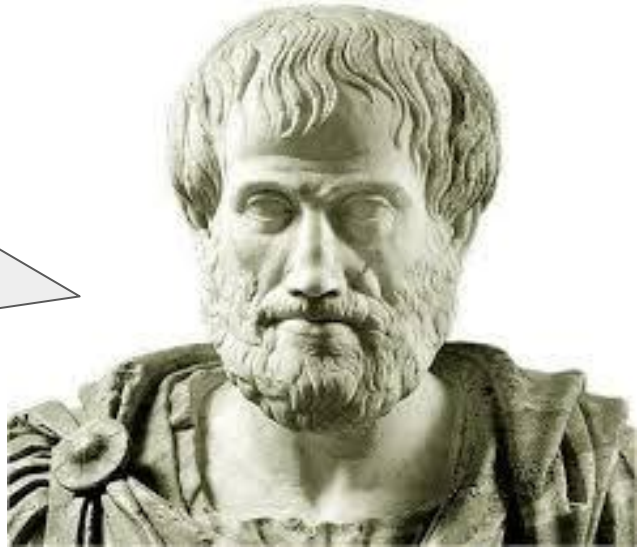
Cambrian-1:
A Fully Open, Vision-Centric Exploration of Multimodal LLMs

12.3 최주연

1. Introduction

"지식을 이해하는 데 감각 경험은 필수일까?"

지식을 얻기 위해서는 **감각 경험과 경험적 관찰**이 중요



1. Introduction

현재 멀티모달 모델 연구에서 제기되는 두 가지 핵심 문제

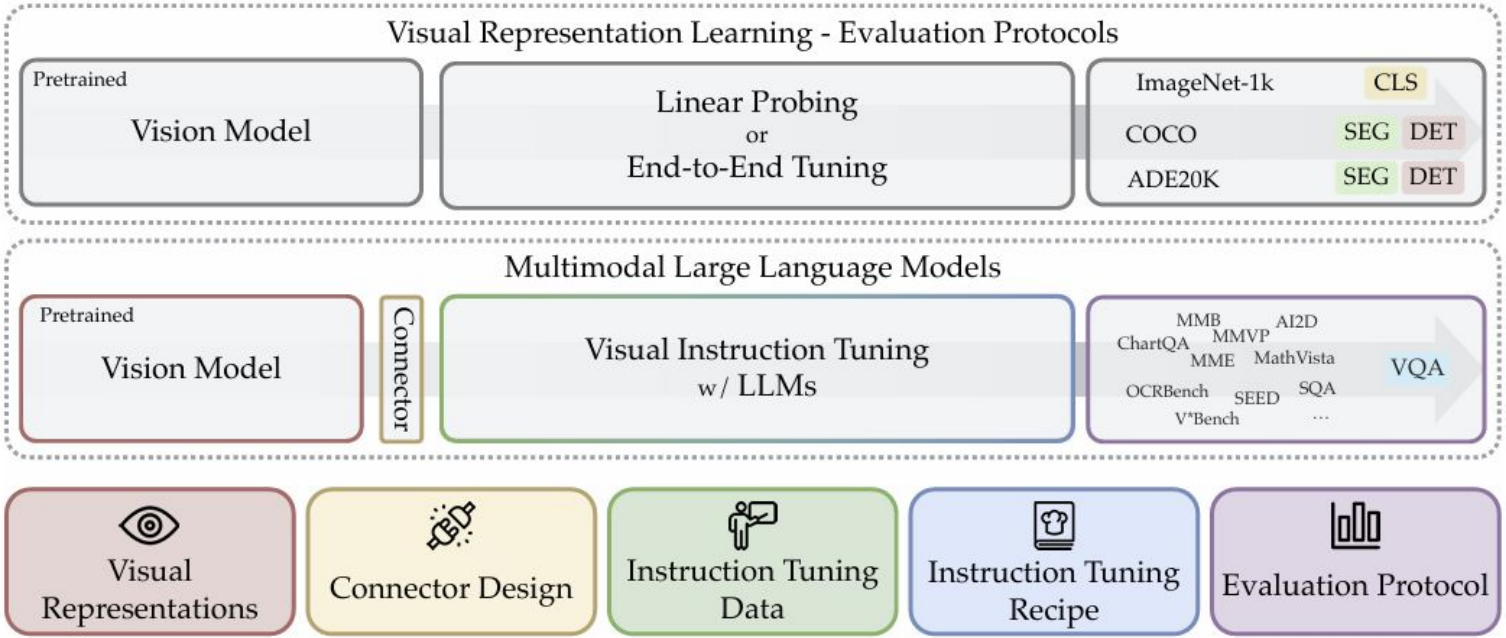
1️⃣ 텍스트 의존적 구조

→ 이미지를 제대로 보지 않아도 그럴듯한 답 생성 (**Shortcut 현상**)

2️⃣ 기존 벤치마크의 한계

→ 시각적 이해(visual grounding)를 제대로 평가하지 못함

1. Introduction



2. Multimodal LLMs: Preliminaries and Related Work

- 대규모 언어 모델(LLM)
- 시각 인코더(Visual Encoder)
- 멀티모달 커넥터(Multimodal Connector)
- 데이터 수집 및 정제 파이프라인(Data Curation Pipeline)
- Instruction Tuning 전략
- 평가 및 벤치마킹 프로토콜(Evaluation & Benchmarking)

2. Multimodal LLMs: Preliminaries and Related Work

Types of **Visual Encoders** in MLLMs



Class Label Supervised
ImageNet-1K [114]

Language Supervised
CLIP [110]

SSL-Contrastive
DINOv2 [104]

SSL-Masking
MAE [53]

Diffusion
Stable Diffusion [113]

Depth Supervised
MiDaS [15]

Segmentation Supervised
SAM [69]

3. Evaluating Visual Representations through MLLMs

질문:
"이 이미지 속 동물은 어디에 있을까?"

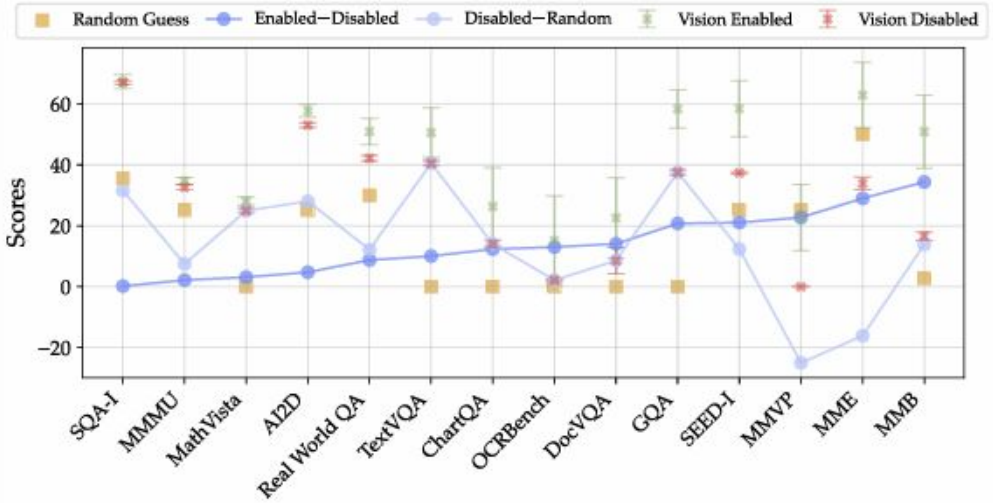
정답: 소파 위

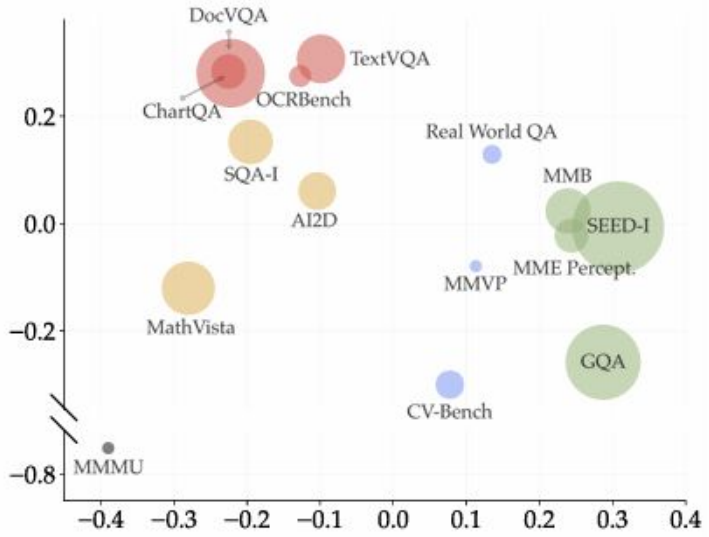
CLIP 기반 MLLM 결과:
"고양이는 소파에 있는 경우가 많아요."

