# The Role of Contrastive Learning in Multimodal Large Language Models (MLLMs)

*Presented by-*
***Aditi Sarker***
*Access ID - hq1351*
***Advisor - Dr. Prashant Khanduri***
*Dept. of Computer Science*
Optimization for Large Scale Machine Learning Lab (OptLML)

WAYNE STATE UNIVERSITY

# Outline

- **Introduction**
- **Contrastive Learning**
- **Literature Review**
- **Challenges**
- **Current and Future Work**
- **Conclusion**

# Introduction

# Introduction

- **LLMs (Large Language Models)** are advanced AI models trained on vast text data to generate **language responses**

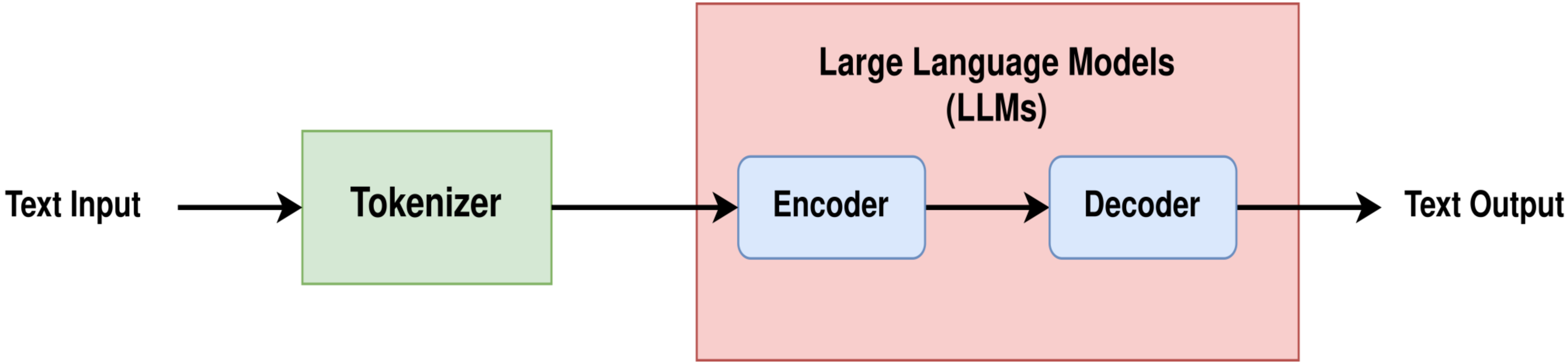



Figure 1: Architecture of LLM

# Introduction(cont...)

- **Multimodal Large Language Models (MLLMs)** are AI models designed to process multiple data types (text, images, audio, video) for advanced reasoning and decision-making

