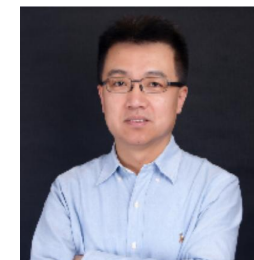


A Multi-Modal Context Reasoning Approach



A Multi-Modal Context Reasoning Approach for Conditional Inference on Joint Textual and Visual Clues

Yunxin Li, Baotian Hu, Xinyu Chen, Yuxin Ding, Lin Ma, Min Zhang



A Multi-Modal Context Reasoning Approach

1. Current VLMs excel at performing multimodal VQA and Reasoning (VCR) for given questions or texts highly attached to the image. However, they usually have inferior performances for conditional inference on joint textual and visual clues, where **the text clue provides the information complementary to the image, or external knowledge**.
2. In contrast, the language models can infer the next-step intent according to the given abstract text information compared to pretrained VLMs, yet they can not understand image information.
3. **Vision-assisted language models** present a feasible method for language models to include visual information.



Premise: [person4] is very friendly.

Actions:

- A. [person2] with a cap wants to sit by the window but [person4] refuses him without any hesitation.
- B. [person2] in a white coat wants to sit by the window but [person4] refuses him without any hesitation.
- ✓ C. [person2] with a cap wants to sit by the window and [person4] exchanges his seat with him generously.
- D. [person2] in a white coat wants to sit by the window and [person4] exchanges his seat with him generously.

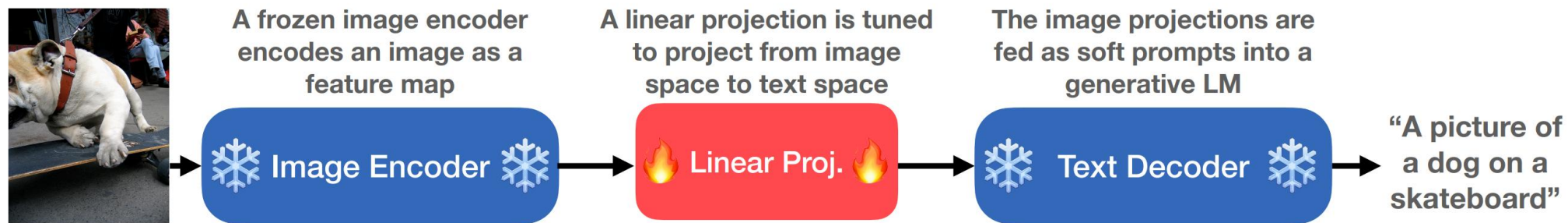
I

II

A Multi-Modal Context Reasoning Approach

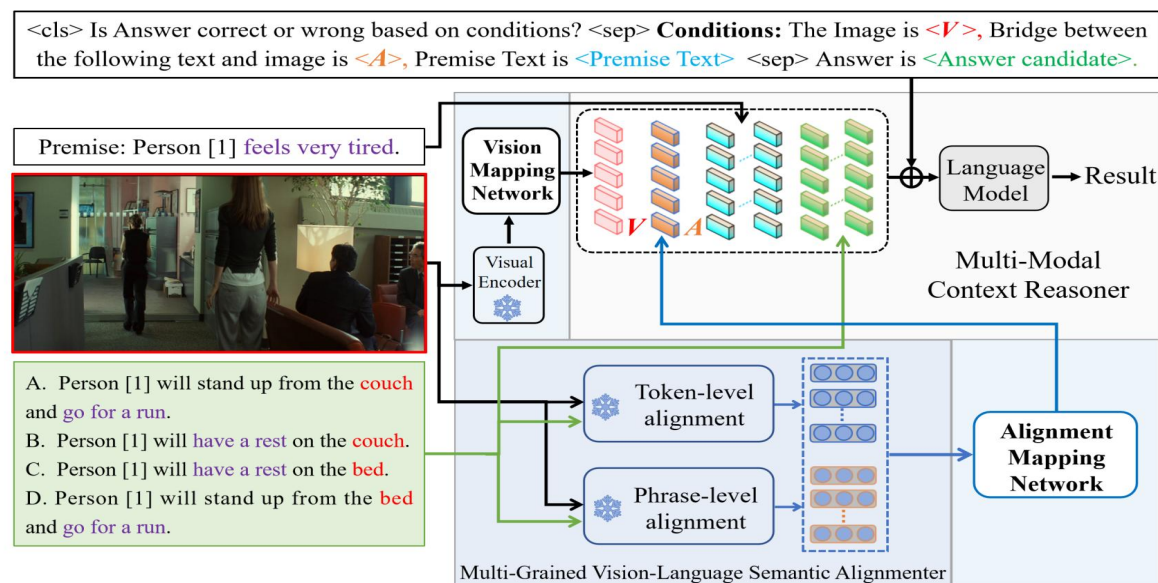
Vision-assisted Language Model

1. It aims to **infuse the visual information into the language model**, boosting the performance of language models on NLP tasks such as open-ended text generation.
2. The images comes from the image corpus or are generated by the powerful text-to-image technical.
3. A simple visual mapping network (**linear project layer or MLP**) is introduced to map the image feature into the langugae space, which usually need to be pretrained via enormous image-text pairs.
4. The language model can incorporate the visual information and perform multimodal tasks via the **visual prefix tuning method**.



