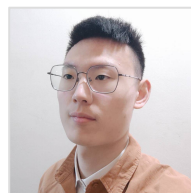


WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences

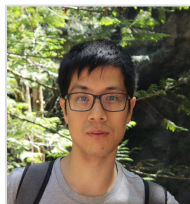
<https://hf.co/spaces/WildVision/vision-arena>



Yujie Lu



Dongfu Jiang



Wenhui Chen



William Wang



Yejin Choi



Bill Yuchen Lin

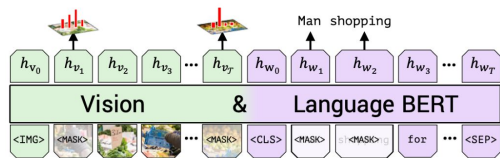


Outline

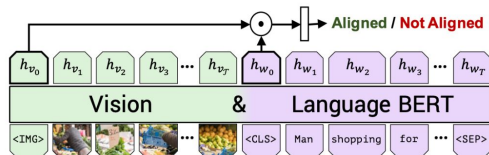
- What is the past/present of VLMs?
- Why is benchmarking these VLMs important and challenging?
 - data-side
 - metric-side
- What we're working on?
 - WildVision-Arena
 - WildVision-Bench
- What are future directions for VLMs and their evaluations?
 - Analyzing model performances from arena user inputs
 - Live benchmarking and leaderboard
- Open-source plan
 - Chat data
 - Feedback data

Evolution of VLMs

Image-Text Fusion

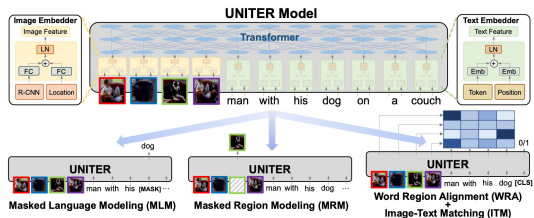


(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

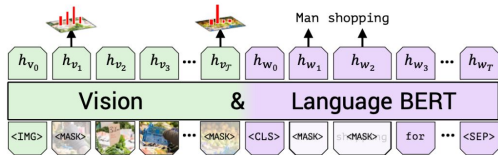
Lu et. al, *ViLBERT*, 2019



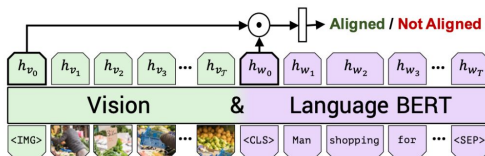
Chen et. al, *UNITER*, 2019

Evolution of VLMs

Image-Text Fusion

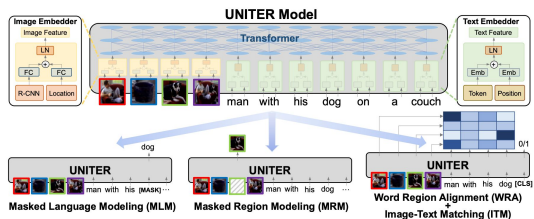


(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

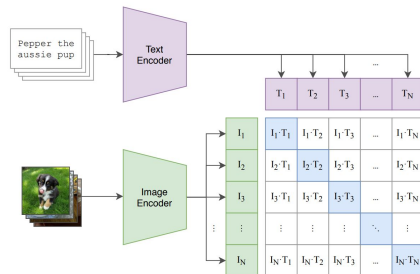
Lu et. al, *ViLBERT*, 2019



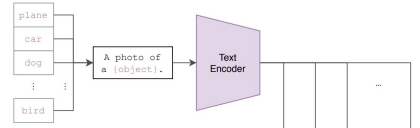
Chen et. al, *UNITER*, 2019

Contrastive Learning

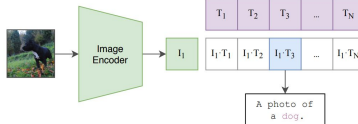
(1) Contrastive pre-training



(2) Create dataset classifier from label text



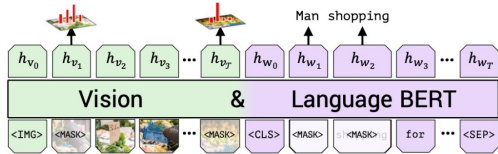
(3) Use for zero-shot prediction



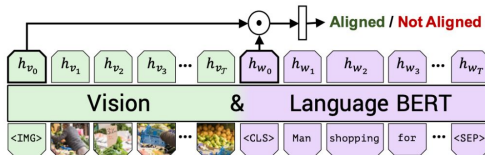
Radford et. al, *CLIP*, 2021

Evolution of VLMs

Image-Text Fusion

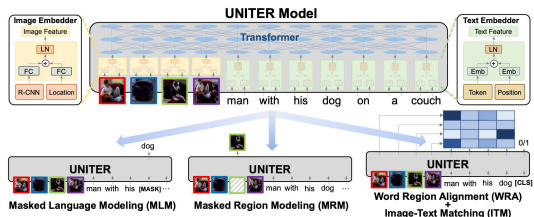


(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

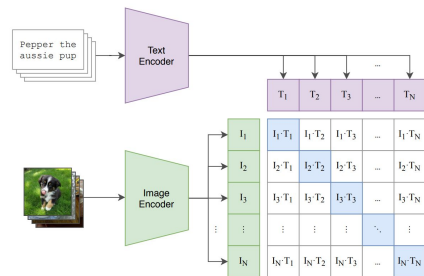
Lu et. al, ViLBERT, 2019



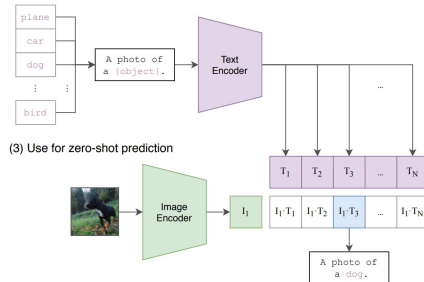
Chen et. al, UNITER, 2019

Contrastive Learning

(1) Contrastive pre-training



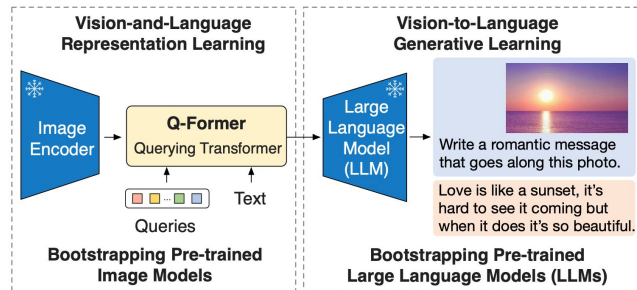
(2) Create dataset classifier from label text



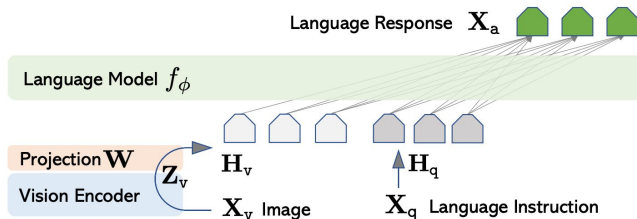
(3) Use for zero-shot prediction

Radford et. al, CLIP, 2021

Large Language Model

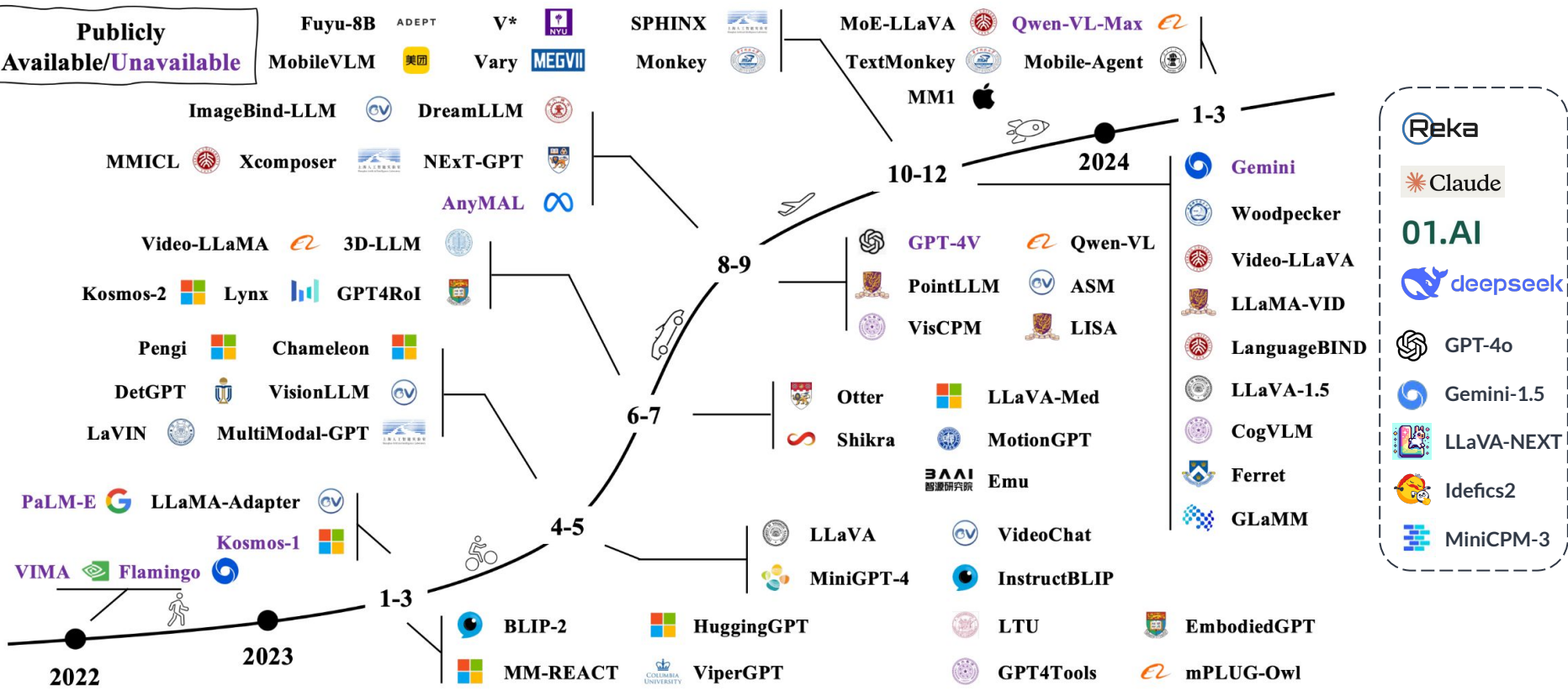


Li et. al, BLIP-2, 2023



Liu et. al, LLaVA, 2023

Publicly Available/Unavailable



Recent Timeline of VLMs

**How do we evaluate the
VLMs?**

A Brief History of VLMs Evaluation

- Task-Specific

- Visual Question Answering

- VQA $\text{Acc}(ans) = \min \left\{ \frac{\#\text{humans that said } ans}{3}, 1 \right\}$

- Grounding Referring Expressions

- RefCOCO $\text{IoU} = \frac{\text{Area}_{\text{intersection}}}{\text{Area}_{\text{union}}}$

- Image Captioning

- NoCaps $\text{CIDEr}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i)$

- Text-Image Retrieval

- Flickr $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

Dataset

VQA v2



Q: Is the boy happy?
GT: Yes

COCO



GT: The man at bat readies to swing at the pitch while the umpire looks on.

Text VQA



Q: What is the largest denomination on table?
GT: 500

Dataset

VQA v2



Q: Is the boy happy?
GT: Yes

Required capability:
Recognition

COCO



GT: The man at bat readies to swing at the pitch while the umpire looks on.

Required capabilities:
Recognition
Language generation

Text VQA



Q: What is the largest denomination on table?
GT: 500

Required capabilities:
Recognition
OCR

Dataset

VQA v2



Q: Is the boy happy?
GT: Yes

Required capability:
Recognition

COCO



GT: The man at bat readies to swing at the pitch while the umpire looks on.

