

# ***MODA: MOdular Duplex Attention for Multimodal Perception, Cognition, and Emotion Understanding***

Zhicheng Zhang<sup>1,2,†</sup>   Wuyou Xia<sup>1</sup>   Chenxi Zhao<sup>1,†</sup>   Yan Zhou<sup>3</sup>   Xiaoqiang Liu<sup>3</sup>  
Yongjie Zhu<sup>3,‡</sup>   Wenyu Qin<sup>3</sup>   Pengfei Wan<sup>3</sup>   Di Zhang<sup>3</sup>   Jufeng Yang<sup>1,2,✉</sup>

<sup>1</sup> Nankai University

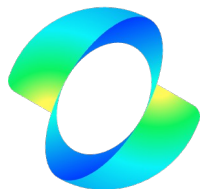
<sup>2</sup> Pengcheng Laboratory

<sup>3</sup> Kuaishou Technology

<sup>†</sup> Work Done at KlingAI

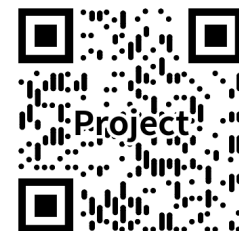
<sup>‡</sup> Project Leader

✉ Corresponding Author

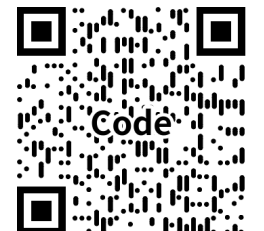


<https://zzcheng.top/MODA>

<https://huggingface.co/KwaiVGI/MODA>



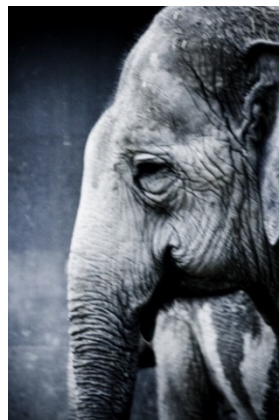
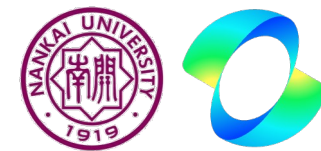
Project



Code

- **Introduction**
- **Rethinking Multimodal LLMs**
- **MODA for Multimodal LLMs**
- **Experimental Results**
- **Conclusion**

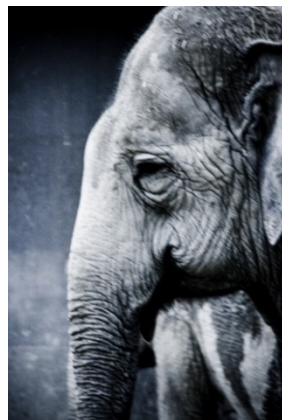
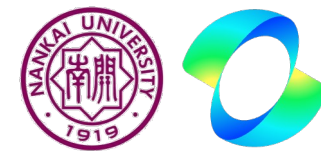
# Introduction



*“A picture is worth a thousand words.”*

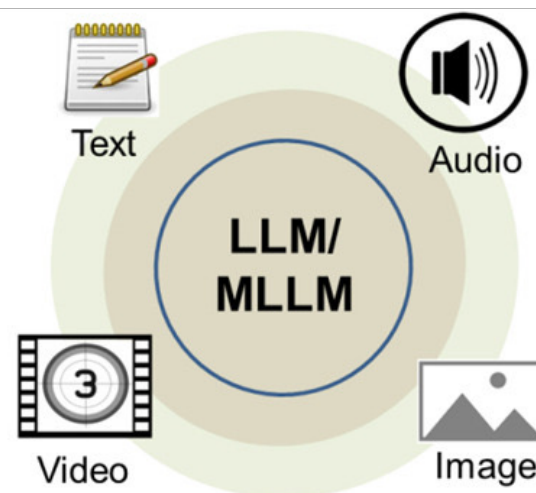


# Introduction



## □ Perception

- Recognition
- Counting
- Detection
- Segmentation

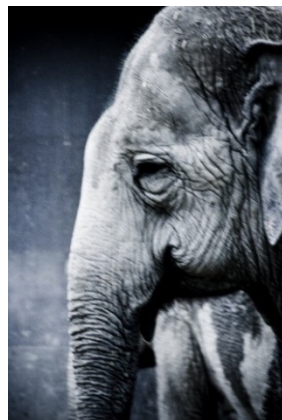
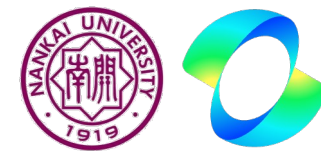


## □ Cognition

- Summarize
- Knowledge
- Memory
- Role-Playing

*The Revolution of Multimodal Large Language Models: A Survey, ACL, 2024*

# Introduction



## □ “*Will it be possible to create sentient robots?*”



- Our answer: It represents a significant step toward Artificial General Intelligence, achievable once robots are equipped with advanced perception and cognitive capabilities.

*125 questions: Exploration and discovery, Science, 2024*

Corleone: A man who doesn't spend time with his family can never be a real man



A conversation from *The Godfather*

SOTA (LLaVA-NeXT): In the dialogue, we see a scene from a film where **two characters** (:three), are engaged in a conversation. The character on the left, who appears to be a man in a suit and tie, is **speaking to the character on the right**, who is wearing a tuxedo with a red flower on the lapel. The character on the right **is looking down** (:looking at Johnny), possibly at a piece of paper or a small object in his hand, and his expression is one of **peace or contemplation** (:gravity).

