

マルチモーダル基盤モデルと 教育への応用

岡崎 直観

東京科学大学 情報理工学院情報工学系

国立情報学研究所 大規模言語モデル研究開発センター 科学主幹

産業技術総合研究所 人工知能研究センター 招聘研究員

マルチモーダル基盤モデルとは

言語や画像、音声など、様々な伝達手段の情報を統合的に処理する基盤モデル

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? Options: (A) Major third <i><image 1></i> (B) Diminished fifth <i><image 2></i> <u>(C) Minor seventh <i><image 3></i></u> (D) Diminished sixth <i><image 4></i>	Question: ...The graph shown is compiled from data collected by Gallup <i><image 1></i>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? Options: (A) 0 (B) 0.2142 <u>(C) 0.3571</u> (D) 0.5	Question: <i><image 1></i> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. Options: <u>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</u> (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x(x)] dx$
Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium	Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium	Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy
Health & Medicine	Humanities & Social Science	Tech & Engineering
Question: You are shown subtraction <i><image 1></i>, T2 weighted <i><image 2></i> and T1 weighted axial <i><image 3></i> from a screening breast MRI. What is the etiology of the finding in the left breast? Options: (A) Susceptibility artifact (B) Hematoma <u>(C) Fat necrosis</u> (D) Silicone granuloma	Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? <i><image 1></i> Option: (A) Oppressor (B) Imperialist <u>(C) Savior</u> (D) Isolationist	Question: Find the VCE for the circuit shown in <i><image 1></i>. Neglect VBE Answer: 3.75 Explanation: ...$IE = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_{E} R_L = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$
Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard	Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy	Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard

MMMUの問題の例 (Liu+ 2024)

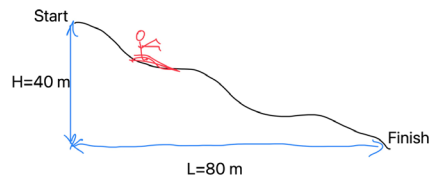
マルチモーダル基盤モデル (Google Gemini)



Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

$$\text{Start: } E = mgL$$

$$\text{End: } E = \frac{1}{2}mv^2$$

$$\Rightarrow mgL = \frac{1}{2}mv^2$$

$$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6 \text{ m.s}^{-1}$$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.



Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh , where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L , instead of the height, H , in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh .

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for v , we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

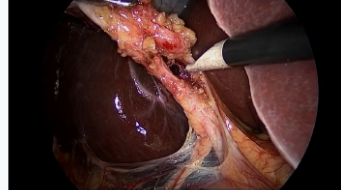
Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

マルチモーダル基盤モデルの医療ドメインへの応用: Med-Gemini

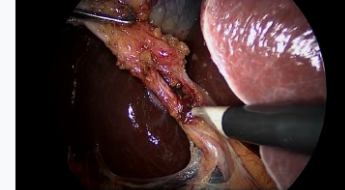
Prompt



...



...



You are a helpful medical video assistant.

Task:

Assess whether Critical View of Safety (CVS) is fully achieved in the provided cholecystectomy video.

The Critical View of Safety (CVS) is fully achieved if the following three criteria are met:

- C1: Clear view of 2 tubular structures connected to the gallbladder.
- C2: A carefully dissected hepatocystic triangle presenting an unimpeded view of only the 2 cystic structures and the cystic plate.
- C3: The lower third of the gallbladder is dissected off the cystic plate.