Required capabilities:
Recognition
Language generation

Text VQA



Q: What is the largest denomination on table?
GT: 500

Required capabilities:
Recognition
OCR



Q: What will the girl on the right write on the board?
GT: 14

Required capabilities:
Recognition
Spatial awareness
OCR
Math



Q: Where is this photo taken?
GT: Qatar

Required capabilities:
Recognition
Knowledge

Me: I'll do it at 8
Time: 8.05
Me: looks like I gotta wait till 9 now



Q: Can you explain this meme?
GT: This meme is a humorous take on procrastination and the tendency to delay tasks until a specific time ...

Required capabilities:
OCR
Recognition
Knowledge
Language generation

Dataset



The original VL problem:

Q: How many apples are there in the image?

A. 4; B. 3; C. 2; D. 1

GT: A

Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):

1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓

2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓

3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗

4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

Dataset



The original VL problem:

Q: How many apples are there in the image?

A. 4; B. 3; C. 2; D. 1

GT: A

Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):

1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓

2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓

3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗

4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third <image 1></p> <p>(B) Diminished fifth <image 2></p> <p>(C) Minor seventh <image 3></p> <p>(D) Diminished sixth <image 4></p>	<p>Question: ...The graph shown is compiled from data collected by Gallup <image 1>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p>(C) 0.3571 (D) 0.5</p>	<p>Question: <image 1> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^1 [f(x) - g(x)] dx$</p> <p>(B) $\int_0^1 [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p>
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact</p> <p>(B) Hematoma</p> <p>(C) Fat necrosis (D) Silicone granuloma</p>	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <image 1></p> <p>Option:</p> <p>(A) Oppressor</p> <p>(B) Imperialist</p> <p>(C) Savior (D) Isolationist</p>	<p>Question: Find the VCE for the circuit shown in <image 1>. Neglect VBE</p> <p>Answer: 3.75</p> <p>Explanation: ...IE = [(V_{EE}) / (R_E)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - I_ER_L = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V</p>
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Dataset



The original VL problem:

Q: How many apples are there in the image?
A. 4; B. 3; C. 2; D. 1 GT: A

Circular Evaluation

4 Passes in Circular Evaluation (choices with circular shift):

- Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
- Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
- Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
- Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third <image 1></p> <p>(B) Diminished fifth <image 2></p> <p>(C) Minor seventh <image 3></p> <p>(D) Diminished sixth <image 4></p>	<p>Question: ...The graph shown is compiled from data collected by Gallup <image 1>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p>(C) 0.3571 (D) 0.5</p>	<p>Question: <image 1> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^1 [f(x) - g(x)] dx$</p> <p>(B) $\int_0^1 [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p>
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact</p> <p>(B) Hematoma</p> <p>(C) Fat necrosis (D) Silicone granuloma</p>	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <image 1></p> <p>Option:</p> <p>(A) Oppressor</p> <p>(B) Imperialist</p> <p>(C) Savior (D) Isolationist</p>	<p>Question: Find the VCE for the circuit shown in <image 1>. Neglect VBE</p> <p>Answer: 3.75</p> <p>Explanation: ...$I_E = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_{ERL} = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$</p>
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Question: Find the missing value in this math puzzle.

Solution:

$$(5 - 4)^3 = 1$$

$$(7 - 3)^3 = 64$$

$$(8 - 2)^3 = 216$$

Similarly, $(11 - 8)^3 = 27$.

So the missing value is 27.

Answer: 27

Category: Math-targeted
Task: Figure question answering
Context: Puzzle test
Grade: Elementary school
Math: Logical reasoning

(a) IQTest

Question: Which function is monotonic in range $[0, \pi]$?

Choices:

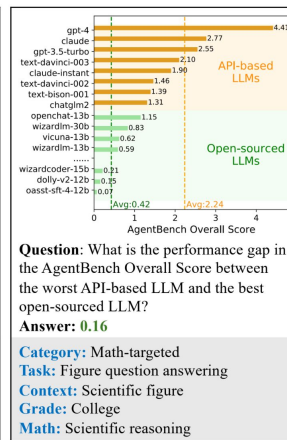
(A) the red one (B) the blue one

(C) both (D) none of them

Answer: (B) the blue one

Category: Math-targeted
Task: Textbook question answering
Context: Function plot
Grade: College
Math: Algebraic reasoning

(b) FunctionQA



(c) PaperQA

Liu et al. MMBench: Is Your Multi-modal Model an All-around Player? 2023

Yue et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. 2023

Lu et al. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. 2023

A Brief History of VLMs Evaluation

- Task-Specific

- Visual Question Answering

- VQA $\text{Acc}(ans) = \min \left\{ \frac{\#\text{humans that said } ans}{3}, 1 \right\}$

- Grounding Referring Expressions

- RefCOCO $\text{IoU} = \frac{\text{Area}_{\text{intersection}}}{\text{Area}_{\text{union}}}$

- Image Captioning

- NoCaps $\text{CIDEr}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i)$

- Text-Image Retrieval

- Flickr $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

- Comprehensive Task

- MME, MMBench, MMVet, MMMU ...

- Multiple choice question answering: Extract (optional), Exact Matching, Accuracy

- Open-ended question answering: using LLM or VLM for single answer scoring/pairwise voting

A Brief History of VLMs Evaluation

- Task-Specific

- Visual Question Answering

- VQA $\text{Acc}(ans) = \min \left\{ \frac{\#\text{humans that said } ans}{3}, 1 \right\}$

- Grounding Referring Expressions

- RefCOCO $\text{IoU} = \frac{\text{Area}_{\text{intersection}}}{\text{Area}_{\text{union}}}$

- Image Captioning

- NoCaps $\text{CIDEr}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i)$

- Text-Image Retrieval

- Flickr $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$

- Comprehensive Task

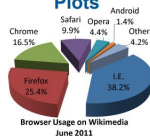
- MME, MMBench, MMVet, MMMU ...
- Multiple choice answer: Extract (optional), Options Mapping, Accuracy
- Open-ended: using LLM or VLM for single answer scoring/pairwise voting

- Real World Task

- VisIT-Bench, RealWorldQA, ...

Dataset

Reasoning Over Plots



How much is the browser usage for Firefox and Safari?

Art Knowledge



Teach me about this painting.

Recognition



Where is this?

Location Understanding



If you are going for a picnic at this location, what items should you carry with you?

Home Renovation



Here is a photo of my bathroom. How can I design it nicer?

Contextual Knowledge of Events



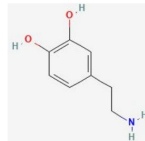
Tell me what is notable or important about the event in this photo.

Figurative Speech Explanation



Someone said that this man is an angel. Why?

Chemical Identification



Which chemical compound does this image represent?

Hazard Identification



If you are driving and come across this scenario, what should you do?

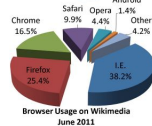
Game Playing



What is the poker hand shown in the picture? Is this a good hand?

Dataset

Reasoning Over Plots



How much is the browser usage for Firefox and Safari?

Art Knowledge



Teach me about this painting.

Recognition



Where is this?

Location Understanding



If you are going for a picnic at this location, what items should you carry with you?

Home Renovation



Here is a photo of my bathroom. How can I design it nicer?

Contextual Knowledge of Events



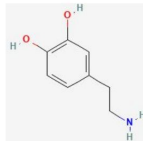
Tell me what is notable or important about the event in this photo.

Figurative Speech Explanation



Someone said that this man is an angel. Why?

Chemical Identification



Which chemical compound does this image represent?

Hazard Identification



If you are driving and come across this scenario, what should you do?

Game Playing



What is the poker hand shown in the picture? Is this a good hand?



Which object is larger the pizza cutter or the scissors? A. The pizza cutter is larger. B. The scissors is larger. **C. They are about the same size.**



Where can we go from the current lane? **A. Turn left.** B. Go straight. C. Turn left and go straight. D. Turn right.



Given this front camera view from our sedan, do we have enough space to drive around the gray car in front of us? **A. Yes.** B. No.



Given the picture, in which cardinal direction is the dinosaur facing? A. North. B. South. **C. East.** D. West.

Metric

Reference-based

BLEU

BERTScore

ROUGE

SPICE

METEOR

CIDER

Human-written
reference required

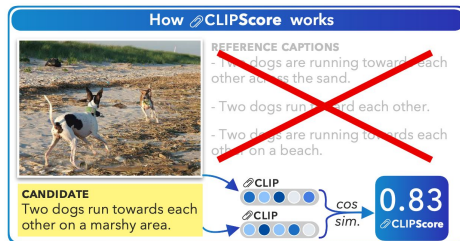
Metric

Reference-based

BLEU
BERTScore
ROUGE
SPICE
METEOR
CIDER

Human-written
reference required

Reference-free



Limited to image-text
matching task

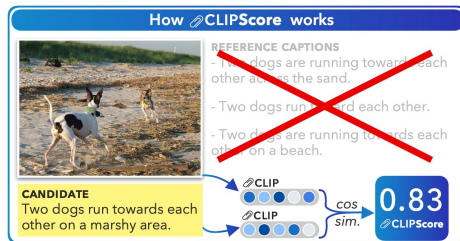
Metric

Reference-based

BLEU
BERTScore
ROUGE
SPICE
METEOR
CIDER

Human-written
reference required

Reference-free



Limited to image-text
matching task

LLM-based

Single-Answer Grading

Input:

A baby is laying down with a teddy bear.

Pred:



Pairwise Comparison

Input:

Pred:

A: A man with a red ...

B: A man with a blue...

Judge
Template



This score is 65
because the image
does not align with
text

A wins because it's
more natural..

Expensive
Consistency
Bias

**What are the limits of current
evals?**

Main Evaluation Results

- Metrics:
 - Avg Score: The average score on all VLM Benchmarks (normalized to 0 - 100, the higher the better).
 - Avg Rank: The average rank on all VLM Benchmarks (the lower the better).
 - Avg Score & Rank are calculated based on selected benchmark. When results for some selected benchmarks are missing, Avg Score / Rank will be None!!!
- By default, we present the overall evaluation results based on 8 VLM benchmarks, sorted by the descending order of Avg Score.
 - The following datasets are included in the main results: MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, A1Z2D, HallusionBench, MMVet.
 - Detailed evaluation results for each dataset (included or not included in main) are provided in the consequent tabs.

Evaluation Dimension

Avg Score
 Avg Rank
 MMBench_V11
 MMStar
 MME
 MMMU_VAL
 MathVista
 OCRBench
 A1Z2D
 HallusionBench
 SEEDBench_IMG
 MMVet

LLaVA_Bench
 CCBench
 RealWorldQA
 POPE
 ScienceQA_TEST

Model Size

<4B
 4B-10B
 10B-20B
 20B-40B
 >40B
 Unknown

Model Type

API
 OpenSource
 Proprietary

Rank	Method	Param (B)	Language Model	Vision Model	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OCR
1	GPT-4o (0513, detail-)				69.9	1.5	82.2	63.9	69.2	61.3	736
2	GPT-4o (0513, detail-)				66.3	3.88	82.8	61.6	62.8	56.5	663
3	Gemini-1.5-Pro				64.4	6	73.9	59.1	60.6	57.7	754
4	GPT-4v (0409, detail-)				63.5	7.25	79.8	56	61.7	54.7	656
5	InternVL-Chat-V1.5	26	InternLM2-20B	InternViT-6B	61.7	7.62	79.7	57.1	46.8	54.7	720
6	GLM-4v				60.8	10.5	78.6	53.2	45.6	45.6	814
7	GPT-4v (0409, detail-)				59.9	12.12	78.8	52.9	62.3	51.4	526
8	Step-1V				59.5	11.75	78	50	49.9	44.8	625
9	InternLM-XComposer2-V	7	InternLM2-7B	CLIP ViT-L/14	58.8	13.38	76.5	55.3	39.7	59.4	675
10	MiniCPM-Llama3-V2.5	8	Llama-3-8B-Instruct	SigLip-400M	58.8	12.75	72	51.8	45.8	54.3	725
11	WeMM	7	InternLM2-7B	SigLip-400M	58.3	13.12	75.7	57	45.3	54.9	628
12	Qwen-VL-Max		QwenLM		58.3	14.5	74.6	49.5	52	43.4	684
13	InternLM-XComposer2-V	7	InternLM2	CLIP ViT-L/14	57.1	15.12	77.6	56.2	41.4	59.5	532