Response from SOTA MLLM

## □ *Bottleneck of Multimodal LLMs*

### □ Dependence on Perception and Cognition

Emotional responses can be distorted due to inaccurate perception and cognitive processing.

### □ Modality Bias Induced by Pretraining

The model tends to over-rely on verbal cues, causing a bias toward the language modality and overlooking fine-grained nonverbal signals.



# Introduction

## □ *Bottleneck of Multimodal LLMs*

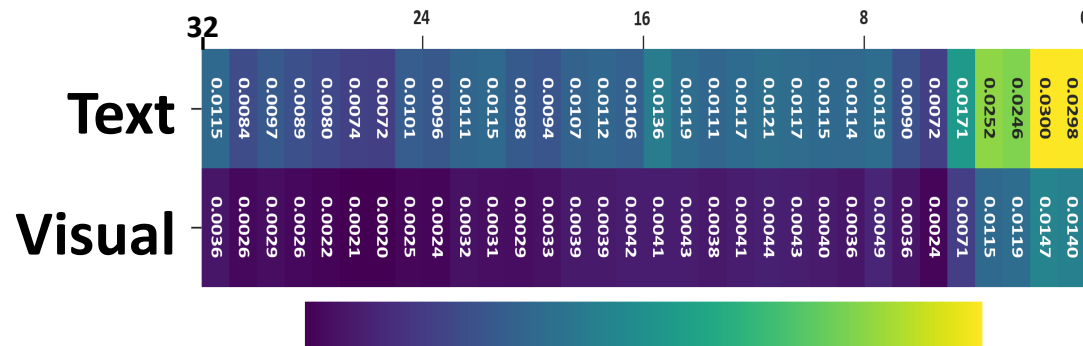
### □ Dependence on Perception and Cognition

Emotional responses can be distorted due to inaccurate perception and cognitive processing.

### □ Modality Bias Induced by Pretraining

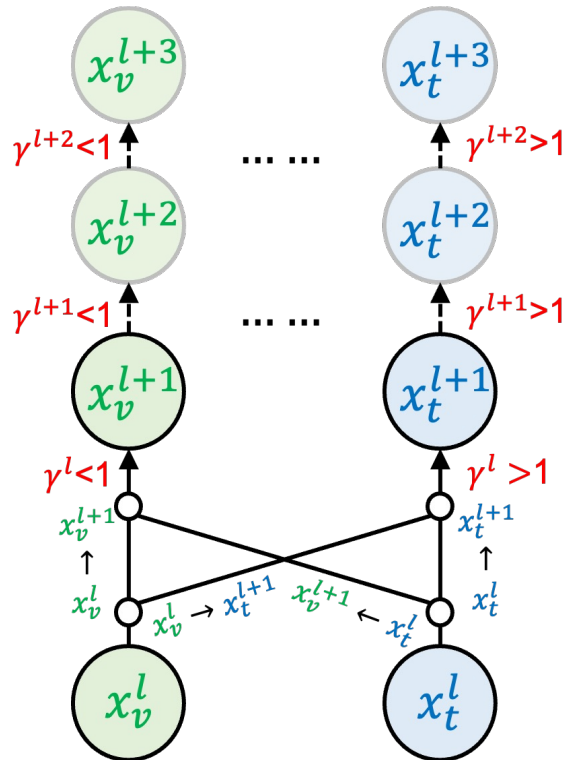
The model tends to over-rely on verbal cues, causing a bias toward the language modality and overlooking fine-grained nonverbal signals.