→ 이미지 근거 X, prior 기반 정답



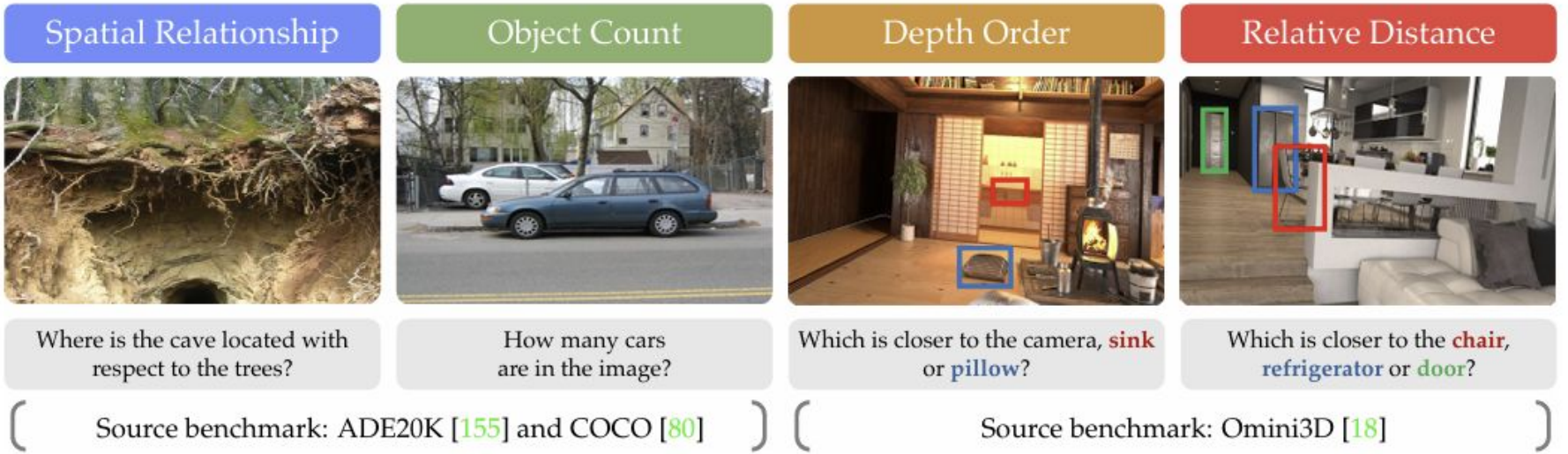
3.1 벤치마크 분석 (Analyzing the Benchmarks)





💡대부분의 벤치마크는 실제 시각 능력을 평가하지 못하며,
그나마 중요한 벤치마크도 샘플 수가 매우 적다.

3.2 새로운 평가 체계 – CV-Bench

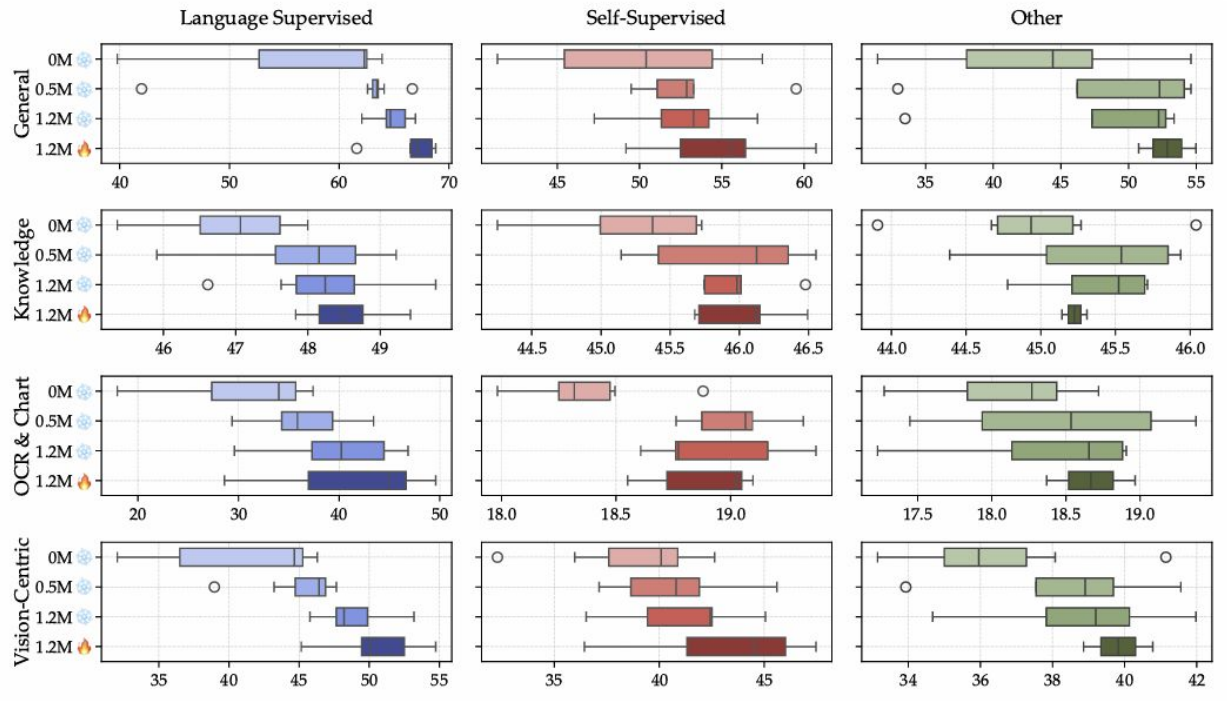


3.2 새로운 평가 체계 – CV-Bench

Type	Task	Description	Sources	# Samples
2D	**Spatial Relationship**	Determine the relative position of an object w.r.t. the anchor object. Consider left-right or top-bottom relationship.	ADE20K COCO	650
	Object Count	Determine the number of instances present in the image.	ADE20K COCO	788
3D	**Depth Order**	Determine which of the two distinct objects is closer to the camera.	Omni3D	600
	Relative Distance	Determine which of the two distinct objects is closer to the anchor object.	Omni3D	600

💡 기존 비전 벤치마크들은 **VQA**형태로 재구성하면
실제 시각 이해 능력을 평가하는 데 효과적으로 활용될 수 있다.

3.3 Instruction Tuning 전략 비교

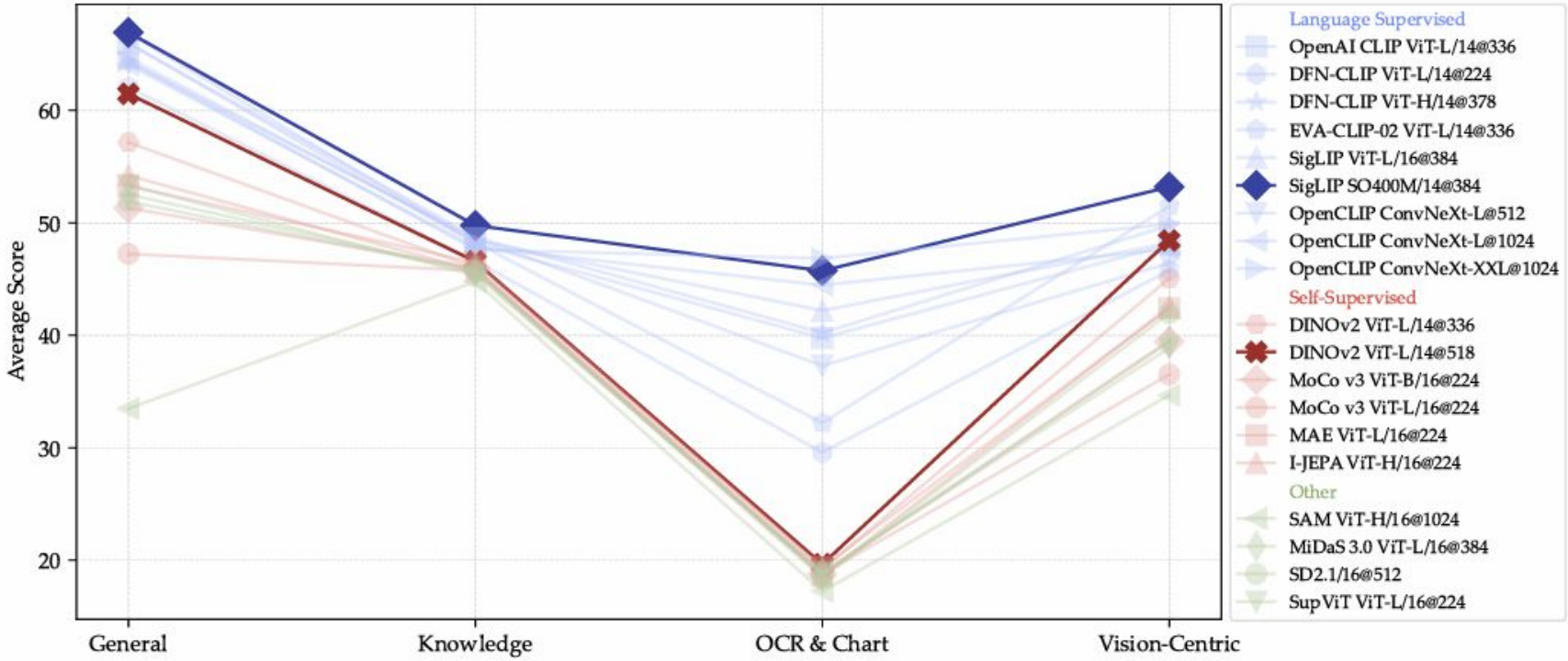


3.3 Instruction Tuning 전략 비교

💡 **2단계 튜닝이 가장 좋고, 데이터가 많을수록 성능이 향상된다.**

💡 **Vision encoder를 unfrozen** 상태로 학습시키는 것이 특히 효과적이다.
→ **SSL** 기반 모델**(DINO2)**은 효과가 극대화됨.