OpenVLM Leaderboard covers

- 73 different VLMs
- 23 multi-modal benchmarks

Results obtained from VLMEvalKit

Open VLM Leaderboard

OpenCompass. A Universal Evaluation Platform for Foundation Models. 2024

OpenVLM Main Leaderboard

[About](#)
[SEEDBench_IMG Leaderboard](#)
[CCBench Leaderboard](#)
[MMBench_TEST_EN Leaderboard](#)
[MMBench_TEST_CN Leaderboard](#)
[MMBench_TEST_EN_V11 Leaderboard](#)
[MMBench_TEST_CN_V11 Leaderboard](#)
[MME Leaderboard](#)
[MMVet Leaderboard](#)
[MMMU_VAL Leaderboard](#)
[MathVista Leaderboard](#)
[HallusionBench Leaderboard](#)
[LLaVABench Leaderboard](#)
[AIZD Leaderboard](#)
[COCO_VAL Leaderboard](#)
[ScienceQA_VAL Leaderboard](#)
[ScienceQA_TEST Leaderboard](#)
[OCRBench Leaderboard](#)
[MMStar Leaderboard](#)
[RealWorldQA Leaderboard](#)
[TextVQA_VAL Leaderboard](#)
[ChartQA_TEST Leaderboard](#)
[OCRQA_TESTCORE Leaderboard](#)
[POPE Leaderboard](#)

→ Fixed Data
 → Not Interactive
 → Contamination Issue

Main Evaluation Results

Metrics:

- Avg Score: The average score on all VLM Benchmarks (normalized to 0 - 100, the higher the better).
- Avg Rank: The average rank on all VLM Benchmarks (the lower the better).
- Avg Score & Rank are calculated based on selected benchmark. When results for some selected benchmarks are missing, Avg Score / Rank will be None!!!

By default, we present the overall evaluation results based on 8 VLM benchmarks, sorted by the descending order of Avg Score.

- The following datasets are included in the main results: MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, AIZD, HallusionBench, MMVet.
- Detailed evaluation results for each dataset (included or not included in main) are provided in the consequent tabs.

Evaluation Dimension

Avg Score
 Avg Rank
 MMBench_V11
 MMStar
 MME
 MMMU_VAL
 MathVista
 OCRBench
 AIZD
 HallusionBench
 SEEDBench_IMG
 MMVet

LLaVABench
 CCBench
 RealWorldQA
 POPE
 ScienceQA_TEST

Model Size

<4B
 4B-10B
 10B-20B
 20B-40B
 >40B
 Unknown

Model Type

API
 OpenSource
 Proprietary

Rank	Method	Param (B)	Language Model	Vision Model	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OCR
1	GPT-4o (0513, detail-)				69.9	1.5	82.2	63.9	69.2	61.3	736
2	GPT-4o (0513, detail-)				66.3	3.88	82.8	61.6	62.8	56.5	663
3	Gemini-1.5-Pro				64.4	6	73.9	59.1	60.6	57.7	754
4	GPT-4v (0409, detail-)				63.5	7.25	79.8	56	61.7	54.7	656
5	InternVL-Chat-V1.5	26	InternLM2-20B	InternViT-6B	61.7	7.62	79.7	57.1	46.8	54.7	720
6	GLM-4v				60.8	10.5	78.6	53.2	45.6	45.6	814
7	GPT-4v (0409, detail-)				59.9	12.12	78.8	52.9	62.3	51.4	526
8	Step-1V				59.5	11.75	78	50	49.9	44.8	625
9	InternLM-XComposer2-V	7	InternLM2-7B	CLIP ViT-L/14	58.8	13.38	76.5	55.3	39.7	59.4	675
10	MiniCPM-Llama3-V2.5	8	Llama-3-8B-Instruct	SigLip-400M	58.8	12.75	72	51.8	45.8	54.3	725
11	WeMM	7	InternLM2-7B	SigLip-400M	58.3	13.12	75.7	57	45.3	54.9	628
12	Qwen-VL-Max		QwenLM		58.3	14.5	74.6	49.5	52	43.4	684
13	InternLM-XComposer2-V	7	InternLM2	CLIP ViT-L/14	57.1	15.12	77.6	56.2	41.4	59.5	532

Open VLM Leaderboard

OpenCompass. A Universal Evaluation Platform for Foundation Models. 2024

Main Evaluation Results

- Metrics:
 - Avg Score: The average score on all VLM Benchmarks (normalized to 0 - 100, the higher the better).
 - Avg Rank: The average rank on all VLM Benchmarks (the lower the better).
 - Avg Score & Rank are calculated based on selected benchmark. When results for some selected benchmarks are missing, Avg Score / Rank will be None!!!
- By default, we present the overall evaluation results based on 8 VLM benchmarks, sorted by the descending order of Avg Score.
 - The following datasets are included in the main results: MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, AIZD, HallusionBench, MMVet.
 - Detailed evaluation results for each dataset (included or not included in main) are provided in the consequent tabs.

Evaluation Dimension

Avg Score
 Avg Rank
 MMBench_V11
 MMStar
 MME
 MMMU_VAL
 MathVista
 OCRBench
 AIZD
 HallusionBench
 SEEDBench_IMG
 MMVet

LLaVABench
 CCBench
 RealWorldQA
 POPE
 ScienceQA_TEST

Model Size

<4B
 4B-10B
 10B-20B
 20B-40B
 >40B
 Unknown

Model Type

API
 OpenSource
 Proprietary

Rank	Method	Param (B)	Language Model	Vision Model	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OCR
1	GPT-4o (0513, detail-				69.9	1.5	82.2	63.9	69.2	61.3	736
2	GPT-4o (0513, detail-				66.3	3.88	82.8	61.6	62.8	56.5	663
3	Gemini-1.5-Pro				64.4	6	73.9	59.1	60.6	57.7	754
4	GPT-4v (0409, detail-				63.5	7.25	79.8	56	61.7	54.7	656
5	InternVL-Chat-V1.5	26	InternLM2-20B	InternViT-6B	61.7	7.62	79.7	57.1	46.8	54.7	720
6	GLM-4v				60.8	10.5	78.6	53.2	45.6	45.6	814
7	GPT-4v (0409, detail-				59.9	12.12	78.8	52.9	62.3	51.4	526
8	Step-1V				59.5	11.75	78	50	49.9	44.8	625
9	InternLM-XComposer2-V	7	InternLM2-7B	CLIP ViT-L/14	58.8	13.38	76.5	55.3	39.7	59.4	675
10	MiniCPM-Llama3-V2.5	8	Llama-3-8B-Instruct	SigLip-400M	58.8	12.75	72	51.8	45.8	54.3	725
11	WeMM	7	InternLM2-7B	SigLip-400M	58.3	13.12	75.7	57	45.3	54.9	628
12	Qwen-VL-Max		QwenLM		58.3	14.5	74.6	49.5	52	43.4	684
13	InternLM-XComposer2-V	7	InternLM2	CLIP ViT-L/14	57.1	15.12	77.6	56.2	41.4	59.5	532

→ Not straightforward to compare VLMs

Open VLM Leaderboard

Main Evaluation Results

- Metrics:
 - Avg Score: The average score on all VLM Benchmarks (normalized to 0 - 100, the higher the better).
 - Avg Rank: The average rank on all VLM Benchmarks (the lower the better).
 - Avg Score & Rank are calculated based on selected benchmark. When results for some selected benchmarks are missing, Avg Score / Rank will be None!!!
- By default, we present the overall evaluation results based on 8 VLM benchmarks, sorted by the descending order of Avg Score.
 - The following datasets are included in the main results: MMBench_V11, MMStar, MMMU_VAL, MathVista, OCRBench, AIZD, HallusionBench, MMVet.
 - Detailed evaluation results for each dataset (included or not included in main) are provided in the consequent tabs.

Evaluation Dimension

- Avg Score
 Avg Rank
 MMBench_V11
 MMStar
 MME
 MMMU_VAL
 MathVista
 OCRBench
 AIZD
 HallusionBench
 SEEDBench_IMG
 MMVet
- LLaVABench
 CCBench
 RealWorldQA
 POPE
 ScienceQA_TEST

Model Size

- <4B
 4B-10B
 10B-20B
 20B-40B
 >40B
 Unknown

Model Type

- API
 OpenSource
 Proprietary

Rank	Method	Param (B)	Language Model	Vision Model	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OCR
1	GPT-4o (0513, detail-)				69.9	1.5	82.2	63.9	69.2	61.3	736
2	GPT-4o (0513, detail-)				66.3	3.88	82.8	61.6	62.8	56.5	663
3	Gemini-1.5-Pro				64.4	6	73.9	59.1	60.6	57.7	754
4	GPT-4v (0409, detail-)				63.5	7.25	79.8	56	61.7	54.7	656
5	InternVL-Chat-V1.5	26	InternLM2-20B	InternViT-6B	61.7	7.62	79.7	57.1	46.8	54.7	720
6	GLM-4v				60.8	10.5	78.6	53.2	45.6	45.6	814
7	GPT-4v (0409, detail-)				59.9	12.12	78.8	52.9	62.3	51.4	526
8	Step-1V				59.5	11.75	78	50	49.9	44.8	625
9	InternLM-XComposer2-V	7	InternLM2-7B	CLIP ViT-L/14	58.8	13.38	76.5	55.3	39.7	59.4	675
10	MiniCPM-Llama3-V2.5	8	Llama-3-8B-Instruct	SigLip-400M	58.8	12.75	72	51.8	45.8	54.3	725
11	WeMM	7	InternLM2-7B	SigLip-400M	58.3	13.12	75.7	57	45.3	54.9	628
12	Qwen-VL-Max		QwenLM		58.3	14.5	74.6	49.5	52	43.4	684
13	InternLM-XComposer2-V	7	InternLM2	CLIP ViT-L/14	57.1	15.12	77.6	56.2	41.4	59.5	532

- No user feedback
- Hard to reflect human preference in real-world
- Not dynamic leaderboard

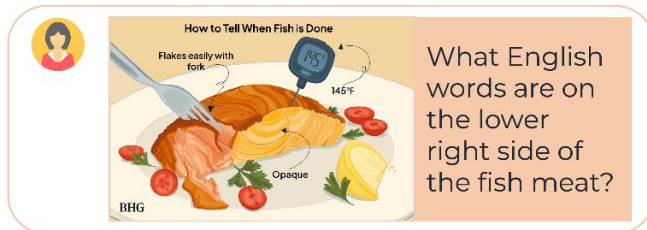
Open VLM Leaderboard

WildVision-Arena

WildVision-Arena



WildVision-Arena



 OpenAI

 Gemini

 Reka

 Claude

 01.AI

 deepseek

 LAVIS

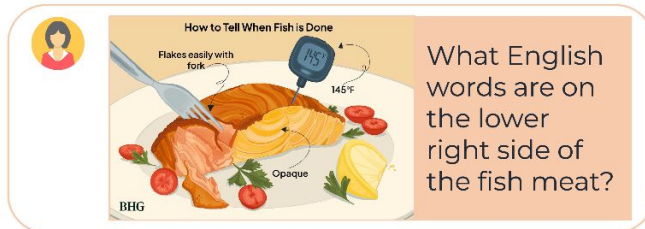
 LLaVA

 Qwen-VL

WildVision-Arena

Rules

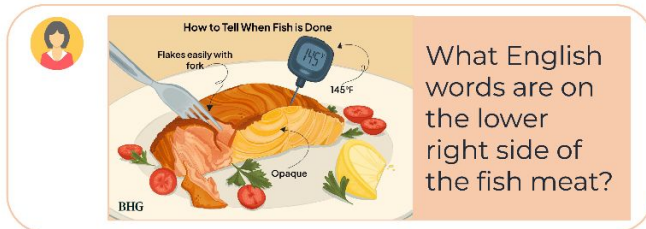
- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



WildVision-Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



OpenAI

Gemini

Reka

Claude

01.AI

deepseek

LAVIS

LLaVA

Qwen-VL

Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.

