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

Zhicheng Zhang [1,2] Wuyou Xia [1] Chenxi Zhao [1,2] Yan Zhou [3] Xiaoqiang Liu [3] Yongjie Zhu [3] Wenyu Qin [3] Pengfei Wan [3] Di Zhang [3] Jufeng Yang [1,2]

[1]Nankai University [2]Pengcheng Lab [3]Kuaishou Technology

ICML International Conference On Machine Learning

Introduction

- **High-Level Tasks Toward AGI**: Building on the advanced capabilities of MLLMs in perception and cognition, emotion understanding becomes a crucial next step for achieving fine-grained multimodal comprehension and driving progress toward Artificial General Intelligence (AGI).
- **MLLMs Struggle with Capturing Fine-Grained Cues**: However, As highlighted in Fig.1, MLLMs exhibit an inability to capture fine-grained cues, such as character eyesights, leading to errors in emotion understanding. This results in SOTA MLLMs achieving only 50:50 accuracy in sarcasm detection tasks, as revealed by benchmark evaluations.
- **Analysis: Attention as the Core Reason**: Our analysis reveals that multimodal tokens mixed by attention in MLLMs are the primary cause of performance issues.

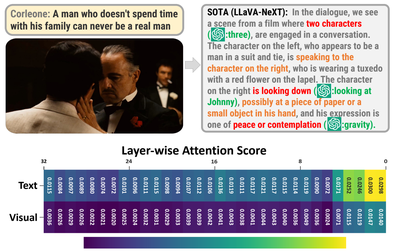

Figure 1. **Bottlenecks of MLLMs in High-Level Tasks.** (tl) Given the detailed image and lines from *The Godfather*, (tr) we highlight incorrect responses, corresponding hallucinated explanations, and attached answers. (b) We visualize attention score across layers, highlighting inconsistent attention across modalities.

Challenge: Disorder Deficit Attention in MLLM

- **Observation on Low Layers**: The attention allocated to visual content is significantly weaker than that for the textual modality, with a difference of considerable magnitude.
- **Observation on High Layers**: A distinct cross-attention bias in the lower layers of the model across its 32 layers.
- **Formulas of Disorder Deficit Attention Problem**: Given the visual tokens x_v^l and text tokens x_t^l in the block l, the multimodal attention builds the link from two parts (*i.e.*, self-modal $x_t^l \to x_t^{l+1}, x_v^l \to x_v^{l+1}$ and cross-modal $x_t^l \to x_v^{l+1}, x_v^l \to x_t^{l+1}$), where the links are commonly implemented by the pair-wise token similarity and weighted sum. However, the modality gap between tokens decrease the magnitude of links, as we observed, the link value of $x_v^l \to x_v^{l+1}$ and $x_t^l \to x_v^{l+1}$ decays exponentially with depth ($\alpha_{v,t\to v}^l \propto \gamma^l, \gamma \neq 1$). This misalignment propagates layer-wise, causing the cumulative error as $\mathbb{E}_{DDA} = \prod_l \gamma^l \epsilon_l$, where ϵ_l denotes the layer-specific alignment error.
- **Discussion**: This aligns with insights from Dong et al. (ICML 2021), which identify rank collapse in pure attention as a key factor worsening imbalance in attention distribution.

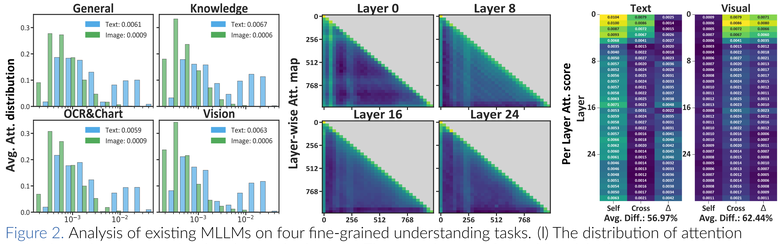

Figure 2. Analysis of existing MLLMs on four fine-grained understanding tasks. (l) The distribution of attention activation values among visual and textual tokens. (c) The attention map for multimodal tokens among stages. (r) The self- and cross-modal attention activation scores with their disparity among the attention layers.

MODA: Modular Duplex Attention

When the gap across modalities arises, we propose to align the tokens from multiple modalities in the attention, which we call modular duplex attention (MODA). MODA first splits multimodal attention into the modality alignment part and the token focus correction part.

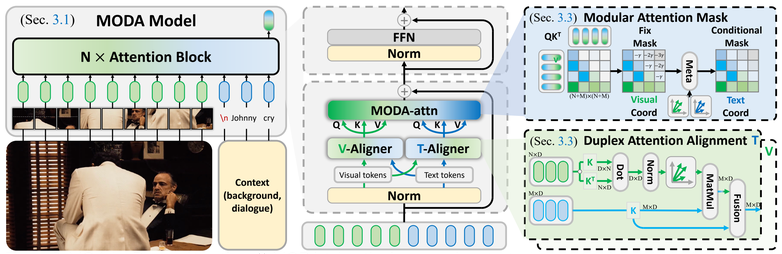

Figure 3. MOdular Duplex Attention. (l) MODA takes the image and contextual prompt as input, including the background and history of the conversation. (c) With MODA, the token flows are justified in each Transformer block of MLLM. MODA modifies the deficient attention scores in a correct-after-align manner via (tr) Modular masked attention and (br) Duplex (V/T)-aligner.