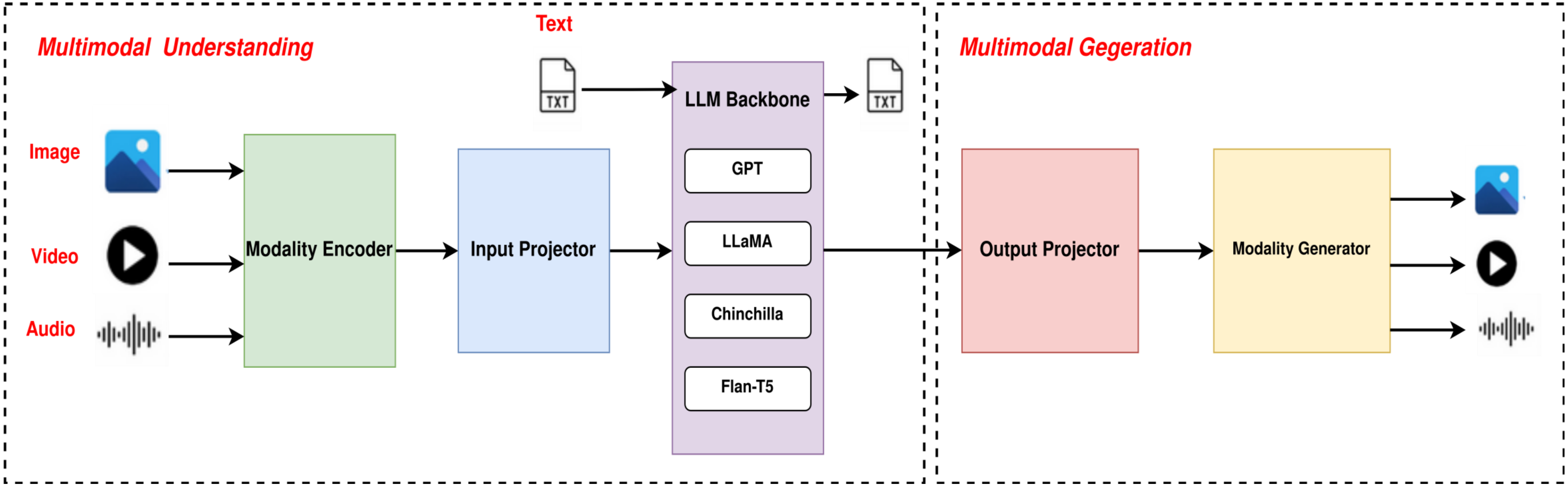


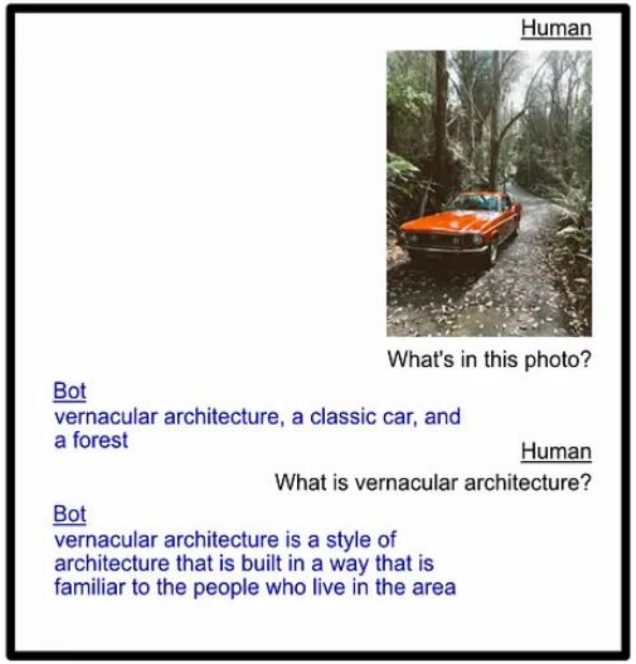

Figure 2: Architecture of Multimodal LLM [image idea taken from here]

Zhang D, Yu Y, Dong J, Li C, Su D, Chu C, Yu D. Mm-llms: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601. 2024 Jan 24.

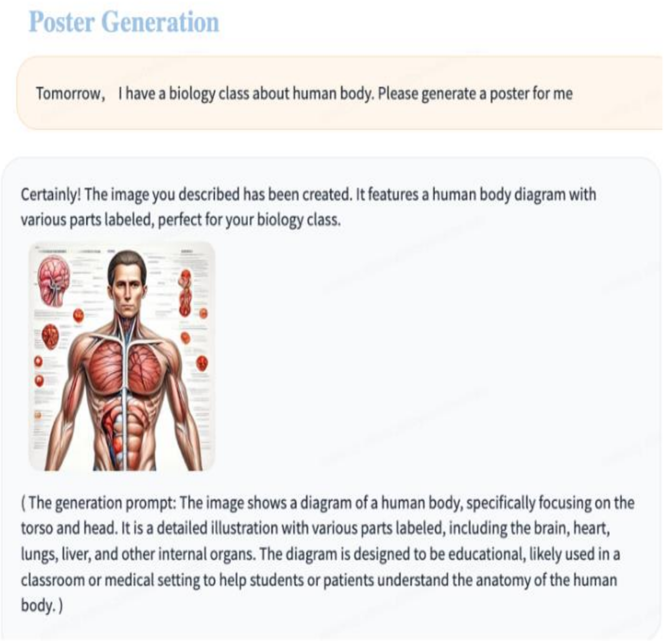
# Introduction(cont...)

