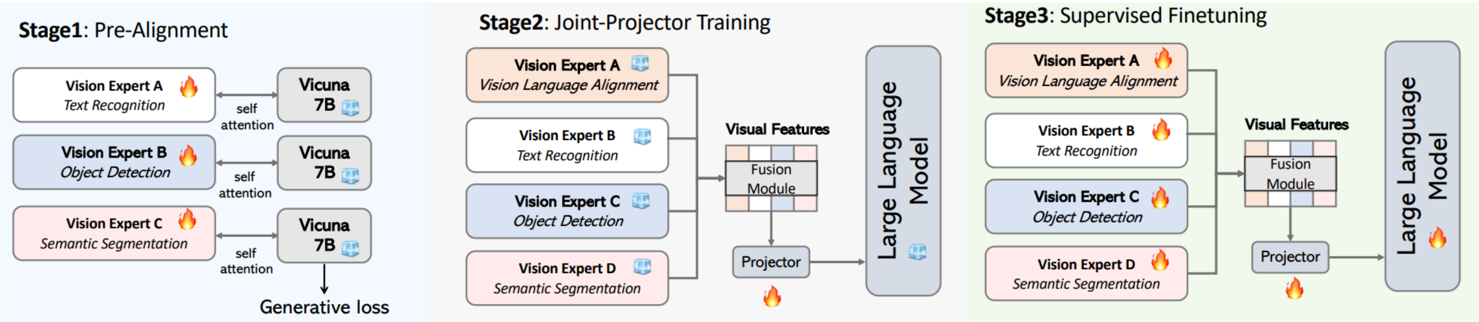
Fusion strategy

- Sequence Append
- Channel Concatenation
- LLaVA-HR
- MiniGemini
- Deformable Attention



Vison-language Pre-Alignment

- 비전 task에 대해서만 사전 학습된 비전 backbone → LLM과 통합할 때 표현 불일치가 발생 → 학습 프로세스를 통합하기 어려움
- Pre-Alignment 학습 단계 제안 → 학습 프로세스 안정화 / 성능 개선 / Vision expert 간 편향 완화