3.3 Instruction Tuning 전략 비교



3.5 다중 비전 모델 조합 (Model Ensembling)

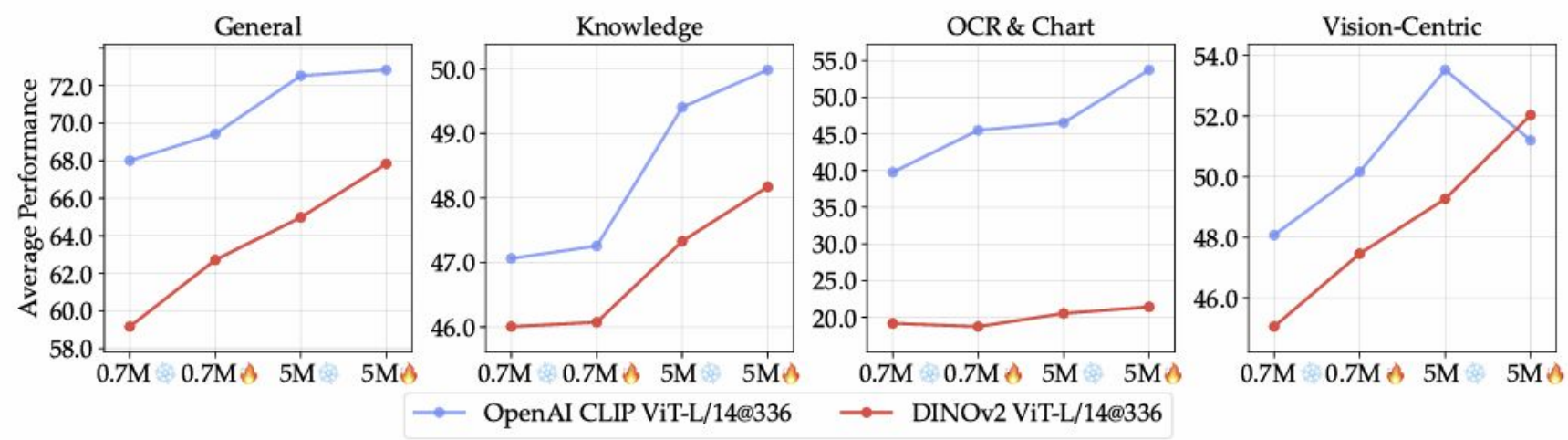
Language Supervised

Model	Architecture	All	G	K	O	V
SigLIP	ViT-SO400M/14@384	1	1	1	2	1
OpenCLIP	ConvNeXt-XXL@1024	2	6	8	1	3
DFN-CLIP	ViT-H/14@378	3	4	2	5	4
OpenCLIP	ConvNeXt-L@1024	4	8	7	3	8
SigLIP	ViT-L/16@384	5	5	4	4	6
OpenAI CLIP	ViT-L/14@336	6	3	6	6	7
EVA-CLIP-02	ViT-L/14@336	7	2	5	8	2
OpenCLIP	ConvNeXt-L@512	8	7	3	7	9
DFN-CLIP	ViT-L/14@224	9	9	9	9	10
DINOv2*	ViT-L/14@518	10	10	10	10	5

Self-Supervised & Other

Model	Architecture	All	G	K	O	V
DINOv2	ViT-L/14@518	1	1	1	1	1
DINOv2	ViT-L/14@336	2	2	3	3	2
MAE	ViT-L/16@224	3	5	2	2	4
I-JEPA	ViT-H/14@224	4	3	6	8	3
SD2.1	VAE+UNet/16@512	5	7	9	9	5
MiDaS 3.0	ViT-L/16@384	6	6	8	5	6
SupViT	ViT-L/16@224	7	4	9	4	8
MoCo v3	ViT-B/16@224	8	8	4	7	7
MoCo v3	ViT-L/16@224	9	9	5	6	9
SAM	ViT-H/16@1024	10	10	10	10	10

3.4. MLLMs as a Visual Representation Evaluator



💡 고해상도 인코더는 **chart·vision-centric** 벤치마크에서 성능을 큰 폭으로 향상시킨다. 그리고 **ConvNet** 기반 아키텍처는 이러한 벤치마크에 특히 적합하다.

3.4. MLLMs as a Visual Representation Evaluator

💡 언어 기반 감독은 분명한 이점을 제공하지만, 충분한 데이터와 적절한 **fine-tuning**이 이루어진다면 **SSL** 기반 모델이 그 격차를 좁힐 수 있다.

3.4. MLLMs as a Visual Representation Evaluator

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Method	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
SigLIP+DINOv2	51.61	1,432.02	61.28	65.99	63.30	68.82	35.69	29.40	60.01	43.00	35.70	60.40	37.54	30.00	53.99	55.52	53.58
SigLIP+DINOv2+ConvNext	54.52	1,503.51	63.83	67.97	63.95	70.40	35.99	29.30	60.69	48.20	36.90	64.97	45.53	34.67	58.69	55.74	60.33
SigLIP+DINOv2+ConvNext+CLIP	54.74	1,479.46	63.32	67.63	64.04	71.39	35.49	29.10	59.88	50.24	39.60	64.55	46.12	32.67	58.95	58.54	60.42
SigLIP+ConvNext	54.53	1,494.97	64.60	67.98	63.58	71.05	34.90	29.80	60.85	50.64	38.00	64.53	46.52	32.00	57.91	58.83	56.58
CLIP+ConvNext	54.45	1,511.08	63.83	67.41	63.63	70.80	35.09	30.40	59.91	51.32	35.00	64.45	47.88	33.33	57.25	56.32	59.08
SigLIP+DINOv2+ConvNext-L	53.78	1,450.64	63.57	67.79	63.63	71.34	34.80	30.20	61.04	49.32	37.70	64.05	45.83	30.00	56.21	58.08	54.33
SigLIP+CLIP+ConvNext-L	54.53	1,507.28	63.23	68.64	63.63	71.10	35.89	30.90	59.97	52.36	38.50	65.40	47.92	28.67	57.25	57.66	55.92

3.5. 여러 비전 인코더 결합

💡 **Self-supervised 모델을 포함한 여러 비전 인코더를 결합하면,**
다양한 벤치마크에서 MLLM 성능을 향상시킬 수 있으며
특히 vision-centric 과제에서 그 효과가 두드러진다.

4. Spatial Vision Aggregator : A New Connector Design

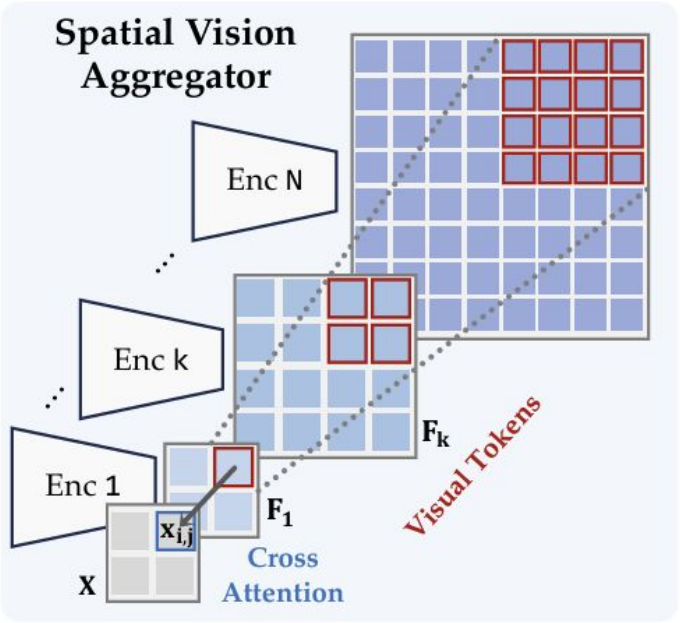
✔ 서로 다른 비전 인코더는 서로 다른 강점을 가진다.

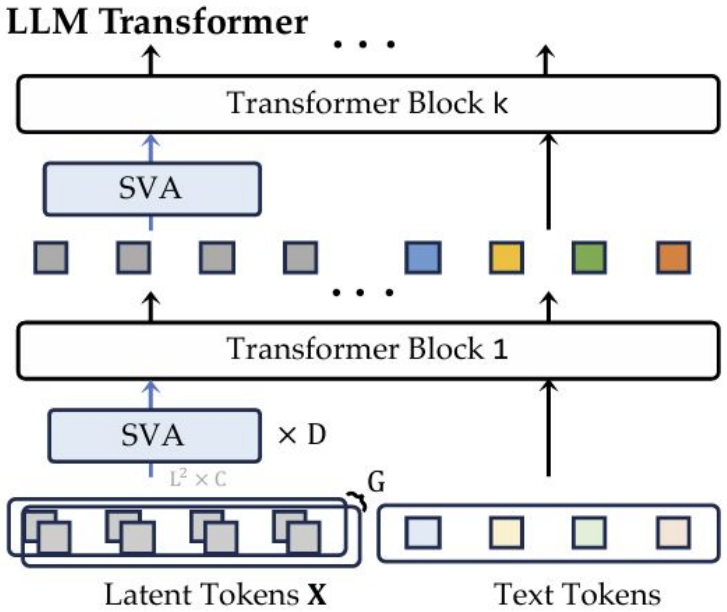
✔ 여러 인코더를 결합하면 성능이 꾸준히 향상되며,

특히 vision-centric task에서 효과가 뚜렷하다.

✔ 단순 결합 방식에는 한계가 있어, 더 정교한 fusion 전략이 필요하다.

4. Spatial Vision Aggregator : A New Connector Design





4. Spatial Vision Aggregator : A New Connector Design

$$\mathbf{q}_{i,j}^{*} = \text{softmax}\left(\frac{\mathbf{q}_{i,j} \cdot [\mathbf{k}_{i,j,1}, \mathbf{k}_{i,j,2}, \ldots, \mathbf{k}_{i,j,N}]^{\top}}{\sqrt{C}}\right)[\mathbf{v}_{i,j,1}, \mathbf{v}_{i,j,2}, \ldots, \mathbf{v}_{i,j,N}] \quad (1)$$

4. Spatial Vision Aggregator : A New Connector Design

Connector	General	Knowledge	OCR & Chart	Vision-Centric
Concat. [127]	67.2	48.9	50.1	52.6
Resampler [59]	63.1	46.5	27.1	42.6
SVA-no-multi-agg	68.0	49.5	55.2	52.6
SVA	**68.5**	**49.7**	**55.5**	**53.2**

Method	General	Knowledge	OCR & Chart	Vision-Centric
Interpolate + MLP	63.4	43.8	28.1	43.7
C-Abstractor [21]	64.4	42.8	26.1	44.3
LDPv2 [35]	62.5	43.9	28.7	43.9
SVA	**65.5**	**44.5**	**31.4**	**46.9**

4. Spatial Vision Aggregator : A New Connector Design

D	OCR & Chart
2	52.1
3	52.4
4	**52.8**

(a) # layers

G	OCR & Chart
1	52.4
2	52.6
3	**53.1**

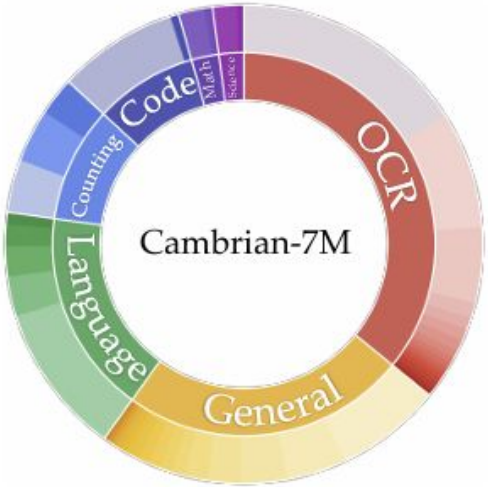
(b) # groups

Multi-agg	OCR & Chart
No	52.4
Yes	**53.3**

(c) Multi-layer aggregation

💡 공간적 **inductive bias**와 **LLM–**시각 피처 간의 깊은 상호작용**(multi-layer aggregation)**은 시각 정보를 더 잘 통합하고 압축하는 데 큰 도움이 된다.