## Layer-wise Attention Score

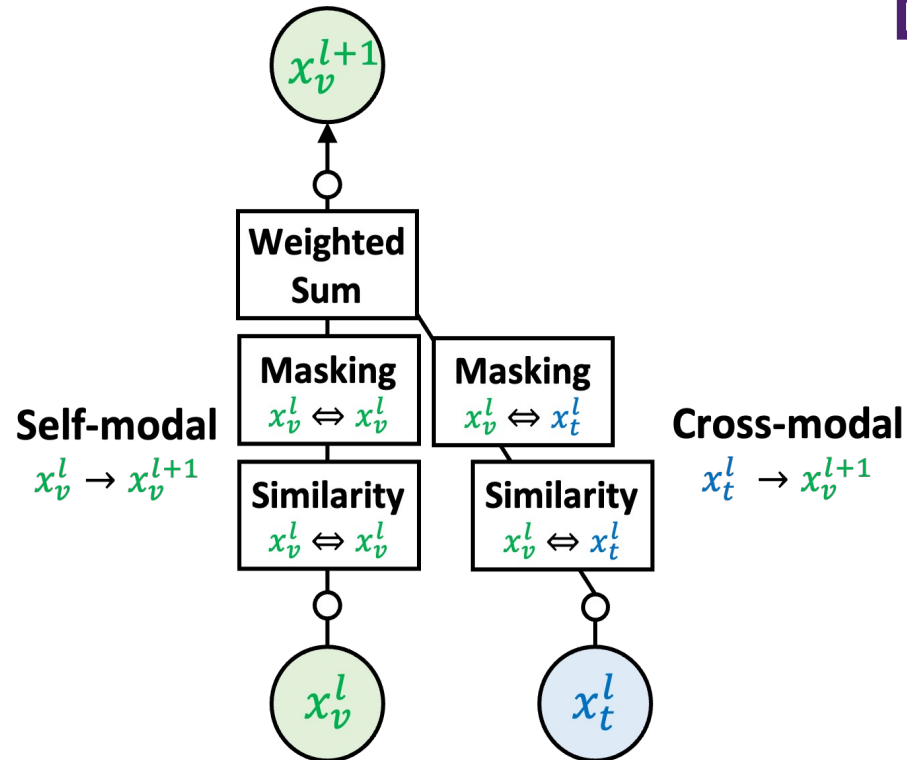


Modality Bias of SOTA MLLM

# Rethinking Previous Works



(a) Graphic diagram of DDA



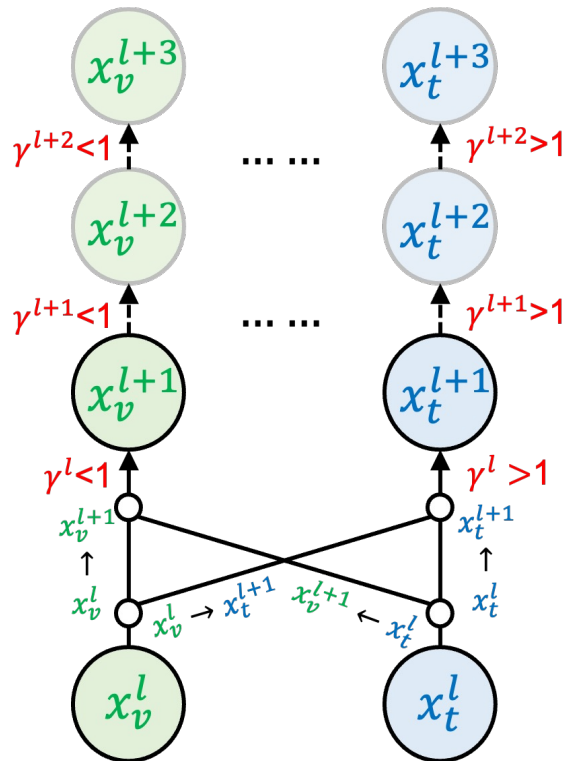
(b) Multimodal attention in the intermediate layer

## □ Attention in MLLMs

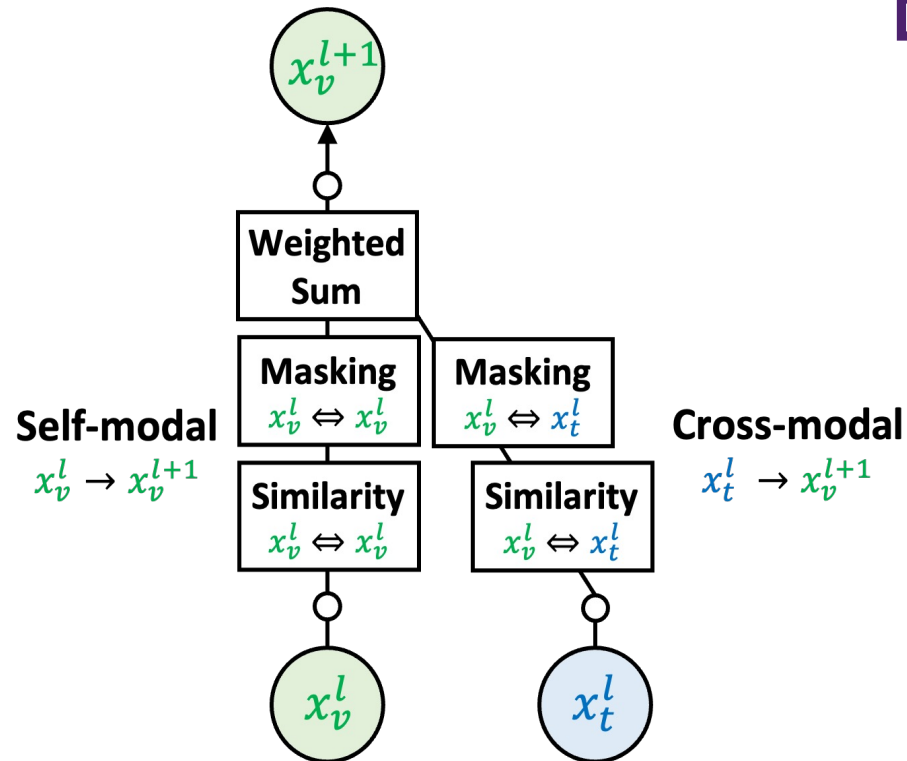
- Token Flow controlled by QK similarity and masking.
- Multimodal Token Flow Across Layers Guided by Cross-Modal and Self-Modal Attention



# Rethinking Previous Works



(a) Graphic diagram of DDA



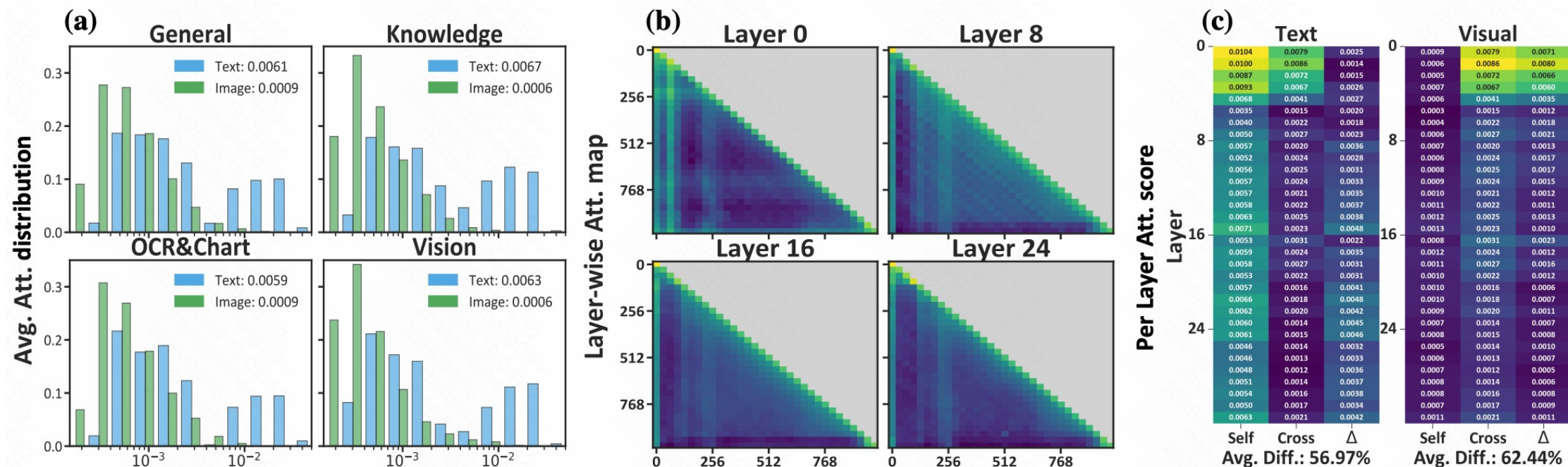
(b) Multimodal attention in the intermediate layer