Instructions:

1. Watch the video carefully, and answer which of the Critical View of Safety (CVS) criteria are met.
2. Output Format:
 - CVS Assessment:
 - C1 - achieved / partially achieved / not achieved (Optional explanation)
 - C2 - achieved / partially achieved / not achieved (Optional explanation)
 - C3 - achieved / partially achieved / not achieved (Optional explanation)

Response

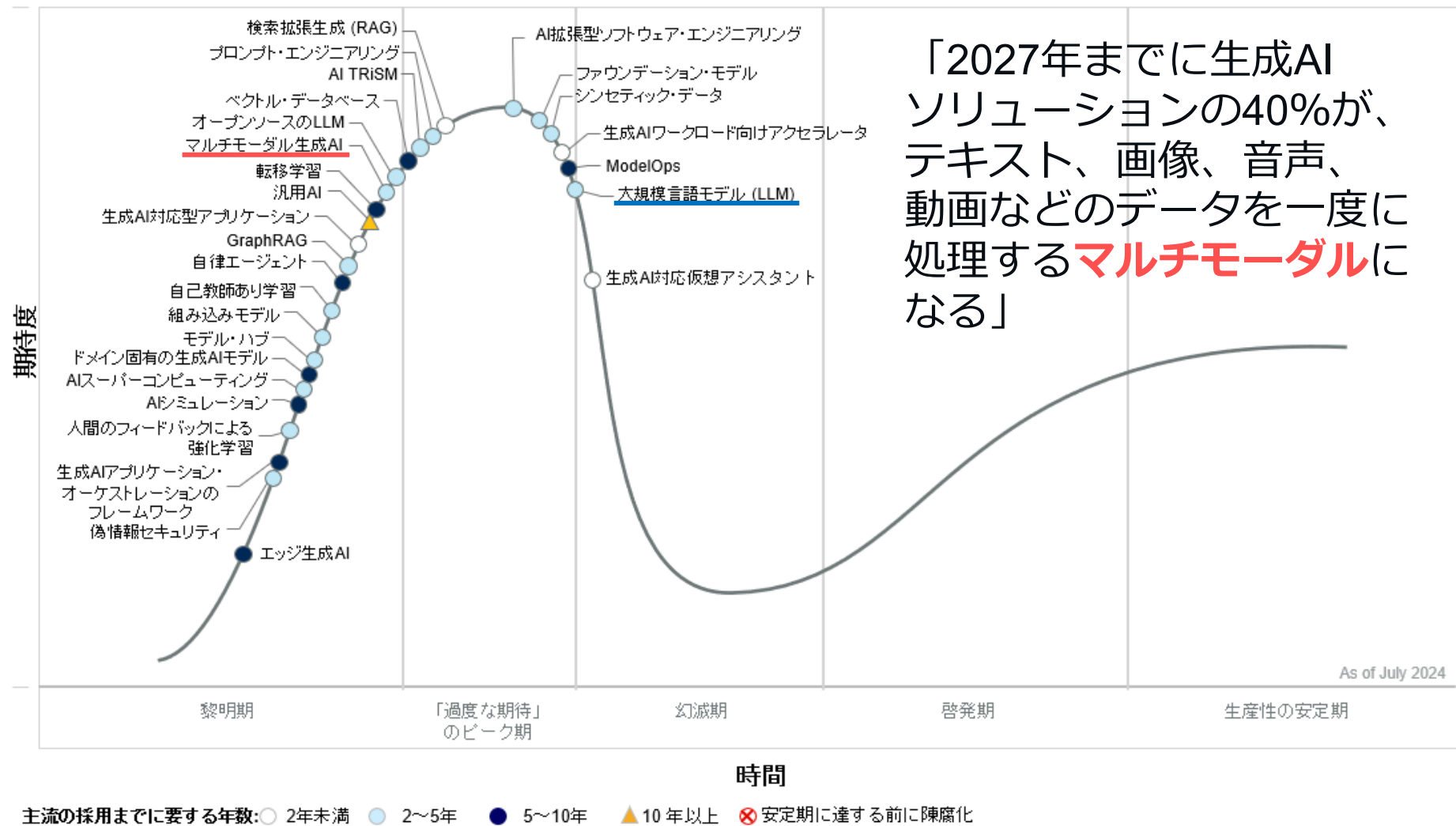
CVS Assessment:

C1 - partially achieved: Two tubular structures are visible but the view is not entirely clear due to some bleeding obscuring the area.