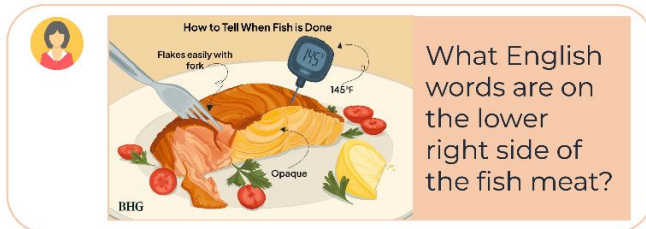
Model B

The English word on the lower right side of the fish meat is "Opaque."

WildVision-Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



 OpenAI

 Gemini

 Reka

 Claude

 01.AI

 deepseek

 LAVIS

 LLaVA

 Qwen-VL

Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.



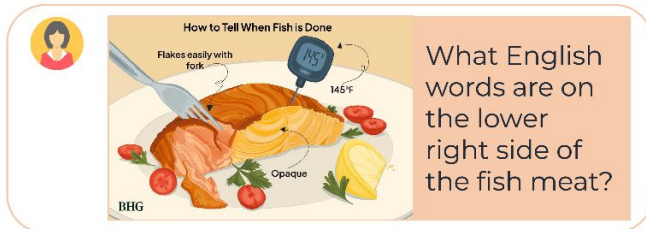
Model B

The English word on the lower right side of the fish meat is "Opaque."

WildVision-Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.



Model B

The English word on the lower right side of the fish meat is "Opaque."

Reason

Both Model A and Model B answer correctly regarding the text.

Vote



A is Better



B is Better



Tie

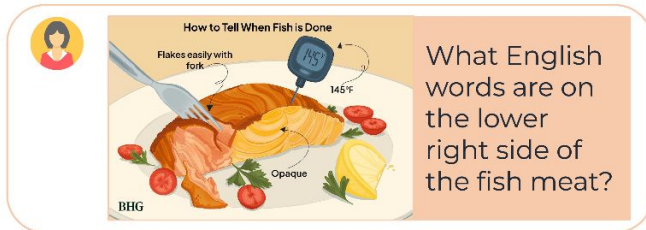


Both are bad

WildVision-Arena

Rules

- Chat with two anonymous models
- Continue to chat until you identify a winner
- Vote for the better one with reason



Model A

On the lower right side of the cooked fish, the word "Opaque" is labeled.



Model B

The English word on the lower right side of the fish meat is "Opaque."

Reason

Both Model A and Model B answer correctly regarding the text.

Vote



A is Better



B is Better



Tie



Both are bad

Model A: Claude-3-Sonnet, Model B: GPT-4V

WVArena Elo Ratings



Submit

Platform

- Model Sample
- Multi-round Multimodal Chat
- User Vote with Reason
- Model Bootstrap and Rank

⌵ Arena 🏆 Leaderboard 📄 About Us

WildVision Arena: Benchmarking Multimodal LLMs in the Wild

by WildVision Team @ AI2 + UCSB + UWaterloo

🚫 Your data will be logged for research purposes. Please do not include any confidential, personal, or other sensitive information. 🚫

🔍 Expand to see all active models

👉 Choose models to sample from


👍 gemini-pro-vision 👍 gemini-1.5-flash-latest 👍 llama-v1.6-34b 👍 llama-v1.6-vicuna-7b 👍 Mistral-CPM-V 👍 coglm-chat-1f 👍 Qwen-VL-Chat

👍 Liform-gen2-qwen-500m 👍 ltray-llava-v1-1f 👍 Bunry-v1.0-3B 👍 claude-3-opus-20240229 👍 claude-3-sonnet-20240229 👍 claude-3-haiku-20240307

👍 deepseek-v1.7b-chat 👍 yi-vl-plus 👍 Reko-Flash 👍 defici2.8b-chatty 👍 gpt-4-vision-preview 👍 gpt-4o

Clear

🖼️ Image 🗨️ Model A 🗨️ Model B



Given my horse's location on this map, what is the quickest route to reach it?

Based on the map, your current location is marked by the pin/stamp icon at Lookout Landing. The quickest route to reach your horse, which is marked with a horse icon at Rito Village, would be to head northwest along the roads/paths shown on the map. It looks like you'll need to traverse the area labeled "Tabantha Frontier" to get from your current spot to Rito Village where your horse is located. The exact path is a bit unclear from this zoomed out view, but generally heading northwest from Lookout Landing should get you to your horse's position at Rito Village by the most direct route possible given the geography shown.

I'm sorry, but I don't have access to your specific map or location information. Could you please provide more context or details so I can assist you better?

Rules

- Upload any image, and send any question to two anonymous models and vote for the better one!
- Or you could also click "Sample Input" to get a random example from public benchmarks such as VisT-Bench.
- Single image multi-round chat is allowed, you can continue to send question until you identify a winner.

Reason for your vote 📄

Please input your reason for response preference here before clicking the model choice button.

👉 A is better 👉 B is better 🏆 Tie 🗳️ Both are bad

🗨️ Your input:

👉 Enter your prompt and press ENTER

Send

📄 Sample Input 🔄 New Round 🔄 Regenerate 🗳️ Share

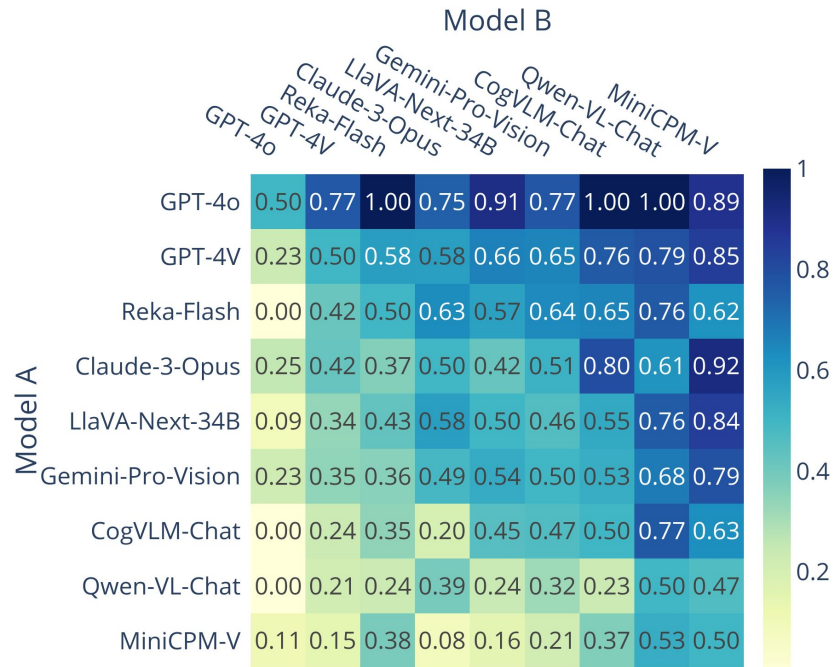
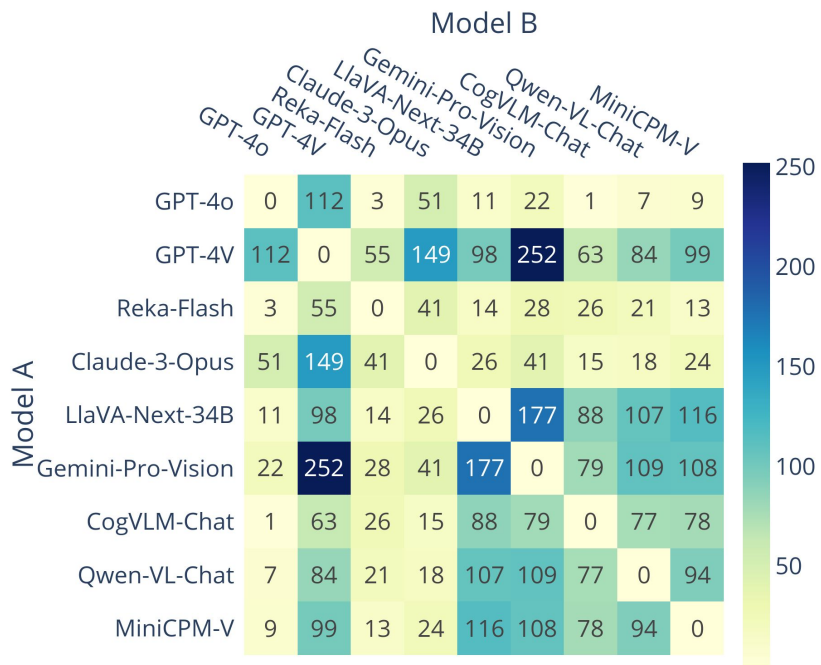
⊖ Parameters

Temperature 0.7

Top P 1

Max output tokens 3024

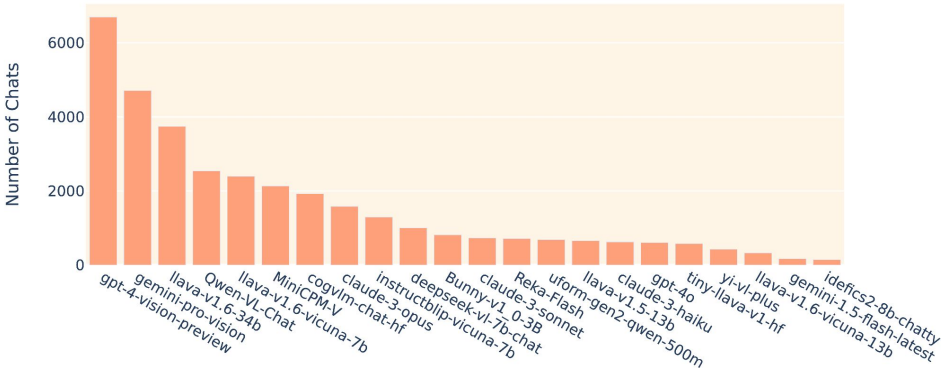
Pairwise Battles between Models



Data Statistics

(date cutoff at May 29, 2024)

- Top-3 Chat Model
 - GPT-4V
 - Gemini-Pro-V
 - LLaVA-v1.6-34B

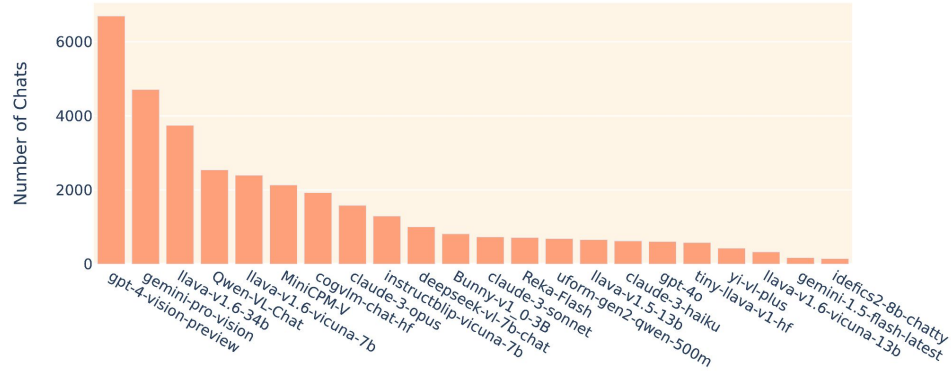


Data Statistics

(date cutoff at May 29, 2024)

- Top-3 Chat Model
 - GPT-4V
 - Gemini-Pro-V
 - LLaVA-v1.6-34B
- 16.4% battles do not get satisfied response

Statistic	Number
Total Votes	8,076
Anonymous	6,636
Non-anonymous	1,440
Left Vote	2,932
Right Vote	2,839
Tie Vote	979
Bad Vote	1,326
Days	102
Total Round	10,884
Avg Round	1.34
Avg Token Input	31.00
Avg Token Output	108.87



Data Statistics

(date cutoff at May 29, 2024)

- Top-3 Chat Model
 - GPT-4V
 - Gemini-Pro-V
 - LLaVA-v1.6-34B
- 16.4% battles do not get satisfied response
- On average, 71 votes per day

Statistic	Number
Total Votes	8,076
Anonymous	6,636
Non-anonymous	1,440
Left Vote	2,932
Right Vote	2,839
Tie Vote	979
Bad Vote	1,326
Days	102
Total Round	10,884
Avg Round	1.34
Avg Token Input	31.00
Avg Token Output	108.87