## **Applications** of Multimodal Large Language Models (MLLMs)


**Image Captioning**




**Visual Question Answering**




**Multimodal assistants**

# Introduction (cont...)

- One of the fundamental **challenge** in MLLMs is **achieving effective alignment** between different modalities

- The most popular approach to addressing this challenge is **contrastive learning**, a self-supervised learning method that enhances modality alignment

- Models such as CLIP [1], ALIGN [2], and BLIP [3] leverage contrastive learning to **solve multimodal tasks**

# Contrastive Learning

# Contrastive Learning

## What is Contrastive Learning?

- A self-supervised technique that learns representations by distinguishing between similar and dissimilar pairs
- **Positive and Negative Pair Sampling:** Determines how data points are contrasted

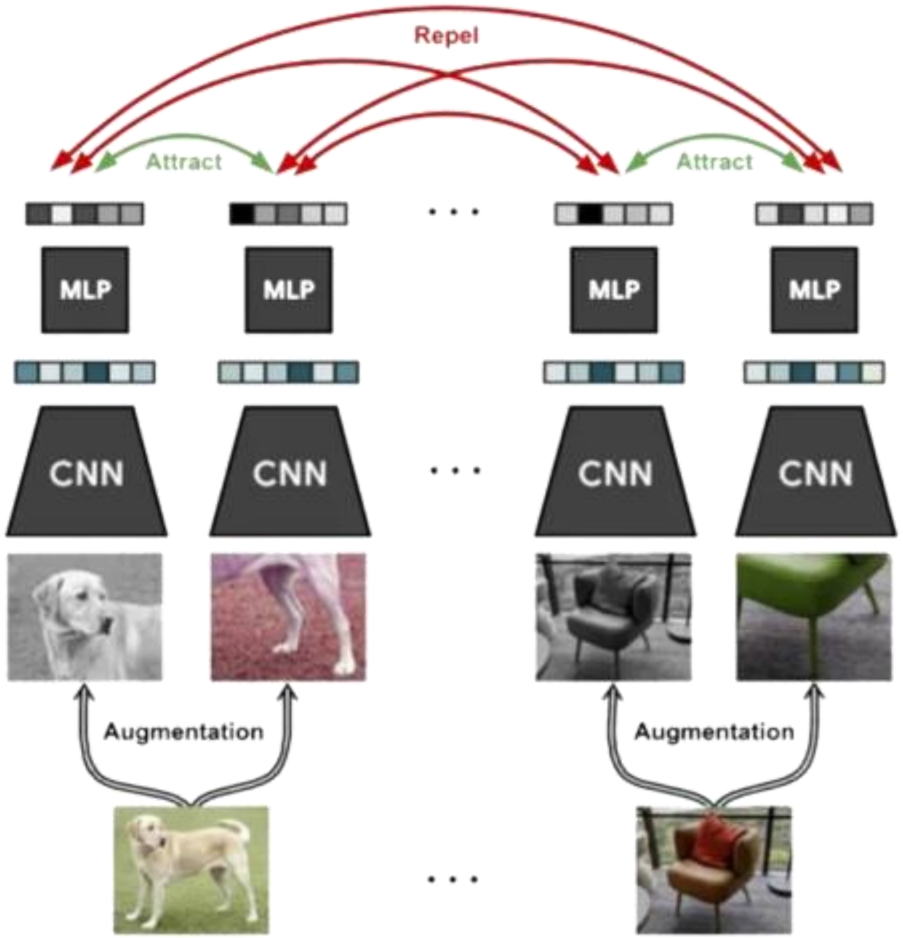



Figure 3: A simple framework for contrastive learning of visual representations. [image idea taken from here]

# Contrastive Learning (cont...)

## Loss Functions

- **InfoNCE Loss (**Information Noise-Contrastive Estimation**):**

  $\longrightarrow$ Maximizes similarity scores between true pairs within a batch

  $\longrightarrow$ The most popular technique used in contrastive learning frameworks

$$L = -\sum_{i=1}^{N} \log \frac{\exp\left(\mathrm{sim}(h_i, h_i^+)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(h_i, h_j^-)/\tau\right)}$$

where $h_i$ and $h_i^+$ are the embeddings of the anchor and positive samples

$h_j$ are the embeddings of negative samples

sim denotes a similarity function, such as cosine similarity

$\tau$ is a temperature parameter that scales the logits

- Other contrastive learning loss functions: Contrastive Loss, Triplet Loss



A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

# Contrastive Learning (cont..)

## Contrastive Learning Models

- **CLIP (Contrastive Language-Image Pretraining)**
  - Learns joint image-text representations with encoders using a contrastive loss
  - Multimodal learning (text and images), zero-shot classification