C2 - not achieved: The hepatocystic triangle is not clearly dissected, and there is not an unimpeded view of only the two cystic structures and the cystic plate.

C3 - not achieved: The lower third of the gallbladder is not dissected off the cystic plate.

今後の成長が期待されるマルチモーダル基盤モデル



2024年9月10日、Gartner、「生成AIのハイブ・サイクル：2024年」を発表。 <https://www.gartner.co.jp/ja/newsroom/press-releases/pr-20240910-genai-hc>

マルチモーダル基盤モデルの科学教育への応用 (Bewersdorff+ 2024)

人間が科学を学習するときも、テキストや図などを用いたマルチモーダル処理である
(参考: マルチメディア学習)

学習支援

- テキストを視覚情報で変換・補完する
- 科学的な言語の理解と使用の支援
- 技術的な言葉を簡略化し、専門用語を説明する
- リサーチクエスチョンや仮説の立案を支援する
- 生データの可視化と解釈
- 視覚的表現に文脈的な説明を添える
- 効果的なコミュニケーションの支援
- 科学現象のアナロジーから視覚ベースのストーリーを生成する

コンテンツ（教材）作成

- 多様な学生のニーズに合わせた学習教材の作成
- 認知負荷を軽減するためのコンテンツ整理
- 生成的な活動を通じた積極性の促進
- 認知を容易にするための視覚的表現の柔軟な適応
- 仮想現実環境のためのコード・コンテンツ生成、マルチモーダル基盤モデルの包括的な統合

宿題とフィードバック

- 学生のレポートの個別評価
- 評価の質と客観性の向上
- 視覚的補助を用いた詳細なフィードバックの提供
- テキストや図画など、さまざまなモダリティに対する即時フィードバック

GPT-4oで実験計画を作成する例

【小学生自由研究テーマNo.129】

水をいっしゅんでこおらせるマジック！

ペットボトルの中の水が、いっしゅんでこおったらビックリするよね。試してみよう！
うまくできたら、なぜこおったのかも考えよう。



対象学年

小学1年生

小学2年生

小学3年生

小学4年生

小学5年生

小学6年生

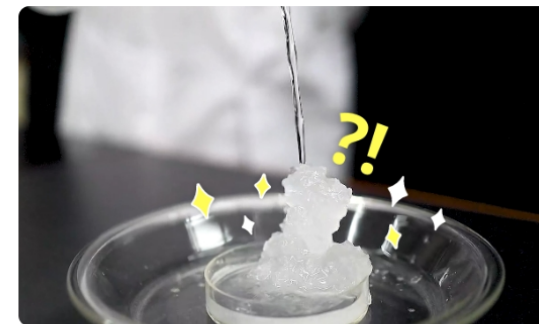
制作日数

1日

カテゴリ

工作

実験



どのような手順で実験をすればよいか、教えてください。また、なぜこのような現象が起こるのか、説明してください。



この写真で行われている実験は、水を「人工雪」として知られる吸水性ポリマーに注いで、雪のように膨らむ現象を観察しているものと思われます。この現象は、吸水性ポリマー（例えばポリアクリル酸ナトリウム）が水分を吸収し、大量の水を保持する能力があるために起こります。以下は、この実験を行う手順と、現象の説明です。

実験手順

1. 用意するもの：吸水性ポリマー（ポリアクリル酸ナトリウムなど）、水、透明の皿やペトリ皿、小さなスプーン。
2. ポリマーを皿に入れる：透明の皿やペトリ皿の中央に少量の吸水性ポリマーを置きます（小さじ1杯程度）。
3. 水を加える：ポリマーの上からゆっくりと水を注ぎます。このとき、ポリマーが急激に膨張し、雪のように見える状態になります。
4. 観察：数分間観察すると、ポリマーが水を吸収してふくらみ、柔らかく白い「雪」のような見た目になります。