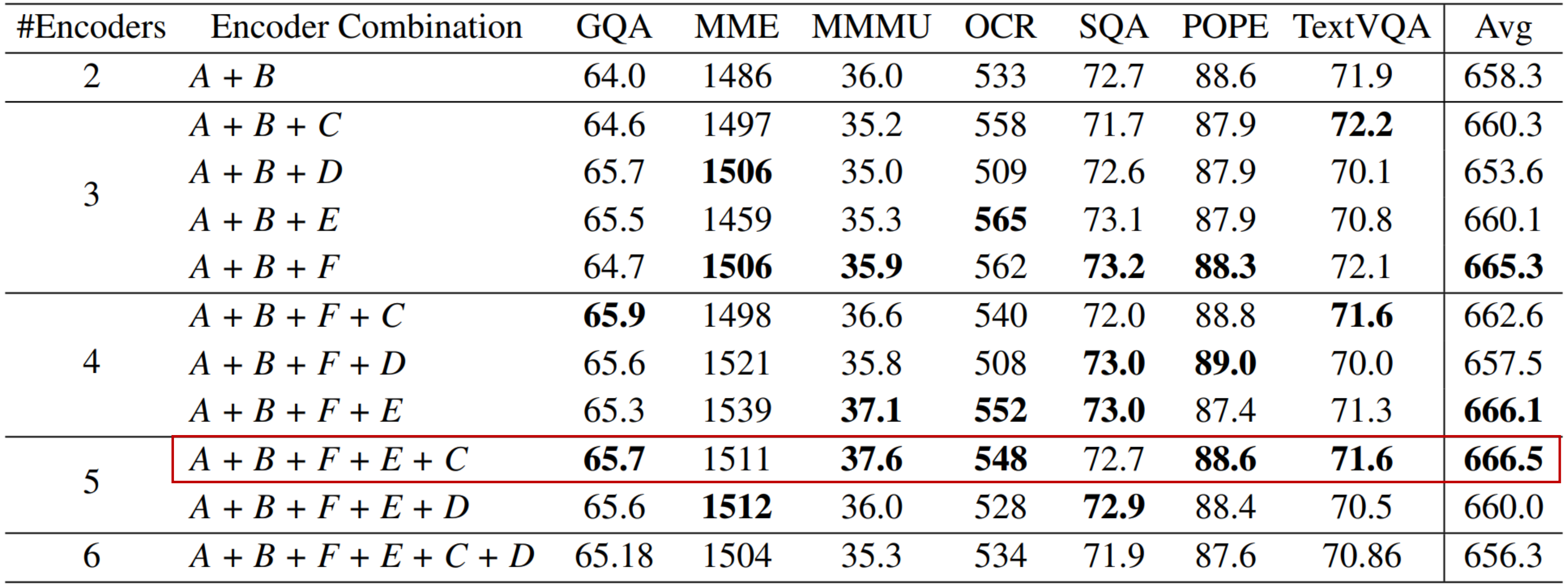
Pre-Alignment 학습

1) LLM 고정, Vision expert를 각자의 projector로 학습
2) Visual Features 결합, Vision expert, LLM 고정, projector만 학습
3) 모델 전체 학습

Design space exploration

Extension to multi-experts

- A: CLIP
- B: ConvNeXt
- C: SAM
- D: DINOv2
- E: Pix2Struct
- F: EVA-02-L



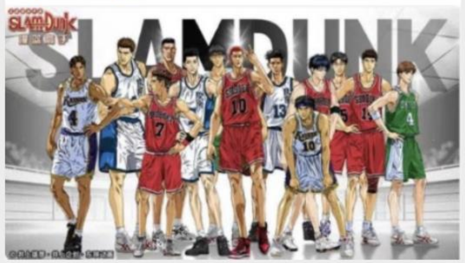
#Encoders	Encoder Combination	GQA	MME	MMMU	OCR	SQA	POPE	TextVQA	Avg
2	$A + B$	64.0	1486	36.0	533	72.7	88.6	71.9	658.3
3	$A + B + C$	64.6	1497	35.2	558	71.7	87.9	**72.2**	660.3
	$A + B + D$	65.7	**1506**	35.0	509	72.6	87.9	70.1	653.6
	$A + B + E$	65.5	1459	35.3	**565**	73.1	87.9	70.8	660.1
	$A + B + F$	64.7	**1506**	**35.9**	562	**73.2**	**88.3**	72.1	**665.3**
4	$A + B + F + C$	**65.9**	1498	36.6	540	72.0	88.8	**71.6**	662.6
	$A + B + F + D$	65.6	1521	35.8	508	**73.0**	**89.0**	70.0	657.5
	$A + B + F + E$	65.3	1539	**37.1**	**552**	**73.0**	87.4	71.3	**666.1**
5	$A + B + F + E + C$	**65.7**	1511	**37.6**	**548**	72.7	**88.6**	**71.6**	**666.5**
	$A + B + F + E + D$	65.6	**1512**	36.0	528	**72.9**	88.4	70.5	660.0
6	$A + B + F + E + C + D$	65.18	1504	35.3	534	71.9	87.6	70.86	656.3

- *CLIP, ConvNeXt, SAM, Pix2Struct, EVA-02 조합이 가장 좋음*

Experiments

Compare results by number of vision experts

Input Image:

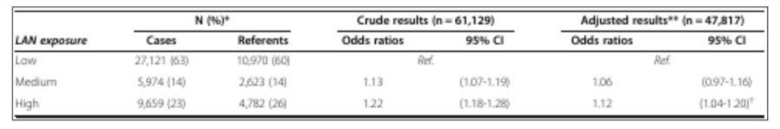


User: *Tell me the locations of all the number 10 jerseys in the picture. Answer the questions in details with a few sentences*

Eagle (baseline): *There are two number 10 jerseys in the picture. One is on the left side of the group, and the other is on the right side*