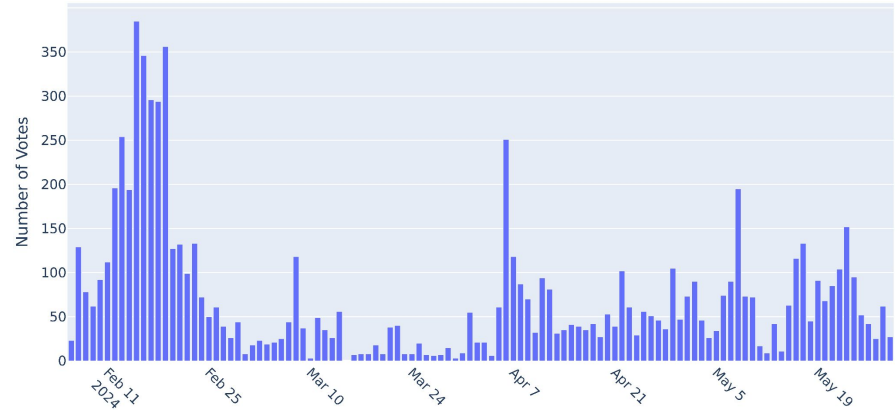
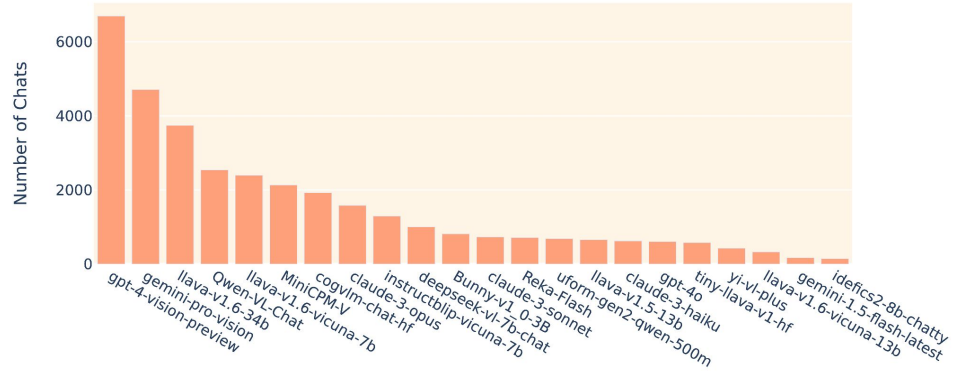


Image [Entertainment-Comics]



[Descriptive-Scene Description] Text Prompt: Whos's in the sky?

Image [People-Portraits]



[Creative-Media Post] Text Prompt: write a social media post with the provided image, saying that I am ready for the new challenge.

Image [Urban-Buildings]



[Recognition-Location] Text Prompt: where is this?

Image [Expert-Science]



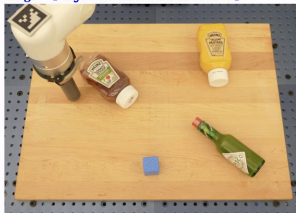
[Analytical-Safety Procedures] Text Prompt: Can you tell me the potential risks and the unreasonable parts in the image?

Image [Natural-Landscapes]



[Recognition-Location] Text Prompt: where was this photo taken?

Image [Objects-Household Tools]



[Descriptive-Object Description] Text Prompt: describe the scene and objects

Image [Entertainment-Web and Mobile Apps Screenshots]



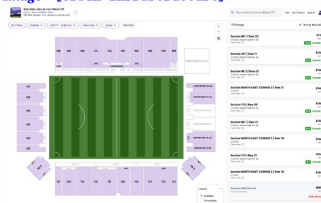
[Interactive-Web Navigation] Text Prompt: I need to download flyer, you will be given screenshot from browser with elements marked with number. give next action to take on web page to download the flyer-give me response in below format example 1 action: [click,scroll,wait], box:1 format action:, box:

Image [Event-Sports]



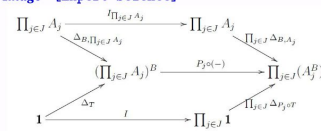
[Descriptive-Scene Description] Text Prompt: this is a football match , every player has an identifier , describe every player action (example : player #501 is running)

Image [Urban-Infrastructure]



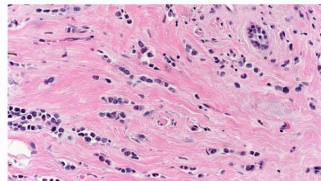
[Interactive-Recommendations] Text Prompt: Which section's ticket would you recommend I purchase?

Image [Expert-Science]



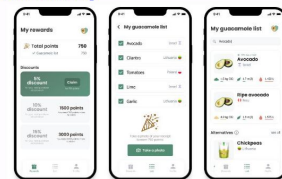
[Interactive-Code Generation] Text Prompt: Give me Latex code to create this diagram

Image [Expert-Health and Medicine]

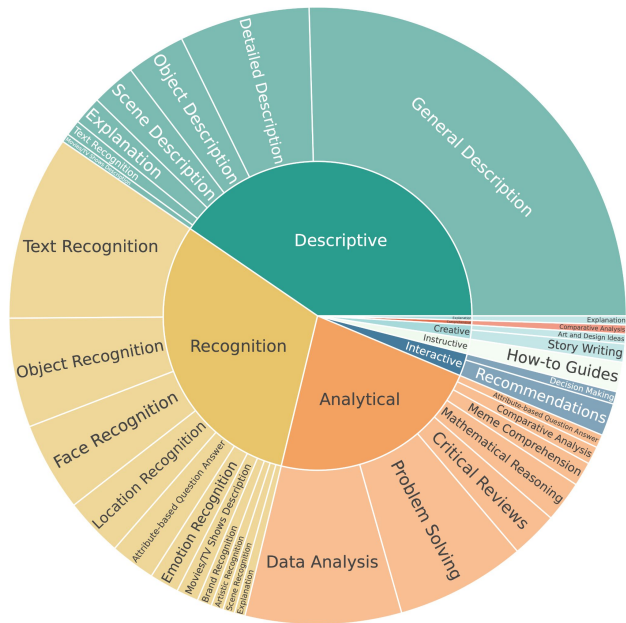


[Recognition-Object] Text Prompt: what type of tumor is this?

Image [Entertainment-Web and Mobile Apps Screenshots]

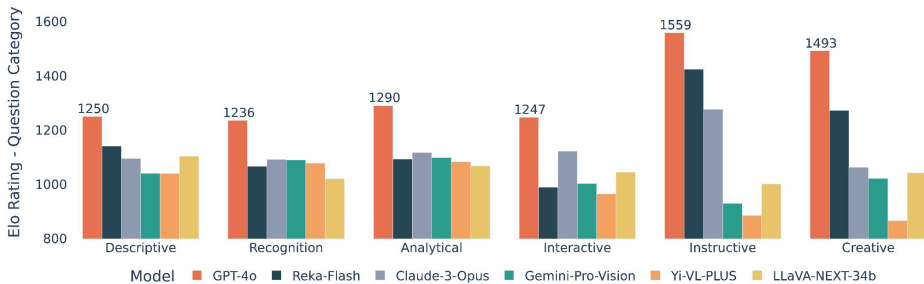


[Analytical-Critical Reviews] Text Prompt: Review each screenshot carefully, focusing on different aspects of usability...



Question Category

- WV-Arena can eval by task types
- GPT-4o consistently outperforms
- Some task types, such as recognition and descriptive are more separative
- Some task types, such as Instructive and creative are less separative



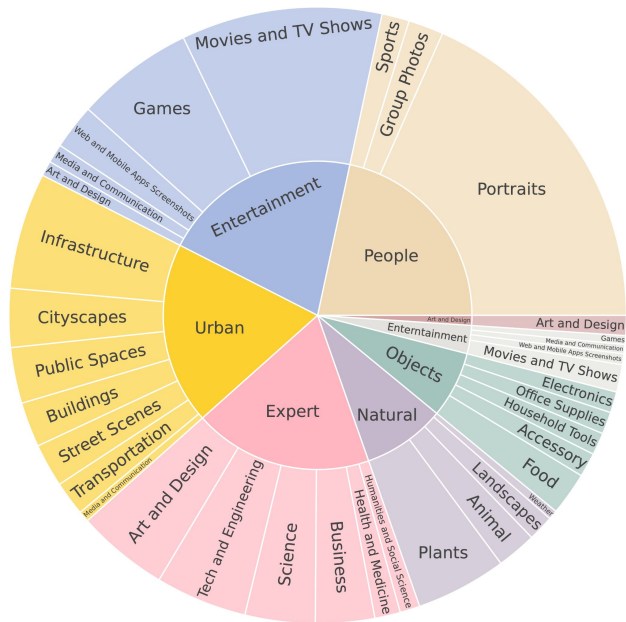
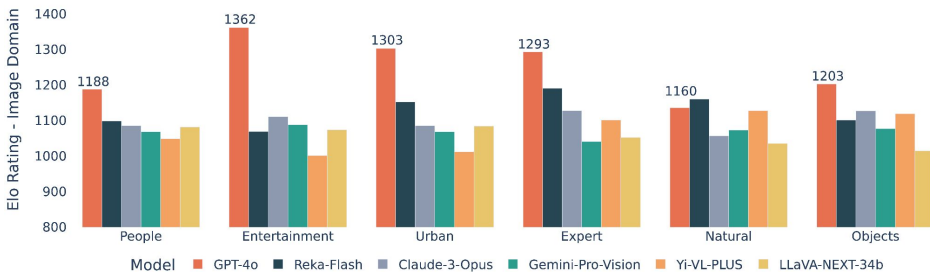


Image Domain

- GPT-4o is less competitive in Natural domain
- Gap between GPT-4o and other models are more obvious in Entertainment domain.
- Model performances are quite varied in Expert domain



WildVision Arena Leaderboard

- Online Elo Rating

$$P(Y_{ij} = 1) = \frac{1}{1 + 10^{(R_j - R_i)/\alpha}},$$

- Statistical Estimation

$$\mathcal{L}(\mathbf{R}) = \sum_{i,j \in \mathcal{N}, i \neq j} (W_{ij} Y_{ij} \log P(Y_{ij} = 1)),$$

Best Second Best Best among proprietary models Best among open-source models.

Models	Size	Elo	Battles	MMMU	Question Category			Image Domain		
					Analyt.	Descri.	Recogn.	Entert.	Objects	Expert
GPT-4O [69]	–	1235	434	62.8	1290	1250	1236	1362	1203	1293
GPT-4-Vision [68]	–	1132	2288	56.8	1154	1169	1099	1177	1109	1178
Reka-Flash [83]	–	1107	513	56.3	1093	1141	1067	1069	1101	1191
Claude-3-OPUS [2]	–	1100	908	59.4	1117	1096	1092	1111	1127	1128
Gemini-Pro-Vision [82]	–	1061	2229	47.9	1099	1041	1090	1088	1077	1041
Yi-VL-PLUS [1]	–	1061	283	–	1084	1040	1078	1001	1119	1101
LLaVA-NEXT [48]	34B	1059	1826	51.1	1068	1104	1021	1074	1015	1052
Gemini-1.5-Flash [81]	–	1055	132	–	1090	1018	1085	1190	990	1127
Claude-3-Sonnet [2]	–	1044	496	53.1	1063	1056	1041	1033	1023	1119
CogVLM-Chat-HF [89]	13B	1016	1024	32.1	950	947	1006	955	930	950
Claude-3-Haiku [2]	–	1002	419	50.2	964	1008	996	1033	1014	1005
LLaVA-NEXT [48]	7B	992	1367	35.1	963	1032	977	992	1023	1001
DeepSeek-VL [51]	7B	979	646	36.6	988	984	953	956	1026	962
Idefics2 [37]	8B	965	100	36.6	818	1003	1011	909	1071	1020
LLaVA-NEXT [48]	13B	956	201	35.9	965	974	1006	975	971	987
Qwen-VL-Chat [5]	10B	930	1328	35.9	898	937	940	923	942	902
Bunny-V1 [23]	3B	921	389	38.2	897	922	878	884	823	823
MiniCPM-V [26]	3B	910	1349	34.7	895	911	925	888	890	840
LLaVA-v1.5 [47]	13B	891	299	36.4	952	838	920	887	827	914
Tiny-LLaVA-v1-HF [111]	3B	879	288	33.1	901	828	821	808	853	894
InstructBLIP [14]	7B	862	807	30.6	834	856	891	840	902	763
UFORM-Gen2-Qwen [86]	500M	827	452	–	911	785	853	768	937	830

Live Leaderboard

🏆 WildVision Arena Leaderboard

Total #models: 23. Total #votes: 9748. Last updated: 2024-06-08 15:15:24 PDT.