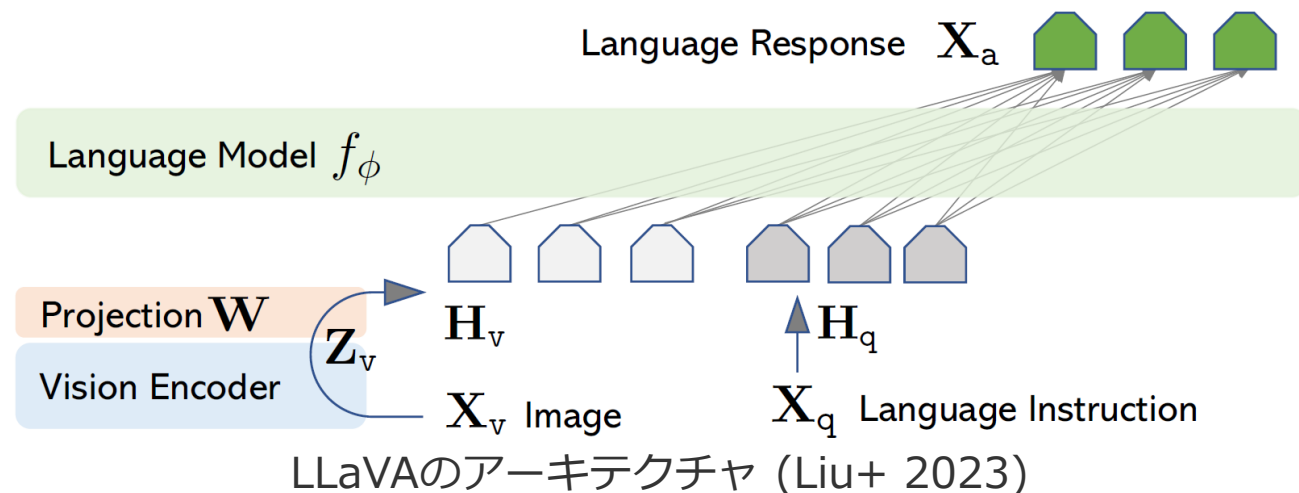
現象の説明

この現象は、吸水性ポリマーの高い吸水力によるものです。吸水性ポリマーは、分子構造に多くの親水基（例えば

マルチモーダル基盤モデルの構築

LLM-jpおよび大規模言語モデル研究開発センター (LLMC) での取り組み

画像と言語のマルチモーダル基盤モデル (VLM): LLaVA (Liu+ 2023)



入力画像の例^[1]

学習手順

1. 言語モデル f_ϕ と画像エンコーダを事前に学習しておく
2. 画像の特徴量 Z_v を言語の空間に変換する行列 W を学習する

$$H_v = WZ_v, \quad Z_v = g(X_v)$$

※画像 X_v 、質問 X_q 、答え X_a が組になった学習データを用いる

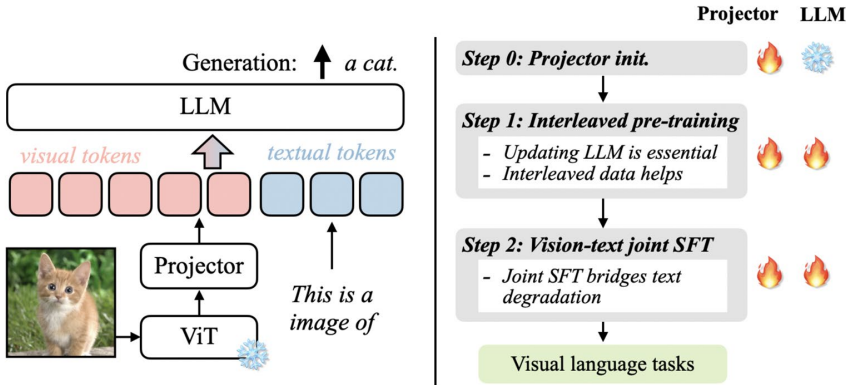
3. 変換行列 W と言語モデル f_ϕ を同時に学習する (画像エンコーダのパラメータは固定しておく)

User: What is unusual about this image?

LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

今回構築したVLMの構成

構成要素	モデル / アーキテクチャ
画像エンコーダ	google/siglip-so400m-patch14-384
プロジェクタ	2層の線形変換
LLM	llm-jp/llm-jp-3-13b-instruct



VILA (先行研究) (Lin+ 2024)

学習段階	日本語データ	英語データ	学習時間
Step 0	日本語画像・テキスト対 (558k)	LLaVA-1.5事前学習データ ³ (558k)	LLM-jp クラスタ 8 x A100, 約14-15時間
Step 1	日本語画像・テキスト対(6M), 日本語版MMC4 (6M)	COYO ⁴ の一部 (6M), mmc4-core ⁵ の一部 (6M)	LLM-jp NVLINK クラスタ 8 x 8 x A100, 約130時間
Step 2	llava-instruct-ja (156k), japanese-photos-conv (12k), ja-vg-vqa ¹ (99k), synthdog-ja ² (102k), 計369k事例	LLaVA-1.5指示チューニングデータ ⁶ (158k), VQAv2 (53k), GQA (46k), OCRVQA (80k), TextCaps (22k), 計358k事例	LLM-jp NVLINK クラスタ 4 x 8 x A100, 約11時間

J Lin, H Yin, W Ping, Y Lu, P Molchanov, A Tao, H Mao, J Kautz, M Shoenybi, S Han. 2024. [VILA: On Pre-training for Visual Language Models](#). CVPR.

構築した日本語の画像・テキストデータセット (1/2)

「東京科学大学」になった旧・東京工業大学、「工業」弔う墓出現
卒業生「無念さを少しは晴らしてくれたのかな」

10/3(木) 20:07 配信 52 5 5 5 5 5

JCASTニュース
ビジネス&メディア ウォッチ



旧・東京工業大学の大岡山キャンパス

2024年10月1日に東京医科歯科大学と統合し「東京科学大学」となった旧・東京工業大学（東工大）の大岡山キャンパス（東京都目黒区）内に、「工業之墓」と書かれた謎の「墓」が出現した。名称が切り替わる9月30日夜から10月1日にかけて写真が複数投稿され、Xで話題となった。

【画像】旧・東京工業大学に出現した「工業之墓」



旧・東京工業大学の
大岡山キャンパス

日本語画像・テキスト対

<https://www.j-cast.com/2024/10/03494892.html>

“また、¥n地下の木挽町広場には、「和」を意識した¥nお店がたくさん入っていて、眺めるだけでも¥n楽しめます¥n”



“もうお雛様が飾られていました¥n¥n”

“投稿者：山本歯科診療所¥n¥n”

“2016.01.18更新¥n¥n”

“### ドクターズスイーツ¥n¥n”

“「ドクターズスイーツ」というのが¥n最近流行っているようです¥n”



主に、糖尿病で糖質の摂取が制限されて¥nいる人でも食べられるお菓子などですが、¥n歯科でもあります！¥n

日本語版MMC4

<http://www.yamamoto-dental-clinic.com/blog/staff/2016/01/>

構築した日本語の画像・テキストデータセット (2/2)



Q: この画像には何が写っていますか？

A: 画像には、白と赤のバスが都市の通りを走っている様子が写っています。バスは比較的大きく、広告が後部に貼られているようです。また、周囲には多くの人や建物があります。

.....

llava-instruct-ja



Q: この画像にはどんな建物が写っていますか？

A: 画像には「LaLaport」と書かれた建物が写っています。

Q: この画像の中で目立つ大きなフィギュアは何ですか？

A: 目立つ大きなフィギュアはガンダムのロボットです。

.....

japanese-photos-conv

<https://huggingface.co/datasets/ThePioneer/japanese-photos>

評価ベンチマーク: Japanese Heron Bench (Inoue+ 2024) の例



Q.映像の状況において、青信号になったら車を発進させてもよいでしょうか？

(In the situation in the video, may I start the car when the light turns green?)