5. Instruction Tuning Data for Training MLLMs



OCR (27.6%)
- Filtered DVQA (1550.0 K)
- DVQA [62] (775.0 K)
- SynthDog [68] (500.0 K)
- ArxivQA [77] (100.0 K)
- OCRVQA [101] (80.0 K)
- ScreenQA [57] (79.0 K)
- WIkiSQL [154] (74.0 K)
- Low-Level Vision [27] (50.0 K)
- DocVQA [98] (39.0 K)
- WTQ [107] (38.0 K)
- ChartQA [97] (28.0 K)
- IconQA [90] (27.0 K)
- Chart2Text [63] (26.0 K)
- TabMWP [89] (23.0 K)
- TextCaps [120] (22.0 K)
- LLAVAR [150] (20.0 K)
- ST-VQA [17] (17.0 K)
- AI2D [66] (15.0 K)
- RenderedText [135] (10.0 K)
- VisText [125] (9.0 K)
- FinQA [31] (6.0 K)
- InfoVQA [16] (2.0 K)
- TAT-QA [159] (2.0 K)
- HiTab [32] (2.0 K)

General (33.3%)
- ALLaVA [25] (700.0 K)
- Q-Instruct [136] (400.0 K)
- LNQA [109] (302.0 K)
- LVIS-Instruct4V [132] (220.0 K)
- LLaVA150K [83] (150.0 K)
- VisualGenome [70] (86.0 K)
- VQAv2 [51] (83.0 K)
- GPT4V Rewritten (77.0 K)
- GQA [58] (72.0 K)
- A-OKVQA [117] (50.0 K)
- AlfWorld [147] (45.0 K)
- ShareGPT [27] (40.0 K)
- RefCOCO [141] (30.0 K)
- VizWiz [52] (20.0 K)
- Visual7W [160] (14.0 K)
- LAION GPT-4V [71] (11.0 K)
- IDK [22] (11.0 K)
- OKVQA [96] (9.0 K)
- HatefulMemes [67] (8.0 K)
- OODVQA [130] (8.0 K)
- SketchyVQA [130] (8.0 K)
- Visualmrc [124] (3.0 K)

Language (23.8%)
- OpenOrca [79] (994.0 K)
- MathInstruct [143] (262.0 K)
- OrcaMath [100] (200.0 K)
- WizardCoder [94] (143.0 K)
- OpenCodeInterpreter [153] (66.0 K)
- Dolly [36] (11.0 K)

Counting (8.5%)
- Filtered CLEVR (350.0 K)
- CLEVR [60] (350.0 K)
- TallyQA [1] (250.0 K)

Code (0.8%)
- Filtered WebSight (790.0 K)
- WebSight [72] (10.0 K)
- DaTikz [14] (47.0 K)
- Design2Code [119] (0.5 K)

Math (3.2%)
- Geo170K [43] (170.0 K)
- RAVEN [149] (42.0 K)
- GeomVerse [65] (9.0 K)
- MathVision [133] (3.0 K)
- Inter-GPS [91] (1.0 K)
- TQA [5] (1.0 K)

Science (2.9%)
- Data Engine (161.0 K)
- PathVQA [54] (32.0 K)
- ScienceQA [92] (12.0 K)

5.1 Data Collection

① 기존 데이터 재활용

- 기존 공개 VQA / OCR / Vision dataset 활용
  (예: PathVQA, DocVQA, GQA, TextVQA 등)
  → **초기 학습용 baseline 확보**

② 자동 생성 기반 데이터 확장 (Internal Pipeline)

- GPT-4 기반 데이터 생성 엔진 구축

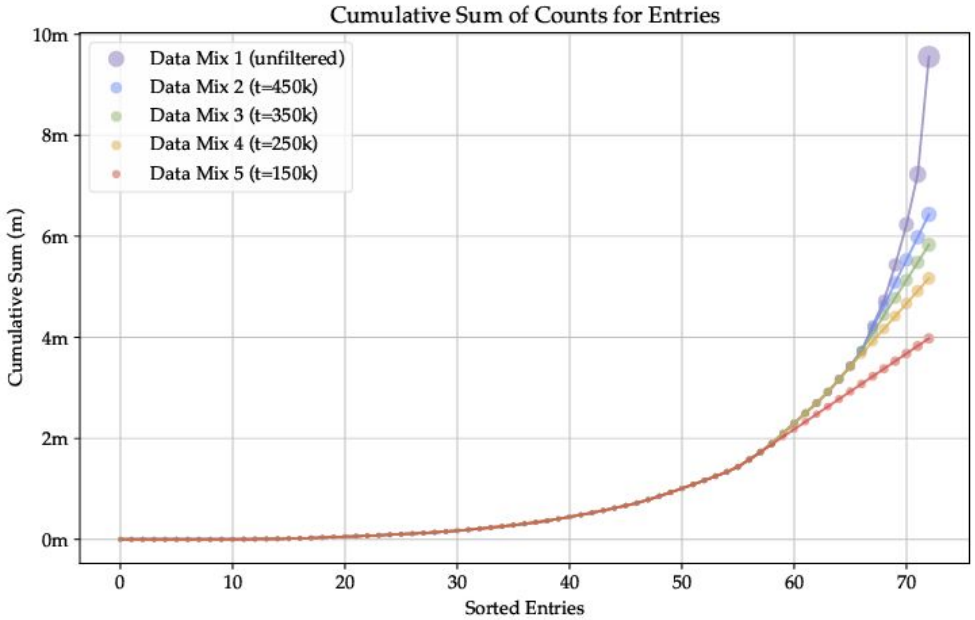
- 프로세스:
  → **(주제 선택 → Wikipedia 검색 → 이미지 매칭 → Q/A 자동 생성)**

- 결과:
  → **과학·지식 기반 시각 질문 161K 생성**
  → 기존 대비 **+400% 확장 효과**

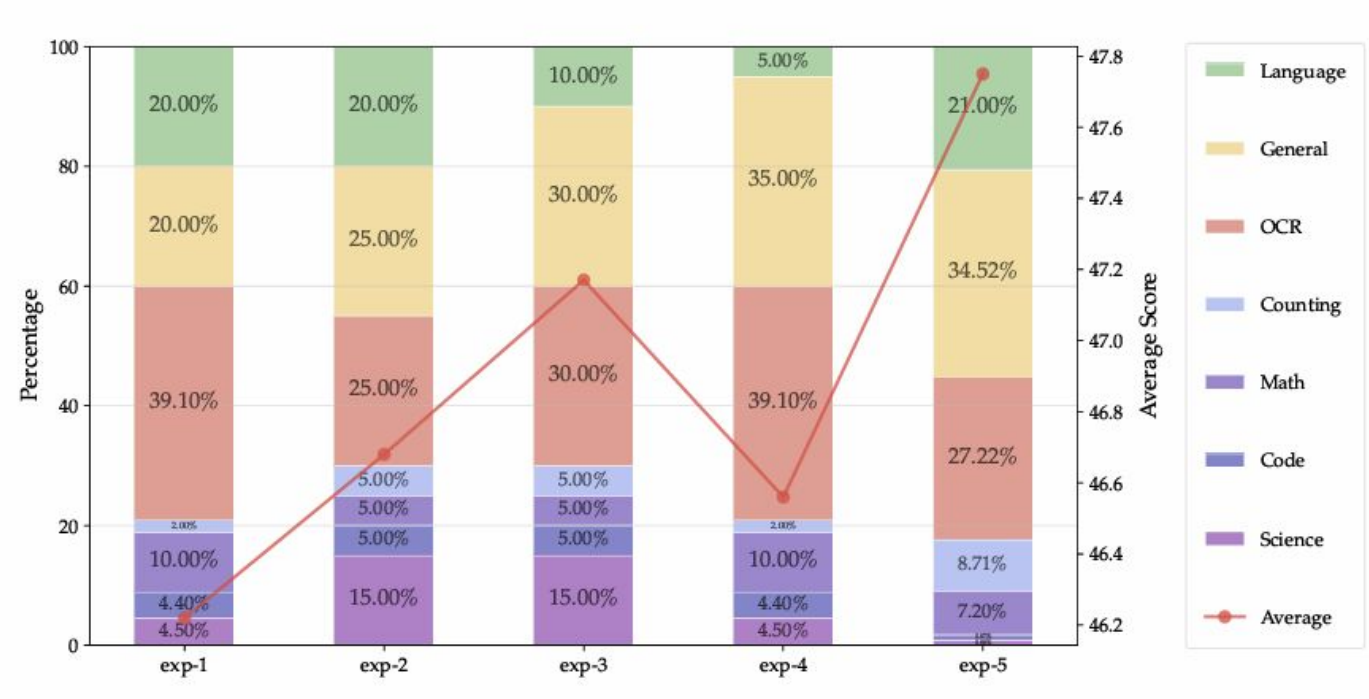
5.2 Data Curation - Data balancing

	Average	General	Knowledge	OCR & Chart	Vision-Centric
150k	53.7	68.0	51.3	45.2	50.5
250k	**54.3**	**68.1**	51.5	45.3	52.2
350k	54.3	67.4	51.4	**46.0**	**52.3**
450k	54.2	68.0	**52.2**	45.5	50.7