Contribute your vote 🗳️ at [vision-arena!](#)

Rank	V-L Model	WV-Arena Elo	95% CI	Battles	MMMU	Org
1	gpt-4o	1233	+25/-21	736	69.1	OpenAI
2	gpt-4o-vision-preview	1129	+12/-11	2427	56.8	OpenAI
3	Reka-Flash	1100	+21/-19	576	56.3	Reka AI
4	claude-3-opus	1098	+18/-18	977	59.4	Anthropic
5	yi-v1-plus	1073	+32/-27	353		01 AI
6	gemini-1.5-flash-latest	1071	+33/-36	231	56.1	Google
7	gemini-pro-vision	1057	+11/-13	2345	47.9	Google
8	llava-v1.6-34b	1056	+14/-12	1899	51.1	UW Madison
9	claude-3-sonnet	1048	+29/-22	581	53.1	Anthropic
10	claude-3-haiku	1015	+21/-27	520	50.2	Anthropic
11	cogvlm-chat-hf	1014	+18/-16	1109	32.1	Tsinghua Univ.
12	llava-v1.6-vicuna-7b	989	+12/-17	1432	35.1	UW Madison
13	deepseek-v1-7b-chat	985	+17/-21	737	36.6	DeepSeek
14	MiniCPM-Llama3-V-2.5	969	+43/-62	100	45.8	OpenBMB
15	idefics2-8b-chatty	968	+52/-41	148	43	Hugging Face
16	llava-v1.6-vicuna-13b	954	+38/-43	201	35.9	UW Madison
17	Bunny-v1_0-3B	932	+27/-24	464	38.2	BAAI
18	Qwen-VL-Chat	931	+19/-17	1346	35.9	Alibaba
19	MiniCPM-V-1_0	911	+16/-18	1369	34.7	OpenBMB
20	tiny-llava-v1-hf	888	+27/-30	334	33.1	"SKLCCSE
21	llava-v1.5-13b	887	+38/-36	299	36.4	UW Madison
22	instructblip-vicuna-7b	861	+24/-22	807		Salesforce
23	uform-gen2-qwen-500m	830	+25/-21	505		Unum

Live Leaderboard

🏆 WildVision Arena Leaderboard

Total #models: 23. Total #votes: 9748. Last updated: 2024-06-08 15:15:24 PDT.

Contribute your vote 🗳️ at [vision-arena!](#)

Rank	V-L. Model	WV-Arena Elo	95% CI	Battles	MMMU	Org
1	gpt-4o	1233	+25/-21	736	69.1	OpenAI
2	gpt-4-vision-preview	1129	+12/-11	2427	56.8	OpenAI
3	Reka-Flash	1100	+21/-19	576	56.3	Reka AI
4	claude-3-opus	1098	+18/-18	977	59.4	Anthropic
5	yi-v1-plus	1073	+32/-27	353		01 AI
6	gemini-1.5-flash-latest	1071	+33/-36	231	56.1	Google
7	gemini-pro-vision	1057	+11/-13	2345	47.9	Google
8	llava-v1.6-34b	1056	+14/-12	1899	51.1	UW Madison
9	claude-3-sonnet	1048	+29/-22	581	53.1	Anthropic
10	claude-3-haiku	1015	+21/-27	520	50.2	Anthropic
11	cogvlm-chat-hf	1014	+18/-16	1109	32.1	Tsinghua Univ.
12	llava-v1.6-vicuna-7b	989	+12/-17	1432	35.1	UW Madison
13	deepseek-v1-7b-chat	985	+17/-21	737	36.6	DeepSeek
14	MiniCPM-Llama3-V-2.5	969	+43/-62	100	45.8	OpenBMB
15	idefics2-8b-chatty	968	+52/-41	148	43	Hugging Face
16	llava-v1.6-vicuna-13b	954	+38/-43	201	35.9	UW Madison
17	Bunny-v1_0-3B	932	+27/-24	464	38.2	BAAI
18	Qwen-VL-Chat	931	+19/-17	1346	35.9	Alibaba
19	MiniCPM-V-1_0	911	+16/-18	1369	34.7	OpenBMB
20	tiny-llava-v1-hf	888	+27/-30	334	33.1	"SKLCCSE
21	llava-v1.5-13b	887	+38/-36	299	36.4	UW Madison
22	instructblip-vicuna-7b	861	+24/-22	807		Salesforce
23	uform-gen2-qwen-500m	830	+25/-21	505		Unum

Wait for Arena-side compute

Wait for user input

Wait for enough voting

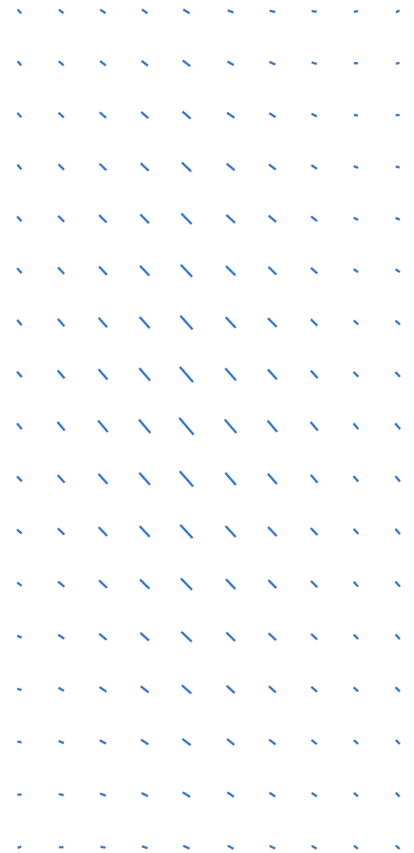
WildVision-Bench

Data Curation

- Public Set
 - 500 samples



Arena Data
20k+ chat 8k+ vote



Data Curation

- Public Set
 - 500 samples



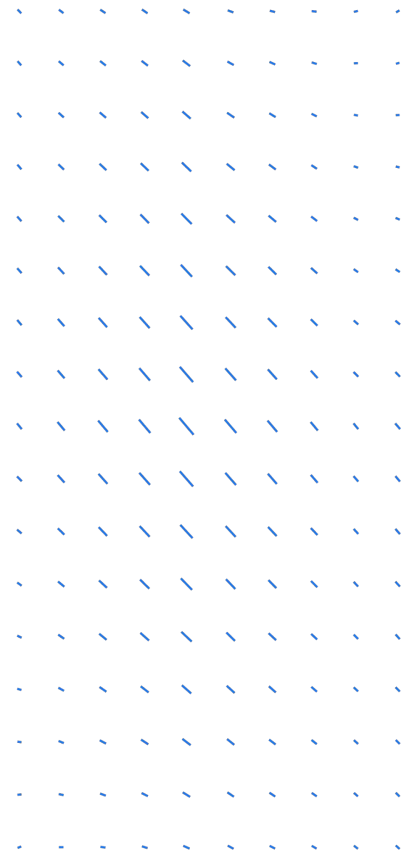
Sample Criteria

- Safety
- Diversity



Arena Data

20k+ chat 8k+ vote



Data Curation

- Public Set
 - 500 samples



Bench Data

500 sample



Sample Criteria

- Safety
- Diversity



Arena Data

20k+ chat 8k+ vote



Local Evaluator

- Arena Human vs. Expert
 - 72.5% agreement and 0.59 Cohen's Kappa
- GPT-4V vs. Arena Human Voting

Arena Human Vote	Left Vote	300	61	34
	Right Vote	102	269	27
	Tie Vote	99	111	41
	GPT-4V Vote	Left Vote	Right Vote	Tie Vote

- Agreement
 - 4-way: left/right/tie/both are bad
 - 3-way: left/right/tie or bad
 - Binary: left/right

Metric vs Human	GPT-4v		
	4-way	3-way	Binary
F1 Score (Macro)	0.4245	0.5143	0.7792
F1 Score (Micro)	0.5747	0.5842	0.7796
F1 Score (Weighted)	0.5407	0.5536	0.7798
Cohen's Kappa Score	0.3404	0.3442	0.5585
Pearson Correlation	0.2906	0.2880	0.5587