GPT-4 (Ref.):

いいえ、青信号になったとしても、交通誘導員が「止まれ」の合図をしているため、車を発進させてはいけません。交通誘導員の指示は信号よりも優先されます。また、片側交互通行の状況下では、誘導員の指示に従う必要があります。

Qwen2-VL-7B:

2 / 10

映像の状況では、青信号が点灯しているため、車を発進させることができます。ただし、安全確認を怠らないように注意してください。

GPT-4o:

8 / 10

像を見る限り、現在道路工事が行われているようです。工事現場では通常、交通誘導員が交通の流れをコントロールします。青信号になったとしても、交通誘導員の指示が優先されます。

この場合は、交通誘導員の指示に従って行動するのが安全です。青信号だからといって自動的に発進させるのではなく、交通誘導員からの指示を確認するようにしてください。

VILA-jp (Ours):

9 / 10

いいえ、青信号になっても車を発進させてはいけません。画像には交通整理をしている作業員が写っており、彼が「止まれ」の合図を出しているため、車は停止する必要があります。信号が青であっても、交通整理の指示に従うことが重要です。

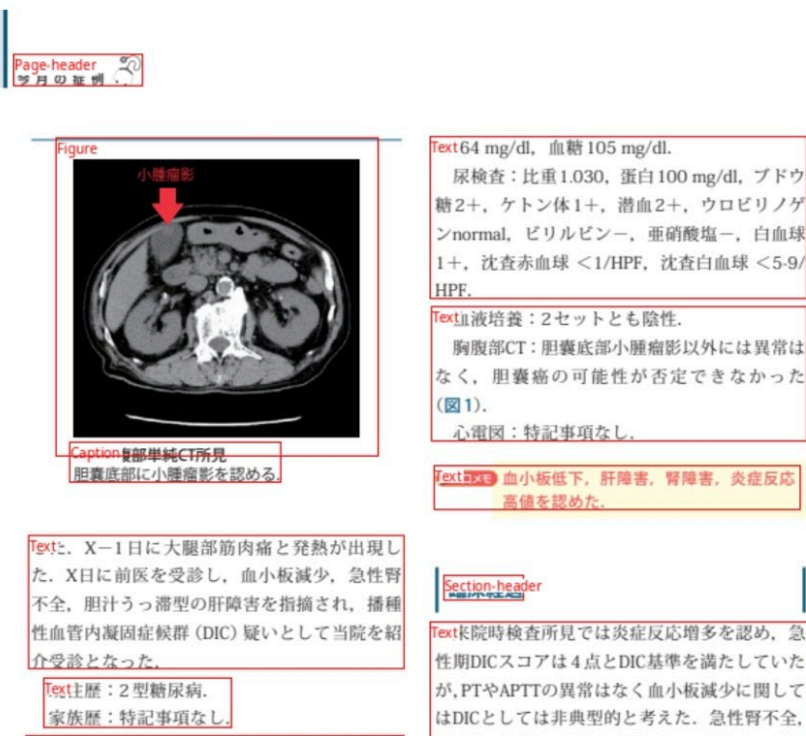
評価結果

モデル	Heron Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
	Judgeスコア (%)	ROUGE-L	Judgeスコア (5点満点)	ROUGE-L	Judgeスコア (5点満点)
<u>Japanese InstructBLIP Alpha (7B)</u>	14.0	20.8	2.42	-	-
<u>Japanese Stable VLM (7B)</u>	24.2	23.3	2.47	-	-
<u>Llama-3-EvoVLM-JP-v2 (8B)</u>	39.3	41.4	2.92	23.5	2.96
<u>LLaVA-CALM2-SigLIP (7B)</u>	43.3	47.2	3.15	17.4	3.21
<u>Llama-3.2-11B-Vision</u>	36.5	27.4	2.77	13.8	2.95
<u>InternVL2-8B</u>	45.2	33.7	2.98	11.6	3.13
<u>Qwen2-VL-7B-Instruct</u>	54.8	45.3	3.53	16.2	3.48
VILA-jp (14B)	57.2	52.3	3.69	16.2	3.62
GPT-4o	87.6	37.6	3.85	12.1	3.58