Eagle (more vision experts): *In the image, there are two players wearing jerseys with the number 10. One of them is standing in the center of the group, while the other is positioned towards the right side of the group. Both players are part of the basketball team, as indicated by their uniforms and the context of the image.*

Input Image:



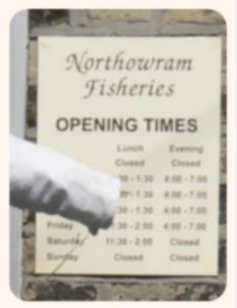
	N (%)*		Crude results (n = 61,129)		Adjusted results** (n = 47,817)	
LAN exposure	Cases	Referents	Odds ratios	95% CI	Odds ratios	95% CI
Low	27,121 (63)	10,970 (60)	*Ref.*		*Ref.*	
Medium	5,974 (14)	2,623 (14)	1.13	(1.07-1.19)	1.06	(0.97-1.16)
High	9,659 (23)	4,782 (26)	1.22	(1.18-1.28)	1.12	(1.04-1.20)†

User: *What's the value of the row "low" and column "referents"?*

Eagle (baseline): *2012-13*

Eagle (more vision experts): *10,970 (60%)*

Input Image:

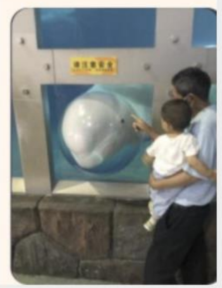


User: *What's the opening time of the restaurant on Sunday?*

Eagle (baseline): *11:30-2:00*

Eagle (more vision experts): *closed*

Input Image:

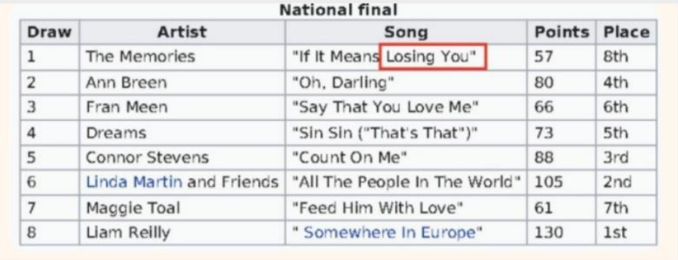


User: *What is the man in the picture doing with his right hand?*

Eagle (baseline): *The man in the image is holding his child with his right hand.*

Eagle (more vision experts): *The man in the picture is pointing at a white object, which is a dolphin, with his right hand.*

Input Image:



National final

Draw	Artist	Song	Points	Place
1	The Memories	"If It Means Losing You"	57	8th
2	Ann Breen	"Oh, Darling"	80	4th
3	Fran Meen	"Say That You Love Me"	66	6th
4	Dreams	"Sin Sin ("That's That")"	73	5th
5	Connor Stevens	"Count On Me"	88	3rd
6	Linda Martin and Friends	"All The People In The World"	105	2nd
7	Maggie Toal	"Feed Him With Love"	61	7th
8	Liam Reilly	" Somewhere In Europe"	130	1st

User: *Describe the information in the red bounding box.*

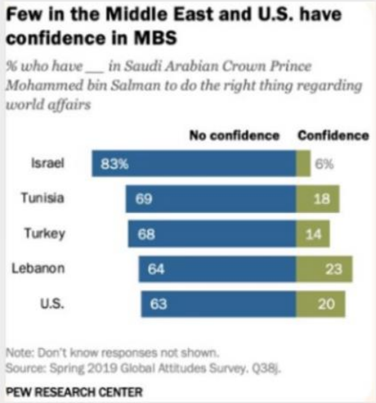
Eagle (baseline): *The song "If It Means Losing You" by The Memories is highlighted in red*

Eagle (more vision experts): *The red bounding box highlights the song "Losing You" by The Memories, which is the national final song.*

Experiments

OCR and document understanding tasks

Input Image:



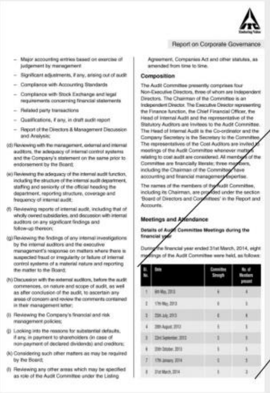
User: *Render the figure in Markdown format.*

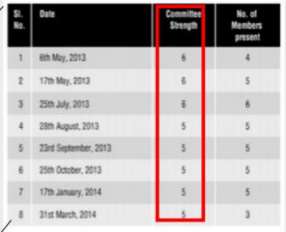
Eagle:

```
Few in the Middle East and U.S. have confidence in MBs

| Country | No confidence | Confidence |
|----------|--------------|------------|
| Israel   | 83% | 6% |
| Tunisia  | 69% | 18% |
| Turkey   | 68% | 14% |
| Lebanon  | 64% | 23% |
| U.S.     | 63% | 20% |
```

Input Image:





Sl. No.	Date	Committee Strength	No. of Members present
1	6th May, 2013	6	4
2	17th May, 2013	6	5
3	25th July, 2013	6	6
4	28th August, 2013	5	5
5	23rd September, 2013	5	5
6	25th October, 2013	5	5
7	17th January, 2014	5	5
8	31st March, 2014	5	3

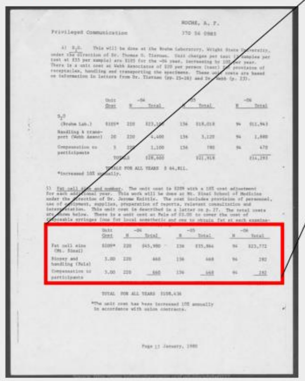
User: *What was the committee strength for the first meeting?*

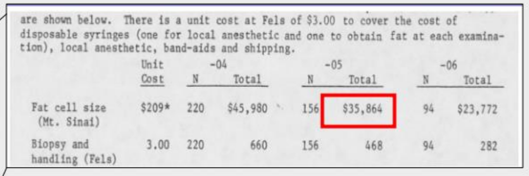
Eagle: *6*

User: *What was the committee strength for the last meeting?*

Eagle: *5*

Input Image:





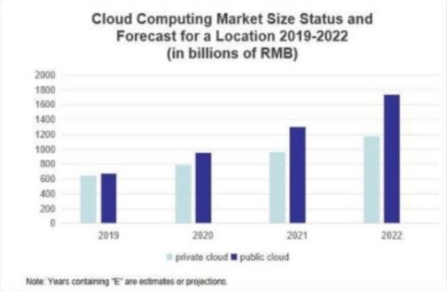
are shown below. There is a unit cost at Fels of $3.00 to cover the cost of disposable syringes (one for local anesthetic and one to obtain fat at each examination), local anesthetic, band-aids and shipping.

	Unit	-04		-05		-06	
	Cost	N	Total	N	Total	N	Total
Fat cell size (Mt. Sinai)	$209*	220	$45,980	156	$35,864	94	$23,772
Biopsy and handling (Fels)	3.00	220	660	156	468	94	282

User: *What's the total cost of Fat cell size (Mt. Sinai) in the -05 year?*

Eagle: *$35,864*

Input Image:



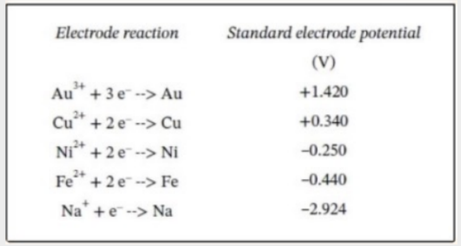
User: *Describe the chart in detail?*

Eagle: *The image presents a bar chart that provides a forecast for the cloud computing market size in China for the years 2019 to2022. The chart is divided into two columns, one for private cloud and the other for public cloud. The years are labeled on the x-axis, while the market size is represented on the y-axis in billions of RMB, The chart shows that the market size for private cloud and public cloud have been steadily increasing over the years. The highest market size is for 2022, which is for public cloud, indicating a significant growth in this sector. The lowest market size is for 2019, which is for private cloud, suggesting that the private cloud market has been growing at a slower pace compared to the public cloud market.*