WV-Bench Score



Submit



Bench Data

500 sample



Sample Criteria

- Safety
- Diversity



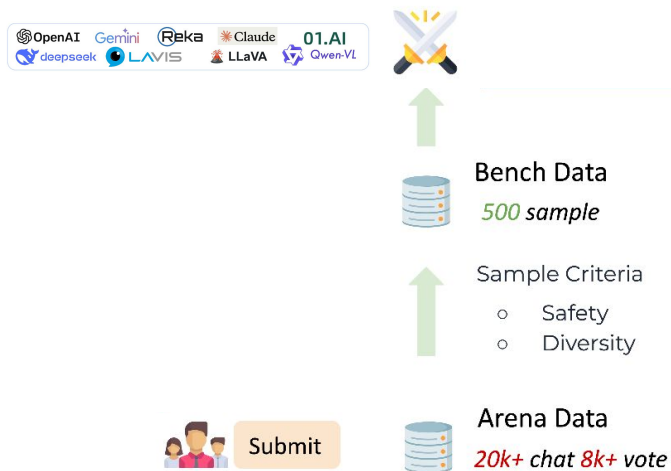
Arena Data

20k+ chat 8k+ vote



WV-Bench Score

- Precomputing



WV-Bench Score

- Precomputing
- Reference Model
 - Claude-3-Sonnet



Reference

Claude-3-Sonnet



Bench Data

500 sample



Sample Criteria

- Safety
- Diversity



Submit



Arena Data

20k+ chat 8k+ vote



WV-Bench Score

- Precomputing
- Reference Model
 - Claude-3-Sonnet
- Judge Model
 - GPT-4o



Judge
GPT-4o

Both models
are correct



Reference
Claude-3-Sonnet



Bench Data
500 sample



Sample Criteria

- Safety
- Diversity



Submit

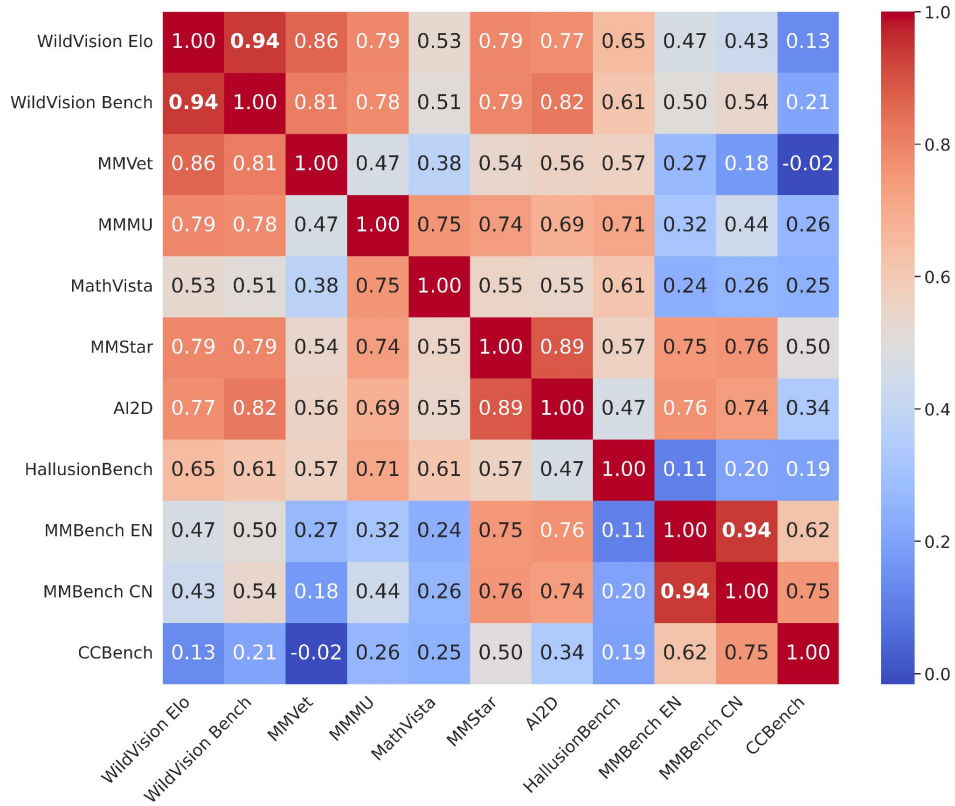


Arena Data
20k+ chat 8k+ vote



WildVision-Bench Leaderboard

Model	Score	95% CI	Win Rate	Reward	Much Better	Better	Tie	Worse	Much Worse	Avg Tokens
GPT-4o [69]	89.41	(-1.7, 2.0)	80.6%	56.4	255.0	148.0	14.0	72.0	11.0	157
GPT-4-Vision [68]	80.01	(-1.9, 2.8)	71.8%	39.4	182.0	177.0	22.0	91.0	28.0	140
Reka-Flash [83]	64.79	(-2.9, 3.0)	58.8%	18.9	135.0	159.0	28.0	116.0	62.0	181
Claude-3-Opus [2]	62.15	(-2.8, 3.4)	53.0%	13.5	103.0	162.0	48.0	141.0	46.0	120
Yi-VL-PLUS [1]	55.09	(-2.9, 3.0)	52.8%	7.2	98.0	166.0	29.0	124.0	83.0	150
LLaVA-NEXT-34B [48]	51.91	(-3.1, 2.4)	49.2%	2.5	90.0	156.0	26.0	145.0	83.0	165
Claude-3-Sonnet [2]	50.00	-	-	-	-	-	-	-	-	120
Claude-3-Haiku [2]	37.70	(-3.2, 4.2)	30.6%	-16.5	54.0	99.0	47.0	228.0	72.0	97
Gemini-Pro-Vision [82]	35.45	(-2.6, 3.2)	32.6%	-21.0	80.0	83.0	27.0	167.0	143.0	66
LLaVA-NEXT-13B [48]	33.69	(-3.8, 2.7)	33.8%	-21.4	62.0	107.0	25.0	167.0	139.0	138
DeepSeek-VL-7B [51]	33.48	(-2.2, 3.0)	35.6%	-21.2	59.0	119.0	17.0	161.0	144.0	119
CogVLM-Chat-HF [89]	31.88	(-2.7, 2.4)	30.6%	-26.4	75.0	78.0	15.0	172.0	160.0	63
LLaVA-NEXT-7B [48]	26.15	(-2.7, 2.3)	27.0%	-31.4	45.0	90.0	36.0	164.0	165.0	139
Idefics2 [37]	23.71	(-2.4, 2.5)	26.4%	-35.8	44.0	88.0	19.0	164.0	185.0	128
Qwen-VL-Chat [5]	17.87	(-2.6, 2.2)	19.6%	-47.9	42.0	56.0	15.0	155.0	232.0	70
LLaVA-v1.5-13B [47]	14.15	(-2.2, 2.2)	16.8%	-52.5	28.0	56.0	19.0	157.0	240.0	87
Bunny-3B [23]	12.70	(-1.8, 1.9)	16.6%	-54.4	23.0	60.0	10.0	164.0	243.0	76
MiniCPM-V [26]	11.66	(-1.8, 2.1)	13.6%	-57.5	25.0	43.0	16.0	164.0	252.0	89
Tiny-LLaVA [111]	8.01	(-1.4, 1.4)	11.0%	-66.2	16.0	39.0	15.0	127.0	303.0	74
UFORM-Gen2-Qwen [86]	7.55	(-1.6, 1.1)	10.8%	-68.5	16.0	38.0	11.0	115.0	320.0	92
InstructBLIP-7B [14]	5.54	(-1.3, 1.5)	7.8%	-72.5	11.0	28.0	15.0	117.0	329.0	47



Benchmark Correlation

VLMs in the wild

Expert annotation as a preview

Hidden Set will be updated periodically to avoid contamination of leaderboard.

Image #44 in WildVision Bench



Text Prompt: Answer shortly what is he doing? And what do we call him?

GPT-4V: He is performing a breakdance move. We call him a breakdancer or b-boy.

LLaVA-v1.6-34B: He is performing a dance move that involves spinning or flipping while balancing on one hand ..., he may be referred to as a dancer or acrobat.

Gemini Pro Vision: He is breakdancing. He is called a b-boy or breaker.

Image #61 in WildVision Bench

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

Text Prompt: Convert the first row of this sudoku matrix into the text format, each cell separated by a space. Use 0 to represent empty cells.

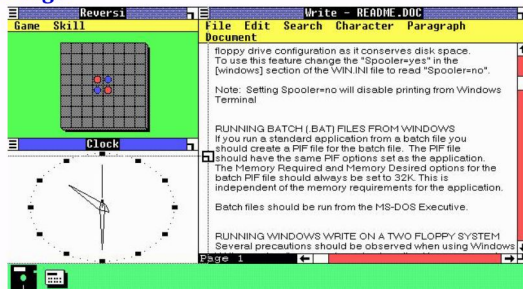
GPT-4V: 6 8 4 2 1 0 0 0 0

LLaVA-v1.6-34B: 6 8 4 0 2 1 0 7 3 4 0 6

Gemini Pro Vision: 6 8 4 2 1 9 8 7 7

Failure Cases

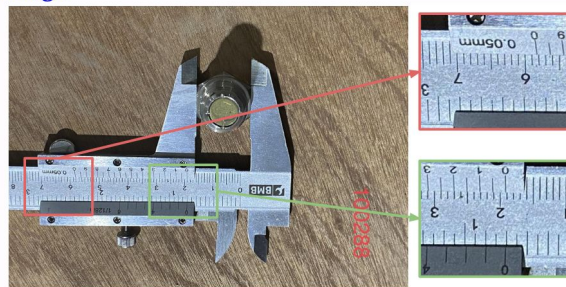
Image



Text Prompt: What operating system is this?

GPT-4V: The image shows a screenshot of Windows 3.x, which is a series of graphical operating environments produced by Microsoft for use on personal computers, prior to the Windows 95 operating system...

Image



Text Prompt: What is the length of the object shown in the image.

Gemini-Pro-Vision: The diameter of the bolt is 7.92 mm.

Image Error #1



Text Prompt: What should I play in this position?

GPT-4V: This chess position is unconventional and humorous because it features a board filled with knights (horses) except for a single queen. In a standard game of chess, this situation would not occur, as each player starts with just two knights. In this specific position, there's not much strategy to discuss because the game doesn't reflect an actual playable scenario. However, if we want to engage with this as a fun exercise in imagination, you could consider moving one of the knights that isn't pinned or creating a threat to the queen. For example, you could move one of the knights on the third rank to attack the queen, such as Nc3-d5 or Ne3-c4. Remember, this is just a playful scenario and not representative of a legitimate chess position. In a real game, the best move would depend on a much broader context of the board and the strategic positions of all the pieces, not just an army of knights.

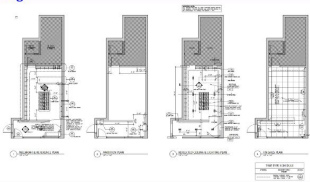
Image Error #3



Text Prompt: What food will these people have very soon?

LLaVA-NEXT-34B: Based on the image, the people are lined up outside a building with a sign that reads "Salsa's Authentic Mexican Restaurant." It seems likely that they will be having Mexican food very soon, which could include items such as tacos, burritos, enchiladas, chiles rellenos, guacamole, and other similar dishes that are typically served at Mexican restaurants. The specific menu items will depend on the restaurant's offerings and the preferences of the customers.

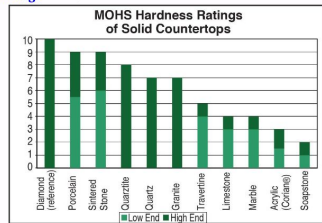
Image Error #2



Text Prompt: Based on the architecture plans is this compliant in the Ontario building code

GPT-4V: I'm sorry, but as an AI, I do not have the capacity to verify the compliance of architectural plans with specific building codes. Building code compliance typically involves a wide range of considerations, including but not limited to accessibility, fire safety, structural integrity, and mechanical systems. Professional architects and engineers, often alongside local code enforcement officials, ensure that designs and construction meet the appropriate codes and regulations. To determine compliance with the Ontario Building Code, the plans would need to be reviewed by a certified professional who has access to and understands the local regulations and requirements. They would check if the design meets all necessary stipulations for things such as room sizes, egress requirements, bathroom accessibility, fire ratings, and many other details. You should consult with a licensed Ontario architect or a building code consultant for an official assessment of compliance.

Image Error #4



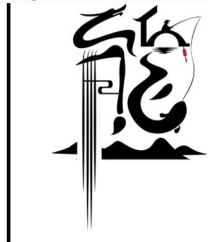
Text Prompt: Based on this chart, what is the MOHS hardness of Marble?

LLaVA-NEXT-34B: The MOHS hardness of Marble, as indicated on the chart, is 3.

Failure Cases

- Visual Recognition
- Visual Reasoning
- Spatial Imagination
- Contextual Understanding
- Expert Domain Knowledge
- Hallucination

Image Error #5



Text Prompt: 图片中是什么字

GPT-4V: 图片中的字体设计巧妙地融合了艺术与文字，呈现出既像文字又像艺术作品的效果。这个特殊的设计使得文字的具体含义不太直观，没有提供清晰的线索来解释这些字符的标准解读。如果你有更多关于这些字的背景信息或上下文，我可以试图进行更多的解释。不过，从图像中直接解读标准文本是非常困难的。

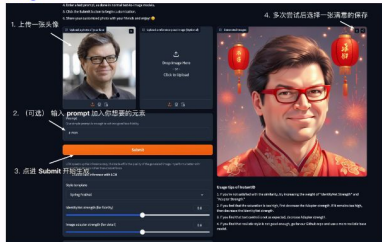
Image Error #7



Text Prompt: which movie is this picture from? what is it about?

LLaVA-NEXT-34B: This image appears to be from the movie 'The Girl on the Train' (2012), which is a psychological thriller film directed by David Riker and starring Abbie Cornish and Caleb Landry Jones. The film is about a young woman who is hired to care for a girl living in a remote Texas farmhouse, but soon discovers that the girl is not what she seems. The story revolves around themes of isolation, identity, and the consequences of one's actions.

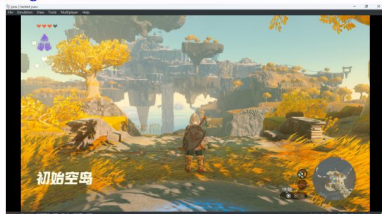
Image Error #6



Text Prompt: 这张图里右侧的照片和左侧相比有什么变化

Gemini-Pro-Vision: 右侧的照片中，人物的脸部表情更加丰富，并且背景中添加了春节的元素，使画面更具喜庆的氛围。

Image Error #8



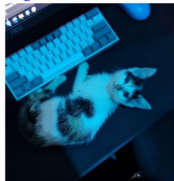
Text Prompt: do u know which game i am playing?

LLaVA-NEXT-34B: You are playing The Legend of Zelda: Breath of the Wild, which is an action-adventure game developed and published by Nintendo for the Wii U and Nintendo Switch consoles. The game is known for its open-world exploration and its unique gameplay mechanics.

Failure Cases

- Visual Recognition
- Visual Reasoning
- Spatial Imagination
- Contextual Understanding
- Expert Domain Knowledge
- Hallucination

Image Error #9



Text Prompt: How many keys does the keyboard have?

GPT-4V: The keyboard in the image features a non-standard layout, which appears to be a compact form factor lacking the numeric keypad commonly found on full-sized variants. Without being able to view the entire keyboard, it's impossible to provide an exact count of the keys...