## □ Disorder Deficit Attention

- Incorrect Focus Due to Inaccurate Similarity and Masking
- Errors Propagate Layer by Layer

**Ignore 19%**  
**Key Concepts**

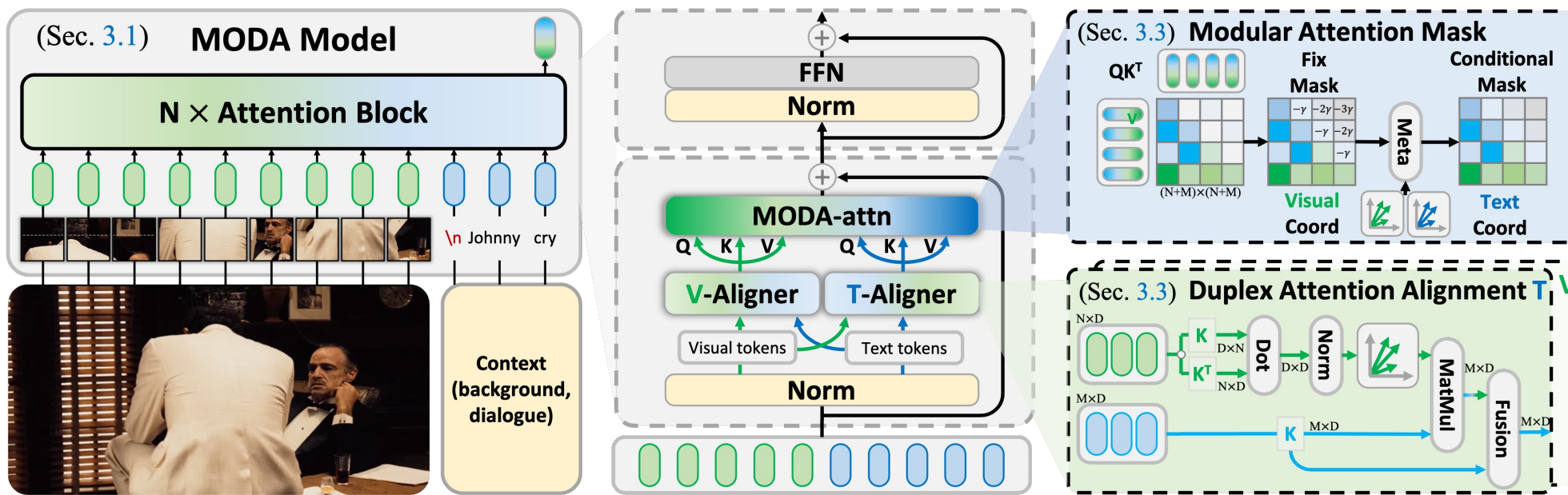
# Rethinking Previous Works



## Key Observation from Experiments

- (a):** The attention allocated to visual content is significantly weaker than that for the textual modality, with a difference of considerable magnitude.
- (b)&(c):** A distinct cross-attention bias in the lower layers of the model across its 32 layers