Experiments

VQA tasks

Input Image:

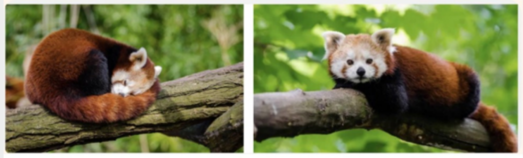


Electrode reaction	Standard electrode potential (V)
Au^{3+} + 3 e^- --> Au	+1.420
Cu^{2+} + 2 e^- --> Cu	+0.340
Ni^{2+} + 2 e^- --> Ni	−0.250
Fe^{2+} + 2 e^- --> Fe	−0.440
Na^+ + e^- --> Na	−2.924

User: *Below are shown, for five metals, reduction reactions and standard electrode potential values. Which of these metals is the least reactive? explain your answers in detail. Option: (A)Au (B) Cu (C) Ni(D) Fe (E) Na*

Eagle: *The least reactive metal among the given options is (A) Au. Explanation: The standard electrode potential (E") is a measure of the tendency of a metal to lose electrons and undergo a reduction reaction. The higher the E" value, the more reactive the metal is. In the given table, the standard electrode potential values are as follows. (A)Au: +1.420 V (B) Cu:+0.340 V (C) Ni:-0.250V (D) Fe:-0.440V (E) Na:-2.924V Comparing these values, we can see that the metal with the highest E"value (+1,420 V) is Au. which indicates that it is the least reactive among the given options. The other metals have lower E" values, making them more reactive than Au.*

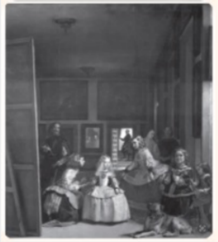
Input Image:



User: *What are the similarities and differences between these two pictures?*

Eagle: The two pictures show a red panda in a natural setting, resting on a tree branch. The similarities between the two images include the presence of the red panda, the tree branch, and the natural environment. The differences lie in the position of the red panda and the angle of the photographs. In the first picture, the red panda is curled up and sleeping on the tree branch, while in the second picture, the red panda is sitting upright on the branch, looking alert and awake. The first photo captures the red panda in a more relaxed state, while the second photo shows the animal in a more active and attentive position.

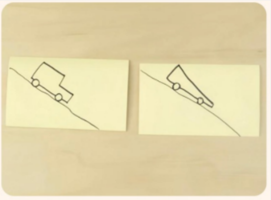
Input Image:



User: *The original placement of this painting was Option: (A)in the Prado Museum in Madrid, showing a willingness to be compared to great Spanish painters of the past (B) in the cathedral in Madrid, indicating a desire to find acceptance with the Catholic Church (C) in the study of king Philip Iv, challenging him to consider the multitude of interpretations the painting represents (D) in the reception room at the Royal Palace in Madrid, where the portraits of the royal family would be viewable by visiting dignitaries*

Eagle: (C) in the study of king Philip Iv, challenging him to consider the multitude of interpretations the painting represents

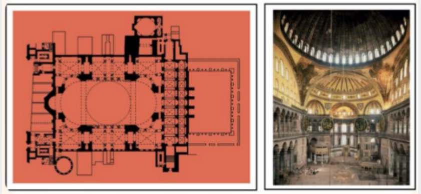
Input Image:



User: *Which car in the picture is more aerodynamic? And explain your answers in detail.*

Eagle: The car in the picture that is more aerodynamic is the one on the right. This is because it has a streamlined shape, with a smooth and curved body that reduces air resistance. The streamlined shape allows the car to move more efficiently through the air, which can result in better fuel efficiency and lower emissions. The car on the left. on the other hand. has amore rectangular shape, which can cause more air resistance and make it less aerodynamic.

Input Image:



User: *Who was the patron of the building?*
Options:
(A) Abd al-Malik
(B) Justinian
(C) Charlemagne (D) shah Jahan

Eagle: (B)