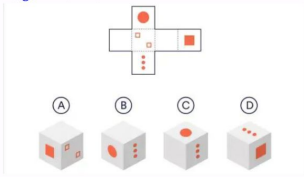
Image Error #11



Text Prompt: what is this

GPT-4V: This is an oversized, artistic representation of a classic hand-held manual egg beater or whisk. It's designed to mimic the appearance of a traditional egg beater with an exaggerated size, likely for decorative or exhibit purposes as suggested by the setting that appears to be a show or exhibition hall. The inclusion of this piece in a public space may indicate it being part of an art installation, event display, or perhaps a promotion related to culinary equipment or innovation.

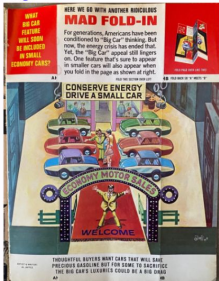
Image Error #10



Text Prompt: By examining the unfolded cube (net) at the top of the image, which one of the four cubes labeled A, B, C, and D can be created by folding this net along the edges?

Gemini-Pro-Vision: The correct answer is C. The unfolded cube (net) at the top of the image has two red squares and three orange circles. Cube C has two red squares and three orange circles in the same arrangement as the unfolded cube.

Image Error #12



Text Prompt: Try to guess the solution to this Fold-in from Mad magazine. A Mad Fold-in is a visual riddle that a human can easily solve by physically folding the page as explained in the instructions in the top-right corner. I know the solution. I want to verify if you are able to solve it. What big car feature will soon be included in small economy cars?

LLaVA-NEXT-34B: The Mad fold-in features a car salesman in front of a sign that says "CONSERVE ENERGY DRIVE A SMALL CAR" with various "big car" features like a tire iron, a spare tire, a picnic basket, and a trunk full of items. When you fold the page as directed, it's revealed that the large "big car" features are actually inside the small car, showing that the small car also has these items, humorously implying that the small car will soon include all the features typically associated with big cars. The joke is that despite their size, small cars will also include all the amenities that were once exclusive to larger vehicles.

Failure Cases

- Visual Recognition
- Visual Reasoning
- Spatial Imagination
- Contextual Understanding
- Expert Domain Knowledge
- Hallucination

Image Error #13



Text Prompt: What is the name of the character with the sunglasses?

GPT-4V: You're looking at an image from the animated film Zootopia. The character with the sunglasses is named Nick Wilde; he's a red fox who is one of the main characters in the movie.

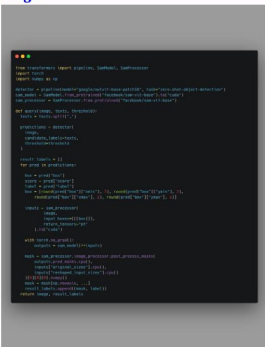
Image Error #14



Text Prompt: Are the workers wearing safety devices?

LLaVA-NEXT-34B: In the image you've provided, the workers are wearing safety vests, which are commonly used in construction and industrial environments to increase visibility and reduce the risk of accidents. These vests are a standard safety device to help workers be more visible to operators of heavy machinery and other workers in bright colors that contrast with the surroundings. The workers also appear to be wearing hard hats, which are required on construction sites to protect the head from potential injuries.

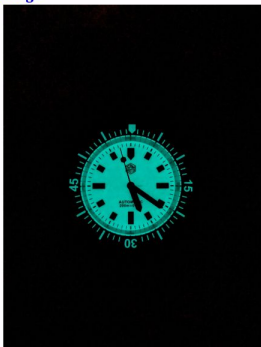
Image Error #15



Text Prompt: What do you see in the image?

Gemini-Pro-Vision: This is a code snippet in Python. It seems to be a function that preprocesses an image for object detection. The function takes an image as input and returns a preprocessed image and a list of bounding boxes.

Image Error #16



Text Prompt: What time is it on that watch?

LLaVA-NEXT-34B: The watch in the image is displaying a time that appears to be 10:11. The hour hand is pointing to the number 10, and the minute hand is pointing to the number 11.

Failure Cases

- Visual Recognition
- Visual Reasoning
- Spatial Imagination
- Contextual Understanding
- Expert Domain Knowledge
- Hallucination

Image Error #17



Text Prompt: where to find food?

Gemini-Pro-Vision: The food is in the refrigerator, which is located in the kitchen. The kitchen is the room with the stove and oven.

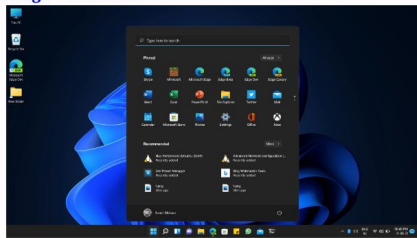
Image Error #19



Text Prompt: What Brand is highlighted

LLaVA-NEXT-7B: The brand highlighted in the image is Head.

Image Error #18



Text Prompt: What time is it?

LLaVA-NEXT-34B: The time is 06:54 PM.

Image Error #20



Text Prompt: What causes this and which is the real one?

Qwen-VL-Chat: It is funny because the guy who is not in a white t-shirt is saying that he is in a white t-shirt.

Failure Cases

- Visual Recognition
- Visual Reasoning
- Spatial Imagination
- Contextual Understanding
- Expert Domain Knowledge
- Hallucination

What's next for Eval

- API support for research purpose
- Multi-facet comparison
- Fine-grained live leaderboard
- Rank bootstrapping from VLMs

What's next for VLMs

- Modality Integration
- Resolution
- Long Context
- Resource-Efficient
- World Knowledge
- Safety
- ...

Arena Data

model_a	model_b	conversation_a	conversation_b	language	image	turn	anony	winner
string · classes	string · classes	list · lengths	list · lengths	string · classes	image · width (px)	int32	bool	string · classes
gpt-4-vision-preview	gemini-pro-vision	[{ "role": "user", "content": "Can you derive Equation 6 from the image?" }, { "role": "...	[{ "role": "user", "content": "Can you derive Equation 6 from the image?" }, { "role": "...	English		1	1	true model_a
gpt-4-vision-preview	gemini-pro-vision	[{ "role": "user", "content": "Given my horse's location on this map, what is the quickest route...	[{ "role": "user", "content": "Given my horse's location on this map, what is the quickest route...	English		1	1	true model_a
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Can you describe the interesting part of the image?" }, { "role": "...	[{ "role": "user", "content": "Can you describe the interesting part of the image?" }, { "role": "...	English		1	1	true model_a
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Can you write codes to load this 3d objects?" }, { "role": "...	[{ "role": "user", "content": "Can you write codes to load this 3d objects?" }, { "role": "...	English		1	1	true model_a
llava-v1.5-13b	gemini-pro-vision	[{ "role": "user", "content": "Please annotate as many as objects in the image as possible. Provide...	[{ "role": "user", "content": "Please annotate as many as objects in the image as possible. Provide...	English		1	1	true model_b
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Write me the codes which load an image editing model from hugging...	[{ "role": "user", "content": "Write me the codes which load an image editing model from hugging...	English		1	1	true model_a
gemini-pro-vision	gpt-4-vision-preview	[{ "role": "user", "content": "Please write me the code to reproduce the charts using python." }]	[{ "role": "user", "content": "Please write me the code to reproduce the charts using python." }]	English		1	1	true model_a
llava-v1.5-13b	gpt-4-vision-preview	[{ "role": "user", "content": "Please convert these 2 sections in the paper to the markdown...	[{ "role": "user", "content": "Please convert these 2 sections in the paper to the markdown...	English		1	1	true model_b
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Please write a interesting tweet with this image if you are an...	[{ "role": "user", "content": "Please write a interesting tweet with this image if you are an...	English		1	1	true model_b
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Please write a interesting tweet with this image if you are an...	[{ "role": "user", "content": "Please write a interesting tweet with this image if you are an...	English		1	1	false model_a
gemini-pro-vision	gpt-4-vision-preview	[{ "role": "user", "content": "My guitar was broken as shown, how can I fix it? Please give me...	[{ "role": "user", "content": "My guitar was broken as shown, how can I fix it? Please give me...	English		1	1	true model_b
llava-v1.5-13b	gpt-4-vision-preview	[{ "role": "user", "content": "My guitar was broken as shown, how can I fix it? Please give me...	[{ "role": "user", "content": "My guitar was broken as shown, how can I fix it? Please give me...	English		1	1	true model_b
llava-v1.5-13b	gemini-pro-vision	[{ "role": "user", "content": "Please solve this physics problem." }, { "role": "assistant" }]	[{ "role": "user", "content": "Please solve this physics problem." }, { "role": "assistant" }]	English		1	1	true tie (bothbad)
llava-v1.5-13b	gemini-pro-vision	[{ "role": "user", "content": "Can you solve the 3 questions in image?" }, { "role": "assistant" }]	[{ "role": "user", "content": "Can you solve the 3 questions in image?" }, { "role": "assistant" }]	English		1	1	true model_a
gemini-pro-vision	llava-v1.5-13b	[{ "role": "user", "content": "Can you solve the 3 questions in the image?" }, { "role": "...	[{ "role": "user", "content": "Can you solve the 3 questions in the image?" }, { "role": "...	English		1	1	true model_a
gpt-4-vision-preview	llava-v1.5-13b	[{ "role": "user", "content": "Can you solve the 3 questions in the image?" }, { "role": "...	[{ "role": "user", "content": "Can you solve the 3 questions in the image?" }, { "role": "...	English		1	1	true model_a
llava-v1.5-13b	gemini-pro-vision	[{ "role": "user", "content": "What are the advantages and disadvantages of this room design...	[{ "role": "user", "content": "What are the advantages and disadvantages of this room design...	English		1	1	true tie (bothbad)

Take-home messages

- **WV-Arena:** Our WildVision-Arena is an interactive evaluation platform that hosts a live leaderboard reflecting crowdsourced user preferences on real-world chats.
- **WV-Bench:** Our WildVision-Bench is a fast-evaluation benchmark that closely aligned with human preferences at 0.94 Spearman's Correlation.
- GPT-4V surpasses many other models in simple visual recognition and reasoning tasks, it still faces challenges with subtle contextual cues, spatial reasoning, visual imagination, and expert domain knowledge.
- Current VLMs exhibit issues with hallucinations and safety when intentionally provoked.
- **Data Release:** We are releasing our chat and feedback data to further advance research in the field of VLMs.

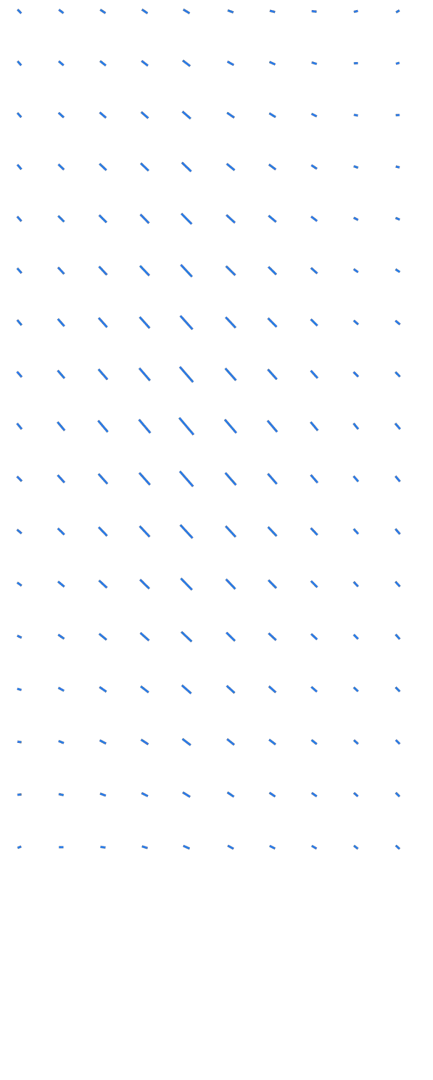
Open-source

Code, Conversation and Feedback data will be updated periodically.

We welcome community effort and call for collaboration and support.



WildVision Arena



WildVision Team

Core contributors:

Yujie Lu, Dongfu Jiang

Advisors:

Wenhu Chen, William Wang, Yejin Choi, and Bill Yuchen Lin



Thank You

Contact: yujielu@umail.ucsb.edu (Yujie Lu)