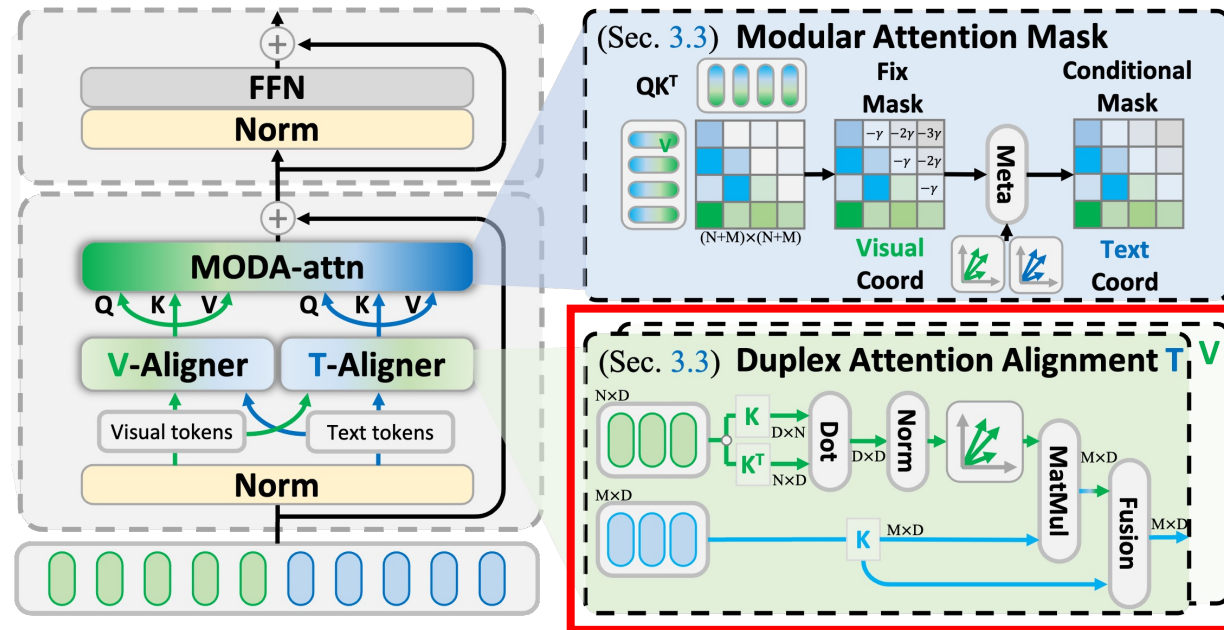
# MODA



## ✓ Contributions

- ❑ From the perspective of attention shift, we identify the key bottleneck in MLLMs and reveal its core cause.
- ❑ A modular and duplex attention mechanism (MODA) for MLLM.
- ❑ Effectiveness for **fine-grained content understanding** in 21 benchmarks.

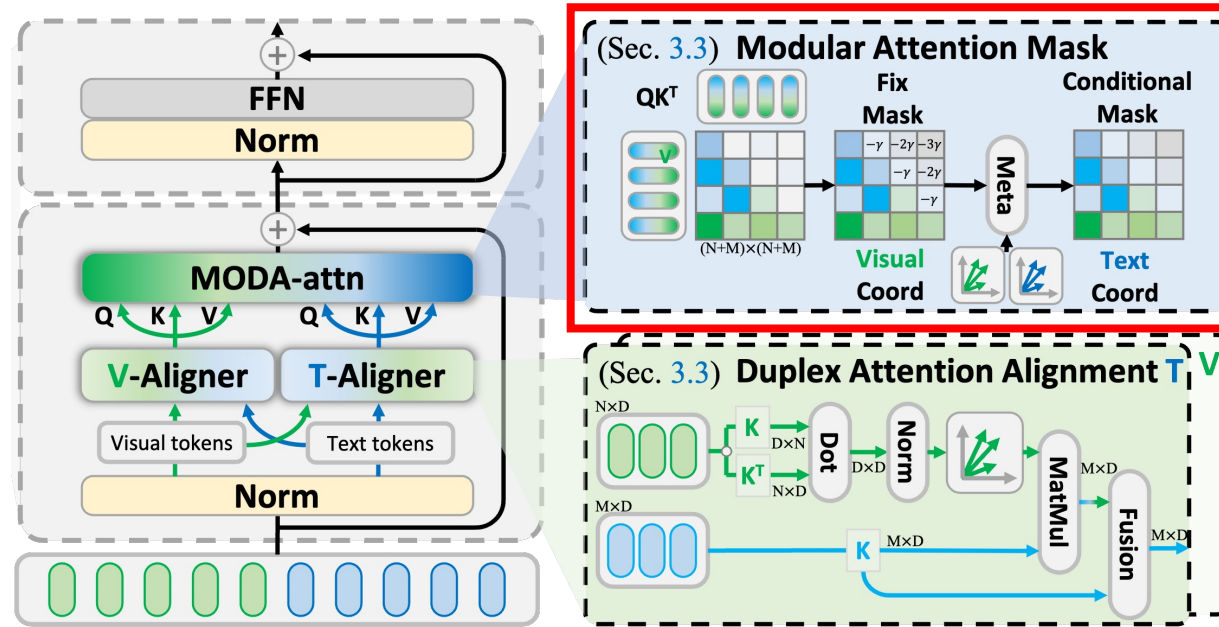




## ✓ Duplex Attention Alignment

- ❑ Modality Space Mapping. **Visual Coordinate** & **Text Coordinate**
- ❑ Kernelized Cross-Modal Transformation. (2 Gram Matrixes)

$$G_{ij}^m = \sum_{k=1}^{N_m} K_{ik}^m K_{kj}^m = K^{m\top} K^m, \quad K^{\bar{m} \rightarrow m} = K^{\bar{m}} \|G^m\|,$$