A Multi-Modal Context Reasoning Approach

1. We first use the visual encoder to gain the image feature and the text are directly fed into the language model.
2. Similar to previous vision-assisted language models, a **vision mapping network** is introduced to map the image feature into the language space.
3. Considering a semantic gap between visual prefixes and text when the language model performs context learning, we devise an **alignment mapping network based on a multi-grained vision-language semantic aligner** to gain the cross-modal alignment prefix.

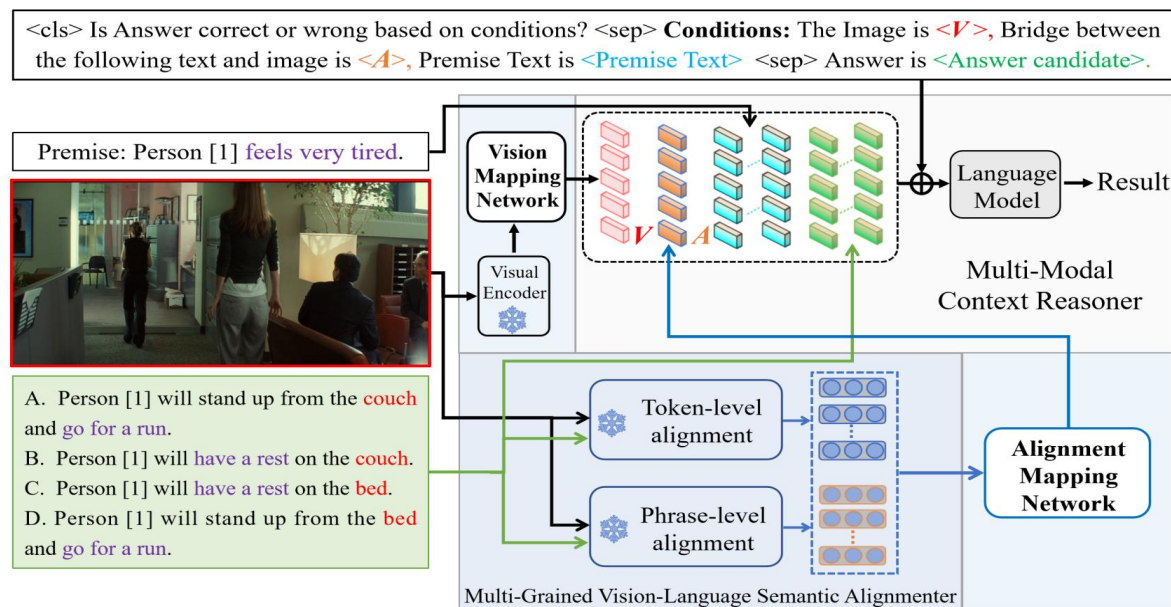


A Multi-Modal Context Reasoning Approach

1. Finally, the two-type prefixes, premise text, and answer candidate are fed into the language model via the **instruction tuning way** to perform multi-modal context reasoning.
2. To make the alignment mapping network capture pivotal multi-view alignment information, we will first train it about one epoch for **alleviating the cold start problem leading to the collapse of the network**.

$$\mathcal{L} = \begin{cases} \mathcal{L}_1, & steps < N_{whole}, \\ \mathcal{L}_f, & steps > N_{whole}, \end{cases}$$

where $steps$ shows the optimization step during training and N_{whole} represents the start of the whole training.



A Multi-Modal Context Reasoning Approach

ModCR achieves state-of-the-art performances on two multimodal reasoning tasks: PMR and VCR.

Method ↓ Types →	Validation	Testing
BERT-B (Devlin et al., 2019)	-	65.2
VL-BERT-B (Lu et al., 2019)	-	75.4
ERNIE-VL-B (Yu et al., 2021a)	-	79.0
UNITER-B (Chen et al., 2020)	-	77.4
Oscar-B (Li et al., 2020)	77.7	76.1
RoBERTa-L (Liu et al., 2019)	77.3	75.0
PromptFuse (Liang et al., 2022)	77.4	76.5
VL-BERT-L (Lu et al., 2019)	-	79.3
ERNIE-VL-L (Yu et al., 2021a)	-	<u>79.9</u>
UNITER-L (Chen et al., 2020)	-	77.0
OFA-L (Wang et al., 2022)	79.9	79.1
MVPTR (Li et al., 2022b)	79.5	78.9
CALeC (Yang et al., 2022)	<u>80.1</u>	78.7
ModCR (frozen VLMs)	85.0	84.3
ModCR (fine-tune VLMs)	85.8	84.7