Duplex Attention Alignment

- **Modality Basis Construction**: For the m^{th} modality, the space bases are given via the normed gram matrix $||\boldsymbol{G}^m|| \in \mathbb{R}^{d\times d}$, where $\boldsymbol{G}_{ij}^m$ is the inner product between tokens i and j:

$$\boldsymbol{G}_{ij}^m = \sum_{k=1}^{N_m} \boldsymbol{K}_{ik}^m \boldsymbol{K}_{kj}^m = \boldsymbol{K}^{m\top} \boldsymbol{K}^m, \tag{1}$$

- **Cross-Modal Token Mapping**: The normed Gram matrix serves as a kernelized mapping function $f : \mathbb{R}^d \to \mathbb{R}^d$, enabling token transformation across modalities. Mapped tokens are computed as $\boldsymbol{K}^{\bar{m}\to m} = \boldsymbol{K}^{\bar{m}}||\boldsymbol{G}^m||$, where $\boldsymbol{K}^{\bar{m}}$ represents the value representation from other modalities $\bar{m}$.

Modular Attention Masking

- **Self- and Cross-Modality Attention**: We modular and assign attention mask into $\boldsymbol{M}^m$ and $\boldsymbol{M}^{\bar{m}}$ responsible for self- and cross-modality, respectively.

$$\boldsymbol{O}_{self} = \text{Softmax}(\frac{\boldsymbol{Q}^m \boldsymbol{K}^{m\top}}{\tau} + \boldsymbol{M}^m)\boldsymbol{V}^m, \tag{2}$$

$$\boldsymbol{O}_{cross} = \text{Softmax}(\frac{\boldsymbol{Q}^m \boldsymbol{K}^{\bar{m}\top}}{\tau} + \boldsymbol{M}^{\bar{m}})\boldsymbol{V}^{\bar{m}}. \tag{3}$$

- **Pseudo-Attention Masking**: A modular attention mask is introduced to store unnecessary attention values as pseudo-attention scores.

$$A_{MM} = \begin{pmatrix} \boldsymbol{q}_1\boldsymbol{k}_1^\top & p_{11} & \cdots & p_{1(n-1)} \\ \boldsymbol{q}_2\boldsymbol{k}_1^\top & \boldsymbol{q}_2\boldsymbol{k}_2^\top & \cdots & p_{1(n-2)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{q}_n\boldsymbol{k}_1^\top & \boldsymbol{q}_n\boldsymbol{k}_2^\top & \cdots & \boldsymbol{q}_n\boldsymbol{k}_n^\top \end{pmatrix} \tag{4}$$

$$p_{base} = 0, p_{ij} = p_{base} - (j-1)\beta \tag{5}$$

- **Modality Location Priors**: The normed Gram matrix guides the model to refine self-modal attention (O_{self}) and enable cross-modal attention (O_{cross}) for efficient interaction.

Experiment: Calibration Results on Multimodal Attention

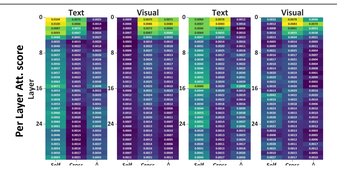

Figure 4. **Analysis of multimodal attention and MODA.** (l) Multimodal attention: average difference between self and cross-modal attention is 56.97% for text and 62.44% for visual. (r) MODA: average difference is 50.31% for text and 41.01% for visual.

Experiment: Perception, Cognition and Emotion

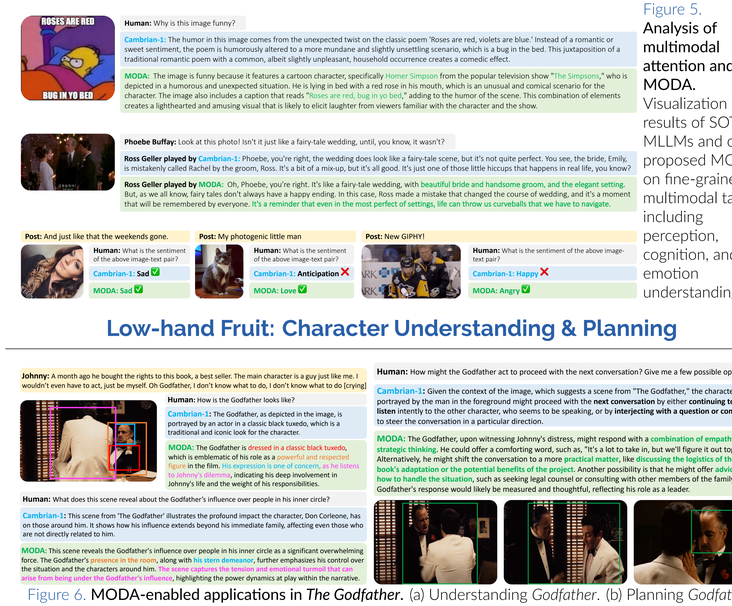

Figure 5. **Analysis of multimodal attention and MODA.** Visualization results of SOTA MLLMs and our proposed MODA on fine-grained multimodal tasks, including perception, cognition, and emotion understanding.

Low-hand Fruit: Character Understanding & Planning

Figure 6. **MODA-enabled applications in *The Godfather*.** (a) Understanding *Godfather*. (b) Planning *Godfather*.

Conclusion

- From the perspective of attention shift, We identify the attention bottleneck in SOTA MLLMs and propose a modular duplex attention mechanism based on our observation.
- We investigate a new MLLM for perception, cognition, and emotion, enabling applications in fine-grained understanding and planning.

Contact
MODA
gloryzzc6@sina.com

Website Project Code