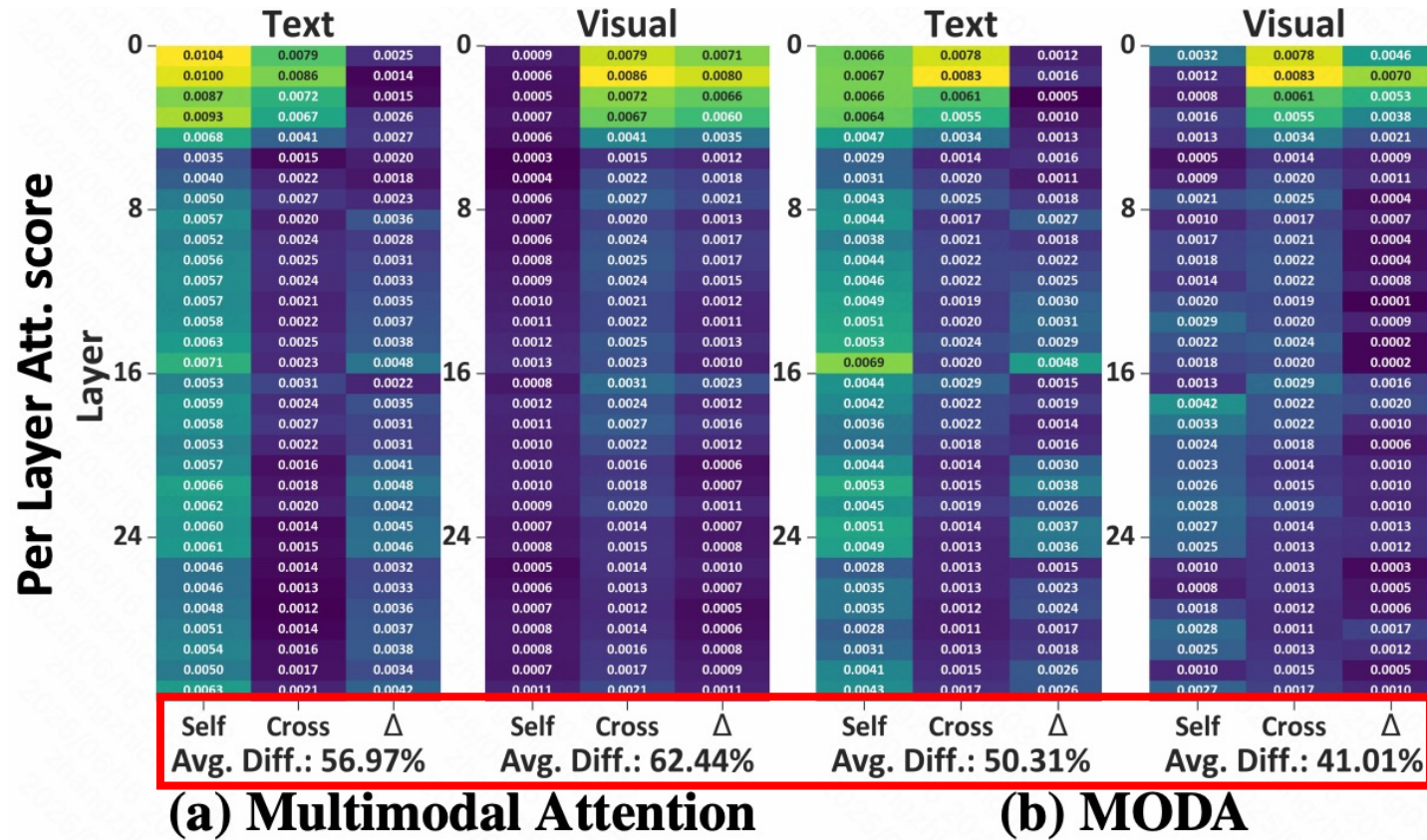
## ✓ Modular Attention Mask

- integrates coordinate information from both visual and text modalities and employs two types of masks — a **Fix Mask** and a **Conditional Mask**.

$$A_{MM} = \begin{pmatrix} q_1 k_1^T & p_{11} & \cdots & p_{1(n-1)} \\ q_2 k_1^T & q_2 k_2^T & \cdots & p_{1(n-2)} \\ \vdots & \vdots & \ddots & \vdots \\ q_n k_1^T & q_n k_2^T & \cdots & q_n k_n^T \end{pmatrix}$$

$$p_{base} = 0, p_{ij} = p_{base} - (j - 1)\beta$$

# Experimental Results



□ **DDA:** MODA can effectively **alleviate the problem of disorder deficit attention**, reducing the attention discrepancy at each model layer from **56% and 62% to 50% and 41%**, respectively.



# Experimental Results

Model	Method	General					Knowledge					OCR & Chart					Vision-Centric				
		Avg	MME <sup>P</sup>	MMB	SEED <sup>I</sup>	GQA	Avg	SQA <sup>I</sup>	MMMU <sup>V</sup>	MathVista <sup>M</sup>	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	CV-Bench <sup>2D</sup>	CV-Bench <sup>3D</sup>
	GPT-4V	63.0	1409.4	75.8	69.1	36.8	65.2	75.7	56.8	49.9	78.2	77.4	78.5	64.5	78.0	88.4	62.4	50.0	61.4	64.3	73.8
	Gemini-1.0 Pro	-	1496.6	73.6	70.7	-	-	79.5	47.9	45.2	-	-	-	65.9	-	-	-	-	-	-	-
	Gemini-1.5 Pro	-	-	-	-	-	-	-	58.5	52.1	80.3	-	81.3	-	73.5	86.5	-	-	67.5	-	-
	Grok-1.5	-	-	-	-	-	-	-	53.6	52.8	88.3	-	76.1	-	78.1	85.6	-	-	68.7	-	-
	MM-1-8B	-	1529.3	72.3	69.9	-	-	72.6	37.0	35.9	-	-	-	-	-	-	-	-	-	-	-
	MM-1-30B	-	1637.6	75.1	72.1	-	-	81.0	44.7	39.4	-	-	-	-	-	-	-	-	-	-	-
<b>Base LLM: Llama-3-Ins-8B</b>																					
	Mini-Gemini-HD-8B	72.7	<b>1606.0</b>	72.7	73.2	64.5	55.7	75.1	37.3	37.0	73.5	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
	LLaVA-NeXT-8B	72.5	1603.7	72.1	72.7	<b>65.2</b>	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
	Cambrian-1-8B	<b>73.1</b>	1547.1	<b>75.9</b>	74.7	64.6	61.3	<b>80.4</b>	42.7	<b>49.0</b>	73.0	71.3	73.3	62.4	<b>71.7</b>	77.8	65.0	51.3	<b>64.2</b>	72.3	72.0
	MODA-8B	72.1	1535.9	73.8	<b>74.9</b>	63.0	<b>61.5</b>	<b>80.4</b>	<b>43.1</b>	48.8	<b>73.6</b>	<b>72.0</b>	<b>74.3</b>	<b>65.2</b>	70.4	<b>78.1</b>	<b>66.0</b>	<b>52.6</b>	64.1	<b>73.5</b>	<b>73.8</b>
<b>Base LLM: Hermes2-Yi-34B</b>																					
	Mini-Gemini-HD-34B	76.2	1659.0	80.6	75.3	65.8	62.4	77.7	48.0	43.4	80.5	68.1	67.6	51.8	74.1	<b>78.9</b>	63.8	37.3	67.2	71.5	79.2
	LLaVA-NeXT-34B	76.0	1633.2	79.3	<b>75.9</b>	<b>67.1</b>	62.5	81.8	46.7	46.5	74.9	67.7	68.7	54.5	69.5	78.1	64.0	47.3	61.0	73.0	74.8
	Cambrian-1-34B	<b>76.8</b>	<b>1689.3</b>	81.4	75.3	65.8	67.0	85.6	49.7	53.2	79.7	71.9	75.6	60.0	76.7	75.5	68.5	52.7	67.8	74.0	79.7
	MODA-34B	76.7	1639.2	<b>82.3</b>	75.8	66.2	<b>69.5</b>	<b>88.1</b>	<b>52.5</b>	<b>54.0</b>	<b>83.4</b>	<b>74.7</b>	<b>79.8</b>	<b>62.7</b>	<b>78.3</b>	78.2	<b>69.9</b>	<b>53.8</b>	<b>68.5</b>	<b>75.8</b>	<b>81.3</b>