5.2 Data Curation - Data Balancing



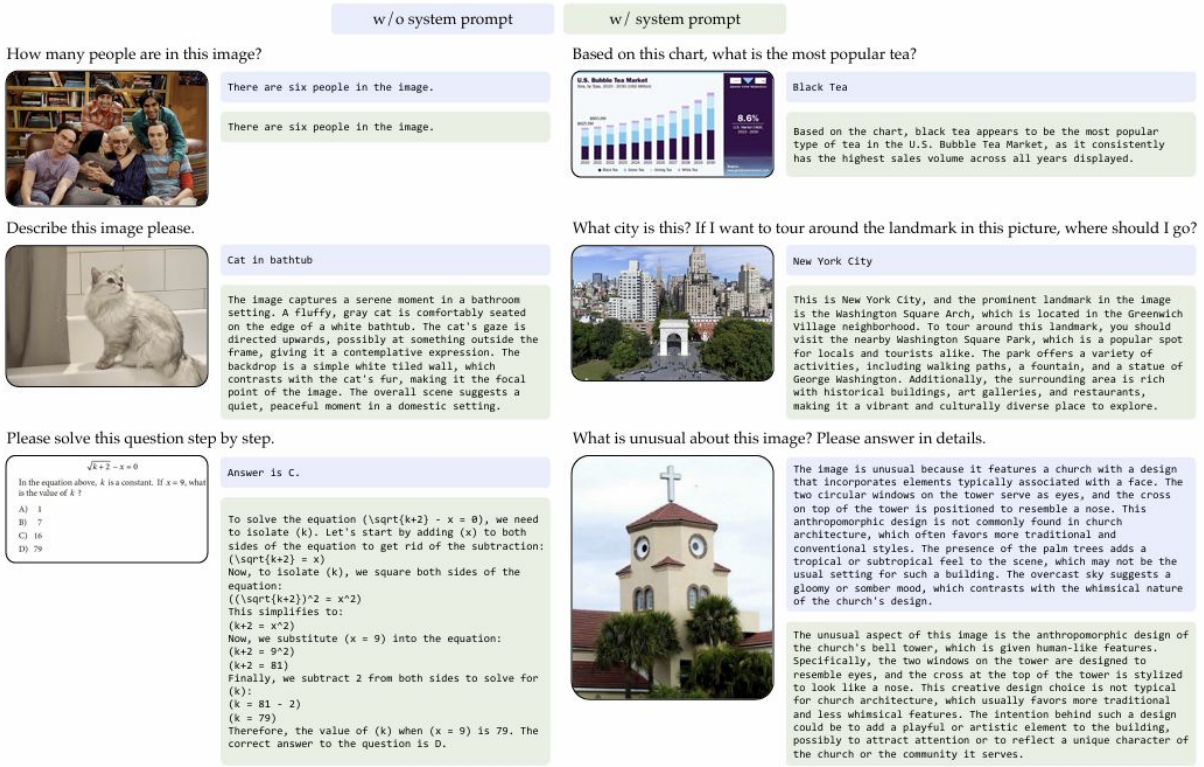
5.2 Data Curation - Data Ratio



5.2 Data Curation - Cambrian-7M

	Average	General	Knowledge	OCR & Chart	Vision-Centric
LLaVA-665K	40.7	64.7	45.2	20.8	32.0
Cambrian-10M	54.8	68.7	51.6	**47.3**	51.4
Cambrian-7M	**55.9**	**69.6**	**52.6**	**47.3**	**54.1**

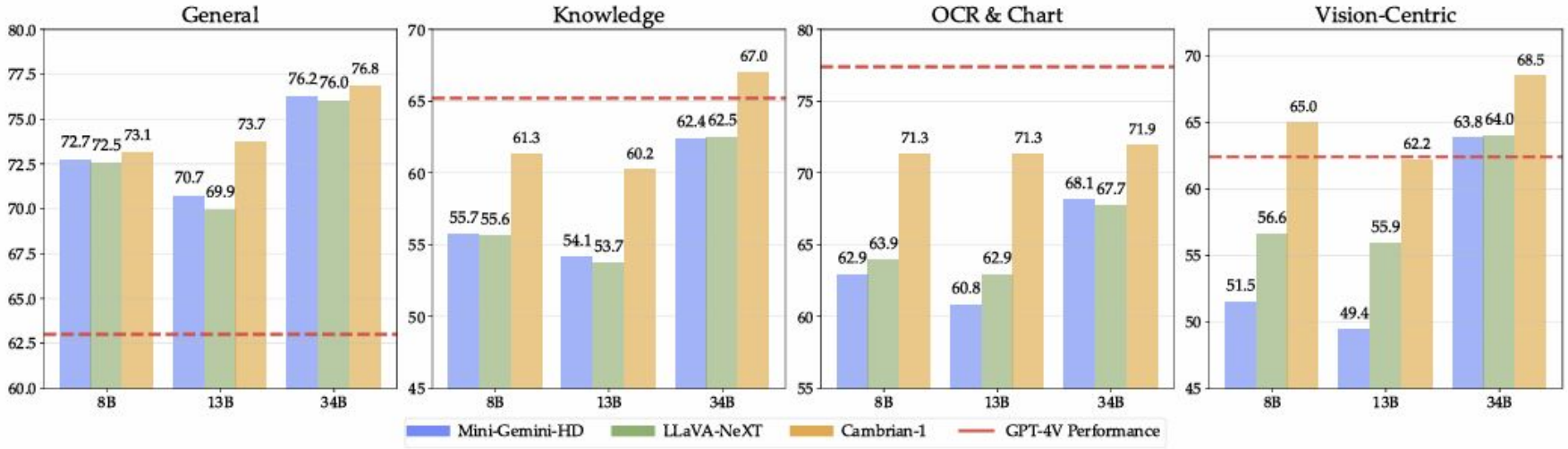
5.3 "Answer Machine Phenomenon" 및 해결



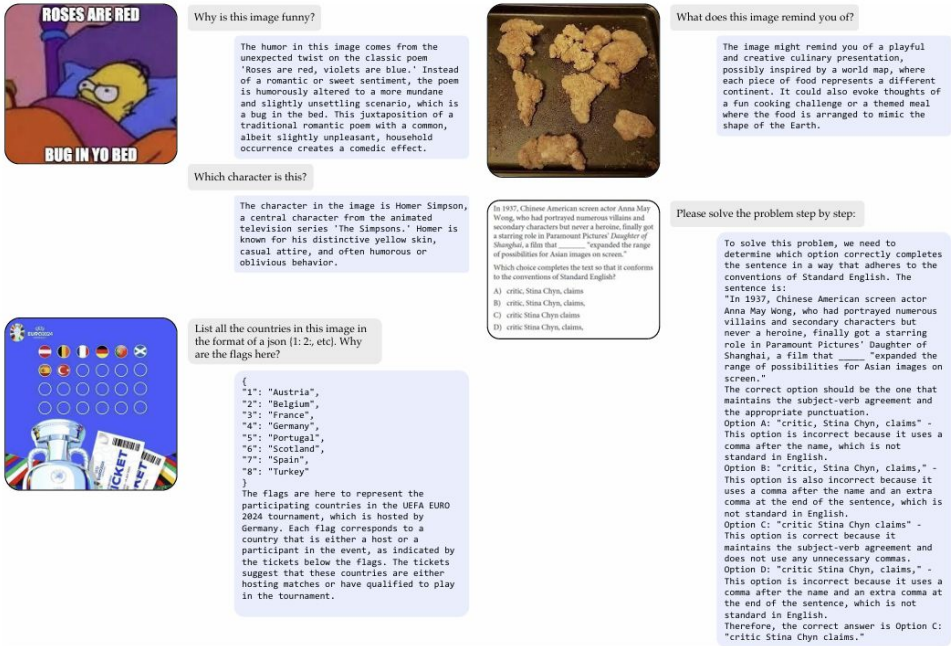
6. State of the Art Performance

Model		General					Knowledge					OCR & Chart					Vision-Centric				
Method	# Vis Tok.	Avg	MME^P	MMB	$SEED^I$	GQA	Avg	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
GPT-4V	UNK.	63.0	1409.4	75.8	69.1	36.8	65.2	75.7	56.8	49.9	78.2	77.4	78.5	64.5	78.0	88.4	62.4	50.0	61.4	64.3	73.8
Gemini-1.0 Pro	UNK.	-	1496.6	73.6	70.7	-	-	79.5	47.9	45.2	-	-	-	65.9	-	-	-	-	-	-	-
Gemini-1.5 Pro	UNK.	-	-	-	-	-	-	-	58.5	52.1	80.3	-	81.3	-	73.5	86.5	-	-	67.5	-	-
Grok-1.5	UNK.	-	-	-	-	-	-	-	53.6	52.8	88.3	-	76.1	-	78.1	85.6	-	-	68.7	-	-
MM-1-8B	144	-	1529.3	72.3	69.9	-	-	72.6	37.0	35.9	-	-	-	-	-	-	-	-	-	-	-
MM-1-30B	144	-	1637.6	75.1	72.1	-	-	81.0	44.7	39.4	-	-	-	-	-	-	-	-	-	-	-
Base LLM: Llama-3-Ins-8B																					
Mini-Gemini-HD-8B	2880	72.7	**1606.0**	72.7	73.2	64.5	55.7	75.1	37.3	37.0	**73.5**	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
LLaVA-NeXT-8B	2880	72.5	1603.7	72.1	72.7	**65.2**	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
Cambrian-1-8B	576	**73.1**	1,547.1	**75.9**	**74.7**	64.6	**61.3**	**80.4**	**42.7**	**49.0**	73.0	**71.3**	**73.3**	**62.4**	**71.7**	**77.8**	**65.0**	**51.3**	**64.2**	**72.3**	**72.0**
Base LLM: Vicuna-1.5-13B																					
Mini-Gemini-HD-13B	2880	70.7	1597.0	68.6	70.6	63.7	54.1	71.9	37.3	37.0	70.1	60.8	56.6	46.6	70.2	69.8	49.4	19.3	57.5	53.6	67.3
LLaVA-NeXT-13B	2880	69.9	1575.0	70.0	65.6	**65.4**	53.7	73.5	36.2	35.1	70.0	62.9	62.2	51.4	67.1	70.9	55.9	36.0	59.1	62.7	65.7
Cambrian-1-13B	576	**73.7**	**1,610.4**	**75.7**	**74.4**	64.3	**60.2**	**79.3**	**40.0**	**48.0**	**73.6**	**71.3**	**73.8**	**61.9**	**72.8**	**76.8**	**62.2**	**41.3**	**63.0**	**72.5**	**71.8**
Base LLM: Hermes2-Yi-34B																					
Mini-Gemini-HD-34B	2880	76.2	1659.0	80.6	75.3	65.8	62.4	77.7	48.0	43.4	**80.5**	68.1	67.6	51.8	74.1	**78.9**	63.8	37.3	67.2	71.5	79.2
LLaVA-NeXT-34B	2880	76.0	1633.2	79.3	**75.9**	**67.1**	62.5	81.8	46.7	46.5	74.9	67.7	68.7	54.5	69.5	78.1	64.0	47.3	61.0	73.0	74.8
Cambrian-1-34B	576	**76.8**	**1689.3**	**81.4**	75.3	65.8	**67.0**	**85.6**	**49.7**	**53.2**	79.7	**71.9**	**75.6**	**60.0**	**76.7**	75.5	**68.5**	**52.7**	**67.8**	**74.0**	**79.7**