その他の取り組み：医学分野でのVLMの応用

J-Stageの医学文献からのコーパス構築

- PDF論文からテキストを検出・抽出し、レイアウトの分析、および読み順を解析する
- PDFの処理にSurya^[1]を用いる



レイアウト解析の例[2]

[1] <https://github.com/VikParuchuri/surya>

[2] 小林 俊諒, 豊嶋 弘一, 山田 英嗣, 南 博仁, 藤枝 敦史, 玉木 茂久. Jarisch-Herxheimer反応が診断の端緒となったWeil病の1例, 日本内科学会雑誌, 112巻, 6号, pp. 1005-1011, 2023.

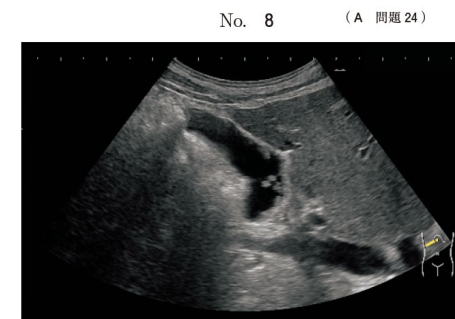
マルチモーダル試験コーパスの構築

- 医師国家試験問題から問題を抽出し、別冊として収録されている画像と対応付ける

24 52歳の女性。2週間前に受けた人間ドックの腹部超音波検査で胆嚢の異常を指摘されたため精査目的で来院した。自覚症状はない。既往歴に特記すべきことはない。身長158 cm、体重64 kg。BMI 25.6。体温36.2℃。腹部は平坦、軟で、圧痛を認めない。血液所見：赤血球458万、Hb 13.7 g/dL、Ht 41 %、白血球7,300。血液生化学所見：総ビリルビン0.9 mg/dL、AST 20 U/L、ALT 18 U/L、LD 148 U/L(基準124~222)、ALP 86 U/L(基準38~113)、 γ -GT 28 U/L(基準9~32)、CEA 1.1 ng/mL(基準5以下)、CA19-9 14 U/mL(基準37以下)。CRP 0.1 mg/dL。腹部超音波像(別冊No. 8)を別に示す。

この患者の治療方針で適切なのはどれか。

- a 経過観察
- b 抗菌薬投与
- c 腹腔鏡下胆嚢摘出術
- d 薬物による抗癌治療
- e 経皮経肝胆嚢ドレナージ



別冊
No. 8

第118回医師国家試験問題 A問題

まとめ

- マルチモーダル基盤モデルは言語や画像など、様々な伝達手段の情報を統合的に扱う
 - 今後、マルチモーダル基盤モデルの研究・開発・公開が進むと予想されている
- マルチモーダル基盤モデルの教育分野への応用はこれから
 - コンテンツ作成、学習者の支援、採点の支援などを「現場」で開発・検証
- 大規模言語モデル開発センターで構築中の画像と言語の基盤モデルを紹介
 - 日本語のベンチマークデータにおいて、比較的良好な性能を示した
 - 今後、イラストを含む試験問題や説明資料の読み取りなどの実タスクに取り組む
 - 共同研究者 (敬称略): 河原、栗田、小田、相澤、金澤、笹川、前田、Zhishen、杉浦
- LLM-jpのマルチモーダルWG
 - マルチモーダル基盤モデルの開発に興味のある方はぜひお声がけください！