□ **SOTA comparison:** In the parameter ranges of **8B** and **34B**, MODA achieved consistency improvements, which can seamlessly replace the existing attention mechanisms and enhance MLLM performance.

# Experimental Results

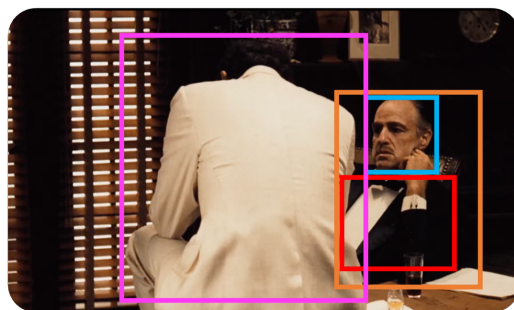
Model		Cognition							
Method	Avg	Instruction Adherence	Fluency	Coherency	Image-Text relevance	Response Accuracy	Personality Consistency	Knowledge Consistency	Tone Consistency
GPT-4 Turbo	1.099	1.055	1.032	1.084	1.097	1.092	1.168	1.103	1.161
Gemini 1.0 Pro	1.021	0.999	1.007	1.028	1.009	1.013	1.052	1.013	1.050
Claude 3 Opus	1.157	1.127	1.070	1.149	1.167	1.146	1.219	1.168	1.213
QWen-VL-Max	1.028	1.014	1.012	1.035	1.034	1.029	1.042	1.021	1.041
<i>Base: Llama-3-Ins-8B</i>									
Mini-Gemini-HD-8B	0.878	0.884	0.942	0.898	0.864	0.853	0.855	0.876	0.852
LLaVA-NeXT-8B	0.968	0.971	0.988	0.980	0.966	0.967	0.966	0.964	0.939
Cambrian-1-8B	0.895	0.901	0.957	0.934	0.886	0.889	0.860	0.892	0.838
MODA-8B	<b>0.972</b>	<b>0.976</b>	<b>0.992</b>	<b>0.985</b>	<b>0.970</b>	<b>0.972</b>	<b>0.970</b>	<b>0.969</b>	<b>0.945</b>
<i>Cognition-Specialized</i>									
MMRole-9B	0.994	0.998	1.000	0.997	<b>0.993</b>	0.987	1.000	<b>0.992</b>	<b>0.988</b>
MODA-8B	<b>0.995</b>	<b>1.000</b>	<b>1.001</b>	<b>0.999</b>	<b>0.993</b>	<b>0.989</b>	<b>1.001</b>	0.991	<b>0.988</b>
MMRole-9B (In-Test)	0.999	1.000	1.000	0.999	0.997	0.989	1.012	<b>0.997</b>	<b>0.997</b>
MODA-8B (In-Test)	<b>1.000</b>	<b>1.002</b>	<b>1.001</b>	<b>1.000</b>	<b>0.998</b>	<b>0.992</b>	<b>1.013</b>	0.996	0.996
MMRole-9B (Out-Test)	0.981	0.992	0.999	0.993	0.979	0.981	0.963	0.977	0.962
MODA-8B (Out-Test)	<b>0.984</b>	<b>0.995</b>	<b>1.002</b>	<b>0.996</b>	<b>0.981</b>	<b>0.983</b>	<b>0.970</b>	<b>0.980</b>	<b>0.965</b>