6. State of the Art Performance



6. State of the Art Performance



7. Discussion

📌 **What We Observed**

- 기존 벤치마크는 점수는 높지만 실제 시각 이해는 잘 측정하지 못함
- 언어 기반 모델 vs Self-supervised 모델 간 성능 차이 존재
- 단순한 정답 모델은 “Answer Machine”으로 남을 위험

📌 **Key Takeaways**

- 다양한 인코더 조합 + 적절한 데이터 구성 → 격차 해소
- Cambrian-1은 일부 상용 모델 수준까지 도달
- 벤치마크 점수가 전부는 아니다 — **사용 경험과 reasoning이 중요**

📌 **Future Directions**

- Higher resolution handling (Any-Resolution)
- RLHF 기반 alignment

8. Appendix

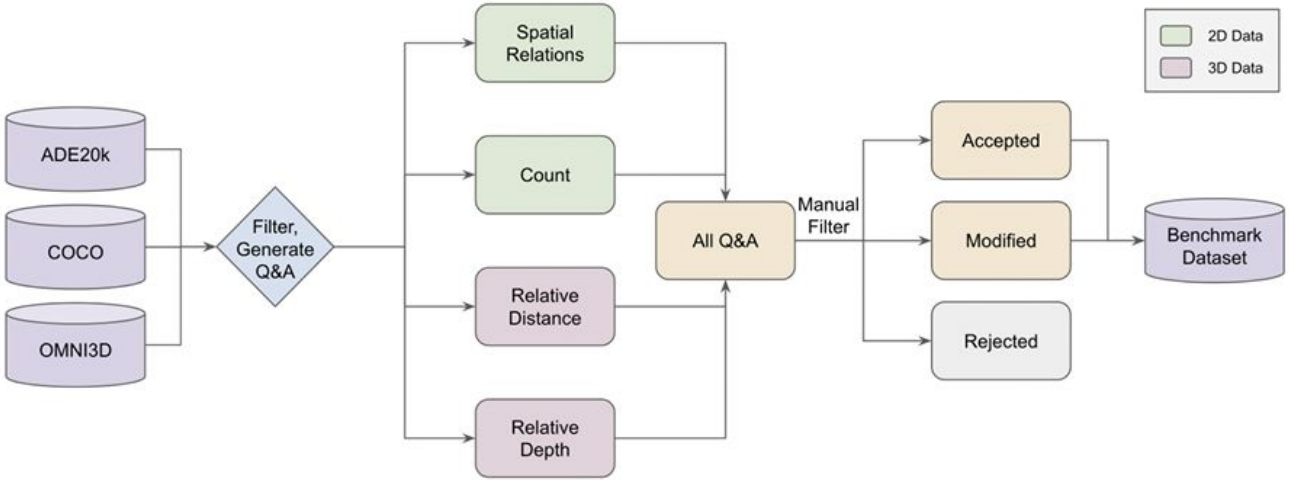


Figure 16 | **CV-CB Benchmark Filtering.** We reformulate classic 2D and 3D CV benchmarks into Q&A questions to evaluate MLLM's visual capabilities.

8. Appendix

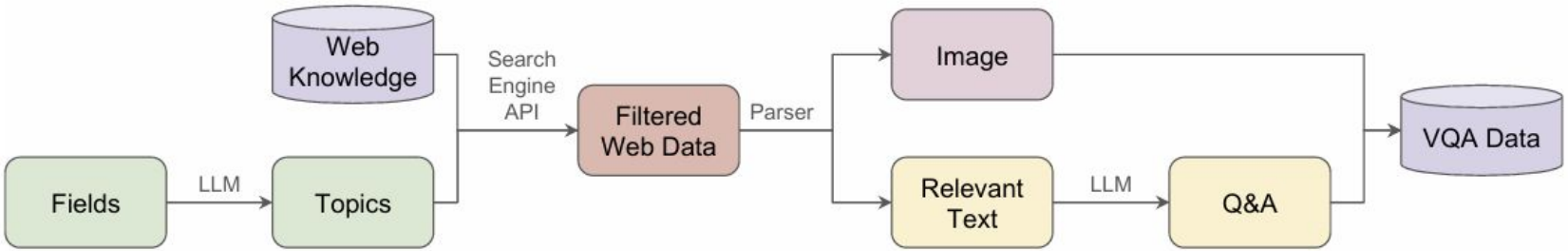


Figure 18 | **Targeted Internet Data Collection Engine.** We build a targeted internet data engine to collect high-quality and large-scale multimodal instruction tuning data for domains like knowledge.

목차

1. Transfusion

1) 멀티모달 생성 모델의 어려움 & 기존 모델들의 해결방법
2) Transfusion이 이 문제를 해결한 방법
3) Transfusion의 Modality에 따른 데이터 처리 방법
4) Transfusion의 손실함수
5) 기존 SOTA와의 비교

2. Gemini 3 Pro

1) 다른 모델들에 대한 차이점
2) Gemini 3 Pro의 핵심원리

3. NanoBanana Pro

1) 다른 모델들에 대한 차이점
2) 여러 예시들

Transfusion

Gemini의 기반이 되는 멀티모달
(텍스트+이미지) 생성 모델



Remove the cupcake on the plate.



Change the tomato on the right to a green olive.



Write the word "Zebra" in Arial bold.



Change this to cartoon style.

멀티모달 생성 모델의 어려움

- 언어 데이터 – **이산적**
  ex) 사과와 바나나의 평균을 구할 수 없음

- 이미지 데이터 – **연속적**
  ex) 연속적인 RGB값으로 이루어져있음

 언어모델(LM)은 다음에 올 단어를 맞추는 Binary Classification 문제를 해결

- Once upon a _____
- That is a picture of a _____
- For breakfast I ate my usual _____

time	0.9
dream	0.03
midnight	0.02
…	
and	1e-100

 이미지 모델(Diffusion)은 노이즈가 낀 신호를 조금씩 수정해가는 문제를 해결

1	2	3	4
3	5	6	3
7	8	9	0
3	1	3	6

[이미지]

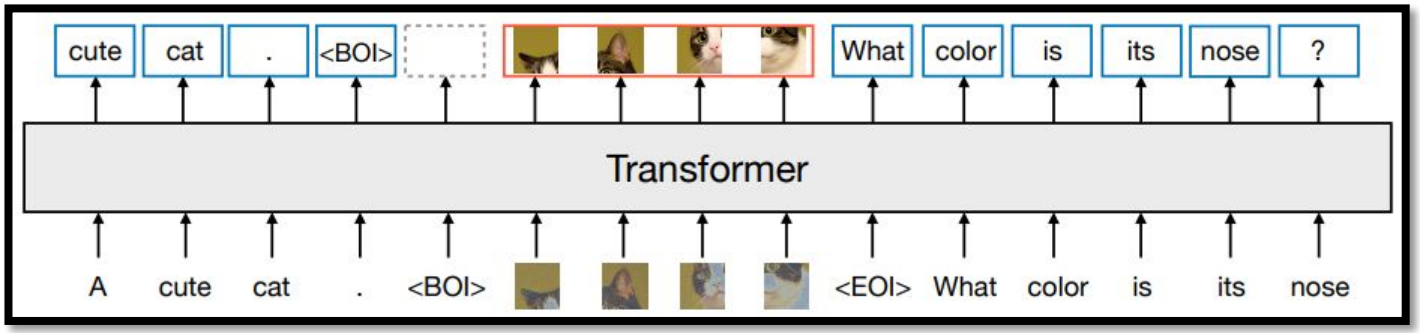
기존 멀티모달 생성 모델이 이 문제를 해결한 방법 (비효과적)

1. **Stitching/Gluing**(따로따로 해서 합치기)
   - Flamingo, LLaVA 등에서 이 방법 사용
   - 이미 학습된 LLM에 이미지 인코더나 디퓨전 모델을 얇은 신경망으로 연결
   - 두 모델은 서로의 언어를 완벽히 이해하지 못해서 비효율적이고 불완전한 학습이 진행됨

3. **Quantization**(억지로 맞추기)
   - Chameleon에서 이 방법을 사용
   - 이미지를 작은 조각으로 나누고, 미리 만들어둔 단어장에서 조각과 가장 비슷한 단어 번호를 매핑
   - 미묘한 색감, 질감 등 세밀한 정보는 사라지고, 모델이 텍스트와 이미지 토큰 사이에서 혼란을 겪어서 텍스트 처리 성능까지 떨어지는 현상 발생
   - 그래도 Transfusion 전까지 SOTA였던 방법론