Model Performance on PMR

Method ↓ Types →	AT ↑	D1 ↓	AF ↓	D2 ↓
BERT-B (Devlin et al., 2019)	65.2	19.8	19.6	4.5
Oscar-B (Li et al., 2020)	76.1	10.2	12.1	1.7
RoBERTa-L (Liu et al., 2019)	75.0	17.7	6.1	1.2
PromptFuse (Liang et al., 2022)	76.5	16.5	5.8	1.2
ERNIE-VL-L (Yu et al., 2021a)	79.9	10.7	8.2	1.2
OFA-L (Wang et al., 2022)	79.1	9.7	9.9	1.3
MVPTR (Li et al., 2022b)	78.9	7.5	11.8	1.8
CALeC (Yang et al., 2022)	78.7	8.6	10.9	1.8
ModCR (frozen VLMs)	84.3	9.2	5.6	0.9
ModCR (fine-tune VLMs)	84.7	7.8	6.8	0.7

Detailed Model Performances on VCR (QR->A)

Method ↓ Types →	Validation	Testing
Oscar-B (Li et al., 2020)	87.3	86.0
RoBERTa-L (Liu et al., 2019)	<u>92.7</u>	<u>91.8</u>
OFA-L (Wang et al., 2022)	90.3	89.4
MVPTR (Li et al., 2022b)	84.2	85.3
CALeC (Yang et al., 2022)	90.8	90.5
ModCR (frozen VLMs)	94.5	93.6
ModCR (fine-tune VLMs)	94.7	94.0


Model Performance on VCR (QR->A)

A Multi-Modal Context Reasoning Approach

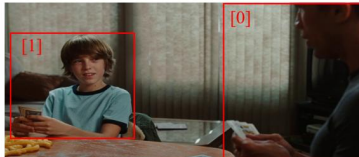
1. Model performances with different lengths of prefix and training strategies.
2. Case study.

Method ↓ Types →	Validation	Testing
CALeC (Yang et al., 2022)	80.1	78.7
RoBERTa-L (Liu et al., 2019)	77.3	75.0
PromptFuse (LV=1, LA=2)	77.4	76.5
ModCR (LV=1, LA=0)	78.1	76.0
ModCR (LV=3, LA=0)	78.2	77.8
ModCR (LV=5, LA=0)	77.3	76.8
ModCR (LV=3, LA=1)	84.9	83.5
ModCR (LV=3, LA=5)	85.8	83.9
ModCR (LV=3, LA=7)	85.3	84.1
ModCR (LV=1, LA=1)	84.0	82.3
ModCR (LV=3, LA=3)	84.8	83.8
ModCR (LV=5, LA=5)	85.0	84.3
ModCR (LV=7, LA=7)	85.1	82.8
ModCR (LV=10, LA=10)	79.7	79.3

MappNet	RoBERTa	VLM	Validation	Testing
✓	×	×	85.7	85.8
✓	✓	×	94.5	93.6
✓	✓	✓	94.7	94.0
✓	×	×	72.2	69.2
✓	✓	×	85.0	84.3
✓	✓	✓	85.8	84.7



Premise: [1] 's personality is very just.
 Prediction: ModCR: **AT**, Oscar/CALeC/OFA/MVPTR: **AF**.
 Answers:
AT: [1] chastises [0] and [2] for missing vital drugs from the refrigerator and **does not** take sides although [2] is a woman. 😊
AF: [1] gives [0] all the heavy work of sorting the medicine in the refrigerator, but only gives [2] some easy work. 😊
 D2: [1] gives [0] all the heavy work of sorting the medicine in **the green refrigerator** , but only gives [2] some easy work. 😊
 D1: [1] chastises [0] and [2] for missing vital drugs from **the green refrigerator** and does not take sides although [2] is a woman. 😊



Premise: [0] and [1] are father and son.
 Prediction: ModCR: **AT**, Oscar/CALeC/OFA/MVPTR: **AF**.
 Answers:
 D2: [0] and [1] in **white shirt** are **brothers**, playing cards together in the room. 😊
AF: [0] and [1] in **blue shirt** are **brothers**, playing cards together in the room. 😊
AT: [0] sitting on the chair and [1] are playing cards together in the room. 😊
 D1: [0] **lying on the bed** and [1] are playing cards together in the room. 😊

Take Home Message

- Large language model is all you need. [Big Model Age](#).
 - Large language models can serve as brain-like actuators, could be attached with human-like eyes and legs. [ToolFormer](#), and [Embodied Intelligence](#).
 - [Incorporating the multi-grained image-text semantic alignment prefix](#) in the language model is very useful for image understanding.
 - Visual Language Model should be improved in terms of [Contextual Reasoning and In-context Learning](#).
-