Model		Emotion							
Method	Avg	MVSA <sup>S</sup> (ACC)	MVSA <sup>S</sup> (F1)	MVSA <sup>M</sup> (ACC)	MVSA <sup>M</sup> (F1)	TumEmo (ACC)	TumEmo (F1)	HFM (ACC)	HFM (F1)
GPT-4V	0.633	0.507	0.570	0.609	0.631	0.608	0.612	0.764	0.765
Gemini 1.0 Pro	0.646	0.634	0.637	0.699	0.657	0.598	0.582	0.674	0.683
Claude 3 Opus	0.628	0.626	0.613	0.635	0.629	0.580	0.574	0.679	0.687
QWen-VL-Max	0.643	0.647	0.645	0.669	0.627	0.565	0.595	0.696	0.701
<i>Base: Llama-3-Ins-8B</i>									
Mini-Gemini-HD-8B	0.482	0.423	0.571	0.487	<b>0.643</b>	0.246	0.395	0.498	<b>0.593</b>
LLaVA-NeXT-8B	0.576	0.591	0.593	0.617	0.607	0.547	0.533	0.572	0.551
Cambrian-1-8B	0.547	0.694	0.661	0.662	0.579	0.439	0.344	0.512	0.487
MODA-8B	<b>0.588</b>	<b>0.702</b>	<b>0.705</b>	<b>0.628</b>	0.619	<b>0.559</b>	<b>0.548</b>	<b>0.585</b>	0.563
<i>Emotion-Specialized</i>									
M <sup>2</sup> CL	-	0.755	0.742	0.732	0.705	0.688	0.687	-	-
MULSER	-	0.757	0.755	0.739	0.738	0.775	0.775	-	-
CMGCN	-	0.733	0.720	0.697	0.683	-	-	0.875	0.841
SPFVTE	-	0.806	0.801	0.799	0.788	-	-	0.883	0.879
MODA-8B	<b>0.841</b>	<b>0.810</b>	<b>0.803</b>	<b>0.802</b>	<b>0.790</b>	<b>0.778</b>	<b>0.778</b>	<b>0.885</b>	<b>0.881</b>

□ **High-level Emotion Tasks.** Besides perception and cognition, MODA performs well on high-level emotion tasks, especially in sarcasm detection that require fine-grained content understanding, surpassing existing models and the 50-50 random guessing baseline.

# Experimental Results

**Johnny:** A month ago he bought the rights to this book, a best seller. The main character is a guy just like me. I wouldn't even have to act, just be myself. Oh Godfather, I don't know what to do, I don't know what to do [crying]



**Human:** How is the Godfather looks like?

**Cambrian-1:** The Godfather, as depicted in the image, is portrayed by an actor in a classic black tuxedo, which is a traditional and iconic look for the character.

**MODA:** The Godfather is **dressed in a classic black tuxedo**, which is emblematic of his role as a **powerful and respected figure** in the film. **His expression is one of concern, as he listens to Johnny's dilemma**, indicating his deep involvement in Johnny's life and the weight of his responsibilities.

**Human:** What does this scene reveal about the Godfather's influence over people in his inner circle?

**Cambrian-1:** This scene from 'The Godfather' illustrates the profound impact the character, Don Corleone, has on those around him. It shows how his influence extends beyond his immediate family, affecting even those who are not directly related to him.

**MODA:** This scene reveals the Godfather's influence over people in his inner circle as a significant overwhelming force. The Godfather's **presence in the room**, along with **his stern demeanor**, further emphasizes his control over the situation and the characters around him. **The scene captures the tension and emotional turmoil that can arise from being under the Godfather's influence**, highlighting the power dynamics at play within the narrative.

❑ **MODA-enabled application:** With MODA, MLLM can analyze micro-expressions, intonation, and cultural background features in real time to build multidimensional character profiles.



# Experimental Results

**Human:** How might the Godfather act to proceed with the next conversation? Give me a few possible options.

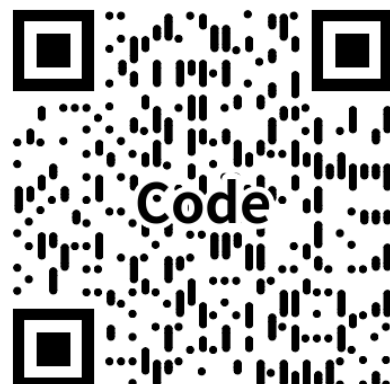
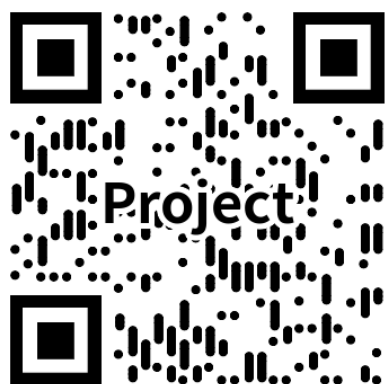
**Cambrian-1:** Given the context of the image, which suggests a scene from "The Godfather," the character portrayed by the man in the foreground might proceed with the **next conversation** by either **continuing to listen** intently to the other character, who seems to be speaking, or by **interjecting with a question or comment** to steer the conversation in a particular direction.

**MODA:** The Godfather, upon witnessing Johnny's distress, might respond with a **combination of empathy and strategic thinking**. He could offer a comforting word, such as, "It's a lot to take in, but we'll figure it out together." Alternatively, he might shift the conversation to a more **practical matter**, like **discussing the logistics of the book's adaptation or the potential benefits of the project**. Another possibility is that he might offer **advice on how to handle the situation**, such as seeking legal counsel or consulting with other members of the family. The Godfather's response would likely be measured and thoughtful, reflecting his role as a leader.



❑ **MODA-enabled application:** By deeply understanding character motivations and emotional contexts, the system can automatically plan dialogue strategies.

# ***MODA: MOdular Duplex Attention for Multimodal Perception, Cognition, and Emotion Understanding***



*Thank you*

<https://zzcheng.top/MODA>  
<https://huggingface.co/KwaiVGI/MODA>