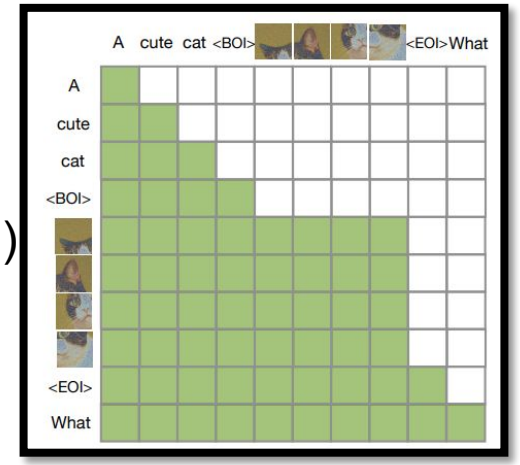
Transfusion이 이 문제를 해결한 방법

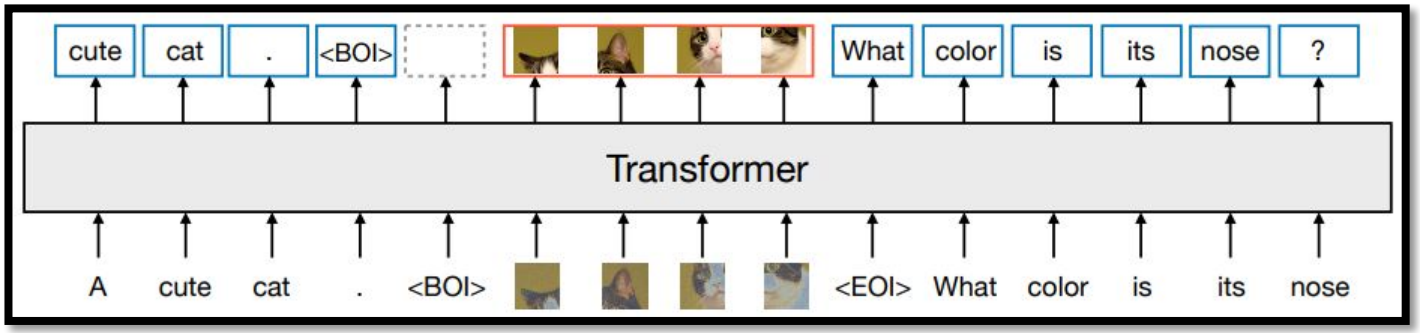
- 이미지를 억지로 이산적으로 변환하지 않고, 모델도 단일 모델을 사용함
- **Transformer 단일 백본** – 텍스트와 이미지 모두를 처리



Transfusion이 이 문제를 해결한 방법

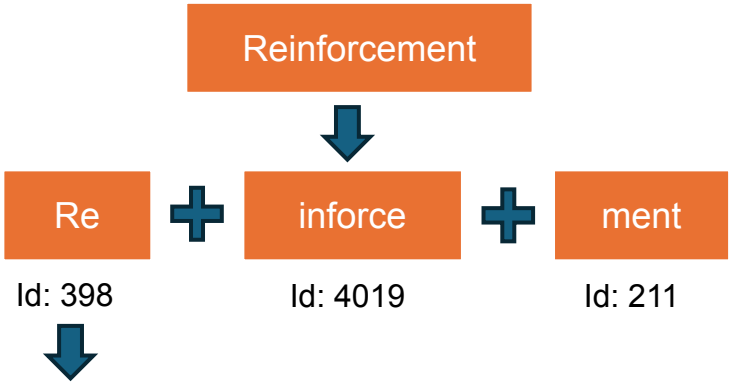
- 텍스트 – **Casual Attention** (이전 토큰만 참조)
- 이미지 – **Bidirectional Attention** (모든 토큰, 패치를 참조)





Transfusion의 Modality-Specific Layers - 텍스트

- 텍스트는 기존의 LM과 동일하게 처리
- **토크나이저**를 통해 각 단어를 토큰화하고, 각 토큰을 특정 정수로 변환
- Ex)

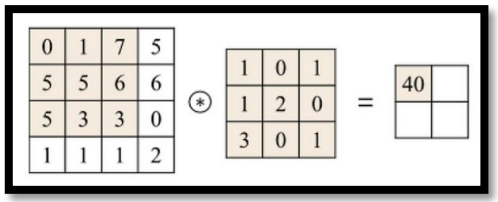


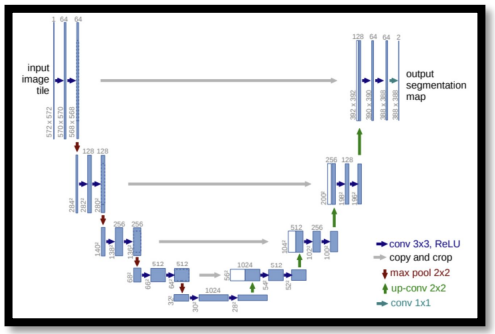
Id에 해당되는 벡터: [0.3, -0.1, …]

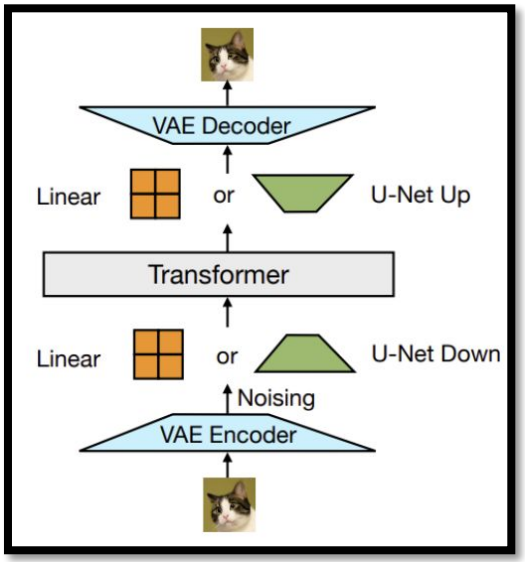
Transformer에 들어감

Transfusion의 Modality-Specific Layers - 이미지

- **VAE**(Variational Autoencoder)를 통한 이미지의 압축 (ex: 256X256 -> 32X32)
- **Patch화** (ex: 32X32 -> 2X2 256개)
- **인코딩/디코딩** - 2가지 방법 비교
  - Linear Layer (단순 선형 변환)
  - **U-Net Block** (핵심 기술!)
- **U-Net Down**
  - Convolutions을 통한 Downsampling
  - ex: 32X32 -> 4X4
- **U-Net UP**
  - Convolution을 통한 Upsampling
  - ex: 1X16 -> 4X4 -> 32X32







이미지 압축 -> 인코딩 -> Patch화 -> Transformer -> 디코딩

Transfusion의 손실함수

- Language Model의 손실함수: Cross Entropy Loss

$$\mathcal{L}_{\text{LM}} = \mathbb{E}_{y_i}\left[-\log P_\theta(y_i|y_{<i})\right]$$

- Diffusion Model의 손실함수: Mean Square Loss

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$$ -> 노이즈를 섞음

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)||^2\right]$$

 모델의 손실함수는 하이퍼파라미터 λ를 통해서 다음과 같이 정의

$$\mathcal{L}_{\text{Transfusion}} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{DDPM}}$$

SOTA(Chameleon 모델)과 비교

- 빨강색: Transfusion, 파랑색: Chameleon

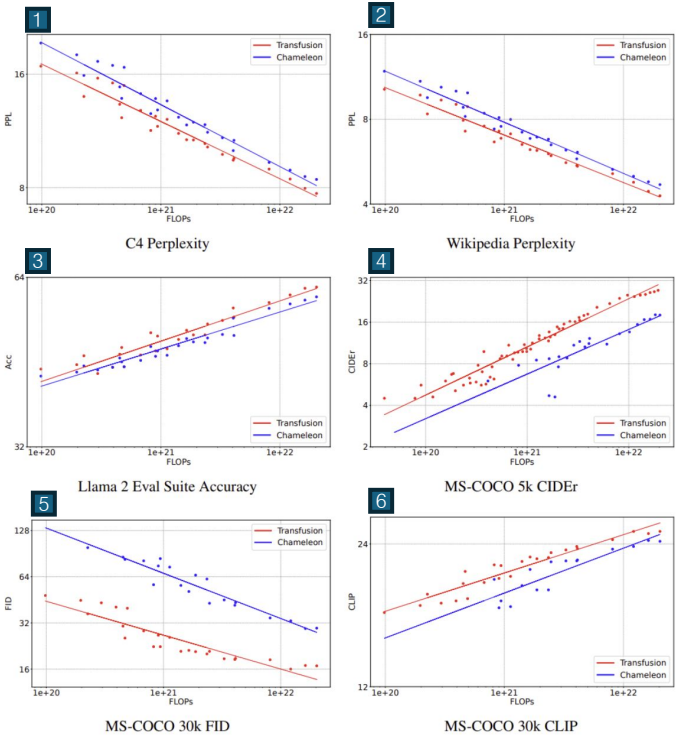
1&2 - Perplexity(텍스트 이해, 낮을 수록 좋음)

3 - Eval Suite Accuracy(텍스트 지능, 높을 수록 좋음)

4 - CIDEr(이미지 이해, 높을 수록 좋음)

5 - FID(가짜 티가 얼마나 나는가, 낮을 수록 좋음)

6 - CLIP Score(텍스트와 이미지의 일치도, 높을 수록 좋음)





Gemini 3 Pro

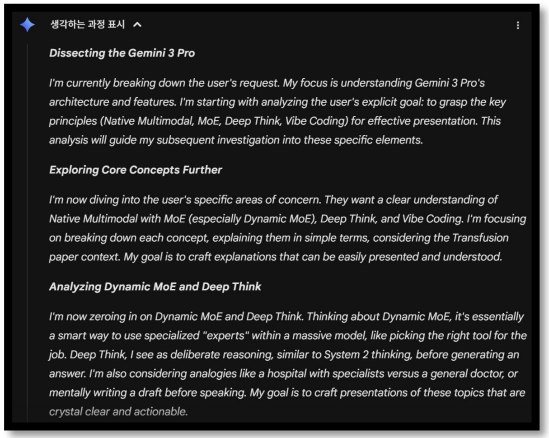
Cambrian-1 + Transfusion + “α”

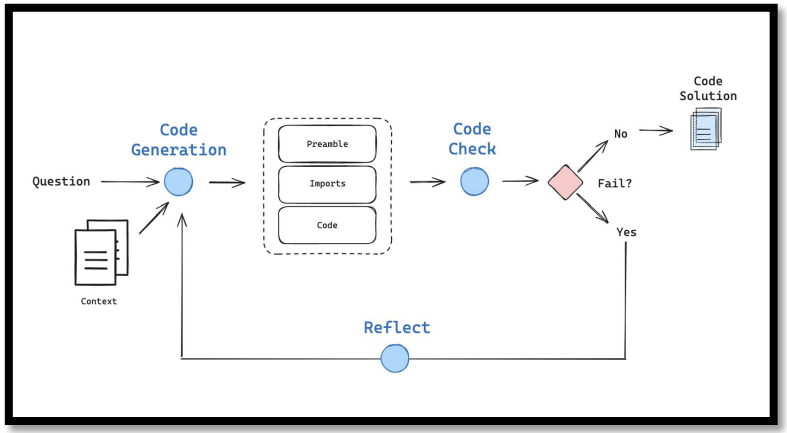
다른 모델들에 대한 Gemini만의 차별점

- **Native Multimodal**: 다른 모델들은 Language Model을 기반으로 이미지도 인식할 수 있게 하는 식이지만, Gemini는 텍스트, 이미지, 오디오를 동시에 학습시킴
- **Long Context Window**: Gemini 1.5 pro 부터 100만(1M) 토큰이라는 파격적인 용량을 선보임
- **Search Grounding**: Gemini 1.5부터 구글 검색엔진을 통해 생성 정보에 대한 팩트체크를 진행함
- **Agentic Capabilities**: 사용자의 의도와 맥락을 깊이있게 파악하는 눈치를 갖춤. Planning, Vibe Coding 등 가능

Gemini 3 pro의 핵심 원리

- **Dynamic MoE(Mixture of Experts) 고도화**: 수학문제, 글쓰기, 코딩 등 다양한 문제에 대해 분야별 전문가를 설정해놔서 문제에 따라 몇 가지 부분 뉴런만 사용하는 매커니즘 고도화
- **Deep Think 기능**: CoT(Chain of Thought)을 통해 답변을 심층 시뮬레이션
- **Vibe Coding 강화**: 추상적인 사용자의 요구에 대해 의도를 파악하고, 코드를 적절히 작성한 후, 코드를 직접 실행해보고 에러가 나면 스스로 고치는 기능 강화





NanoBanana Pro

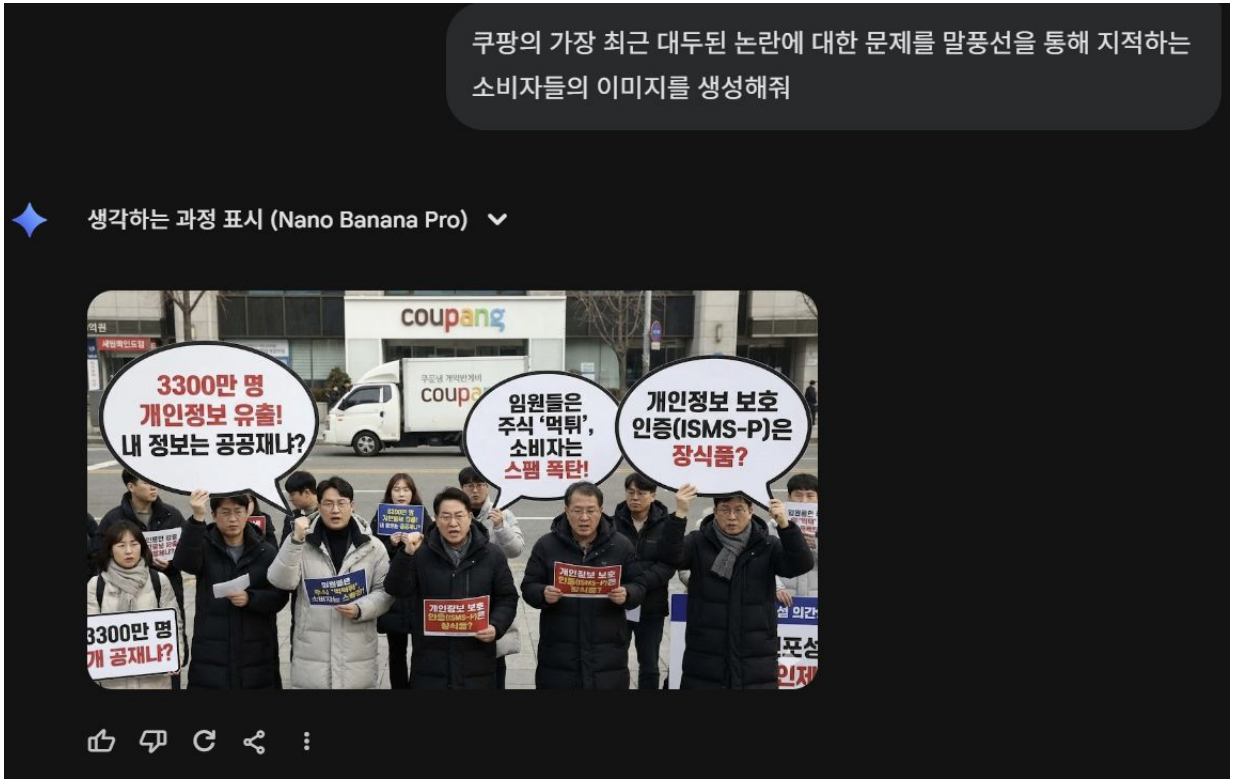


다른 모델들에 대한 NanoBanana Pro만의 차별점

- **실시간 Google Search**: 학습 데이터셋에 의존하는 다른 모델들과 다르게 실시간으로 구글 검색을 거쳐서 이미지를 생성
- **텍스트 처리 기능**: Text & Layout Rendering Engine을 통해 텍스트를 별도로 처리해서 오타가 거의 나지 않고, 다양한 폰트/스타일을 지원
- **추론 능력**: 복잡한 논리/공간 추론 가능
- **생성 프로세스**: 그냥 생성하는(Black Box) 다른 모델들과 다르게 생각과 계획을 거치고 이미지를 생성

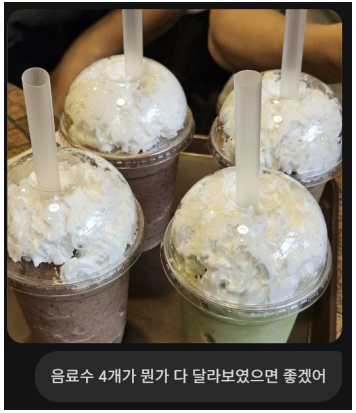
 이미지의 질이 훨씬 좋아졌을 뿐만 아니라 마케팅 포스터, 인포그래픽, UI 디자인 등에도 폭넓게 사용할 수 있게 됨

실시간 Google Search & 텍스트 처리 기능 예시

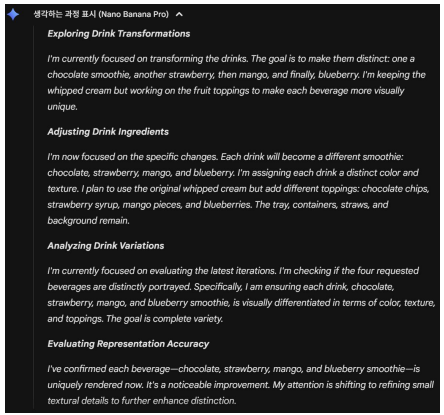


추론 능력 & 생성 프로세스 예시

입력

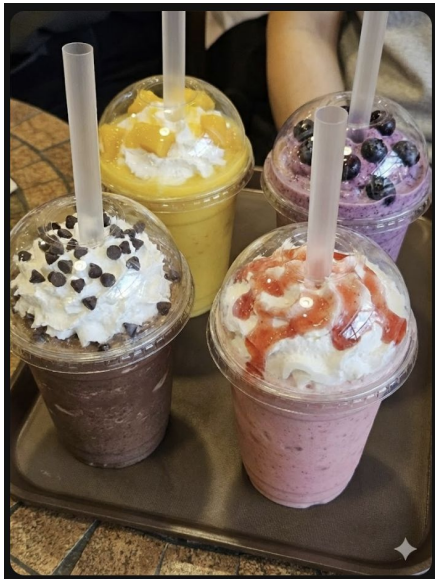


추론



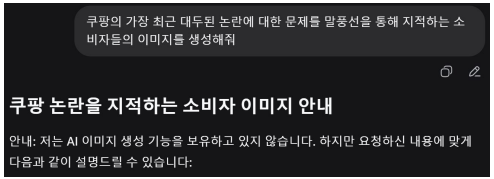
- **음료 변형 살펴보기**
- 하나는 초콜릿 스무디, 다른 하나는 딸기, 그다음은 망고, 마지막으로 블루베리입니다. 휘핑크림은 그대로 두고 과일 토핑을 추가하여 각 음료가 시각적으로 더욱 독특하게 보이도록 작업하고 있습니다.
- **음료 재료 조정**
- 각 음료에 고유한 색상과 질감을 부여하고 있습니다. 기존 휘핑크림을 사용하되 초콜릿 칩, 딸기 시럽, 망고 조각, 블루베리 등 다른 토핑을 추가할 계획입니다. 트레이, 용기, 빨대, 배경은 그대로 유지됩니다.
- **음료 변형 분석**
- 초콜릿, 딸기, 망고, 블루베리 스무디 등 각 음료가 색상, 질감, 토핑 측면에서 시각적으로 차별화되도록 하고 있습니다.
- **표현 정확도 평가**
- 초콜릿, 딸기, 망고, 블루베리 스무디 등 각 음료가 이제 고유하게 표현된다는 것을 확인했습니다. 눈에 띄는 개선입니다. 저는 차별화를 더욱 강화하기 위해 작은 질감 디테일을 다듬는 데 집중하고 있습니다.

출력



경쟁 AI 모델들과 비교

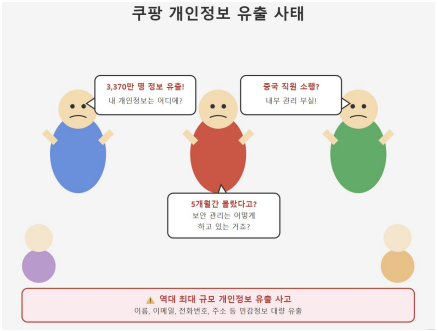
DeepSeek



ChatGPT



Claude – Sonnet 4.5



VS

Gemini - NanoBanana