- CLIP is the **most popular** backbone in MLLMs, enhancing tasks like image-text retrieval, multimodal reasoning, and few-shot learning

- MLLMs like **OpenFlamingo**, **LLaVA** leverage CLIP for improved multimodal performance

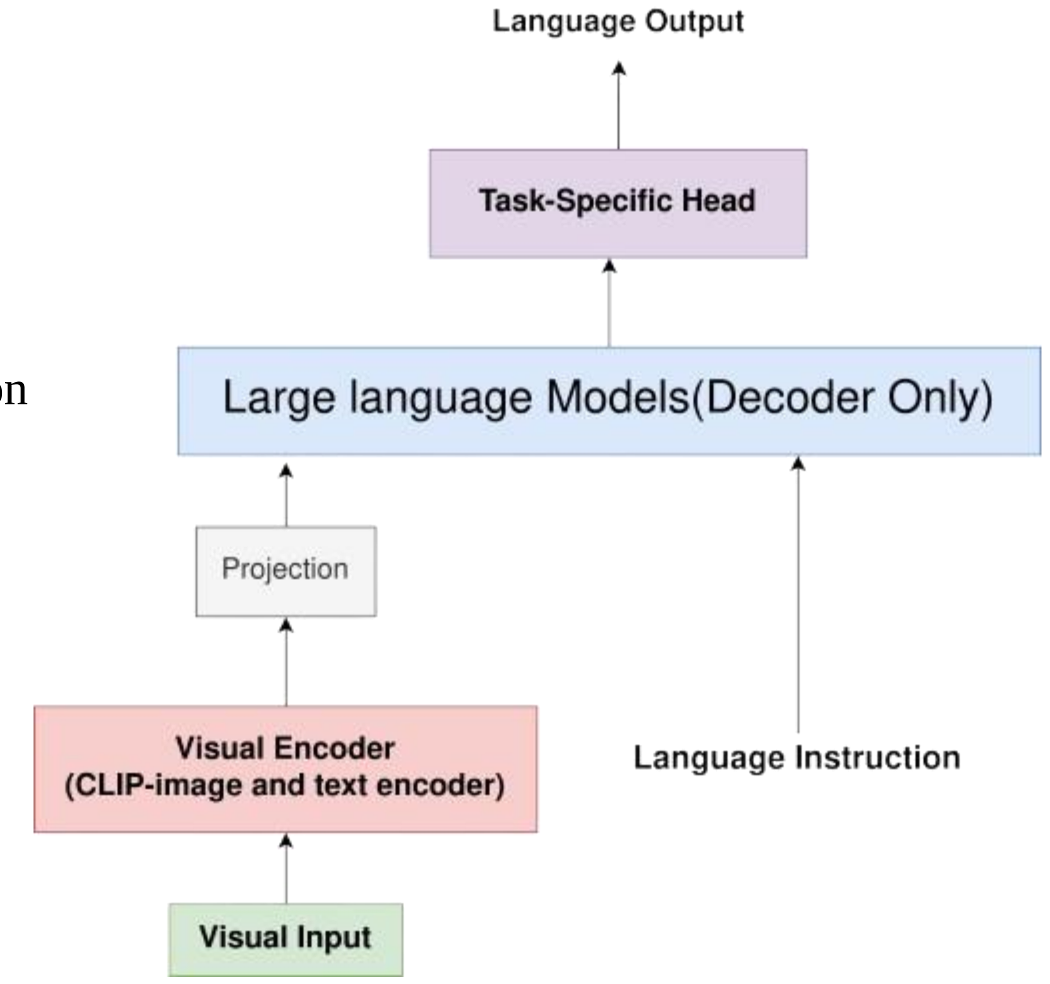



Figure 4: CLIP in MLLMs



A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763

# Literature Review

# Contrastive Learning Based Pretraining

## Example of MLLMs Using Contrastive Learning(CLIP Model)

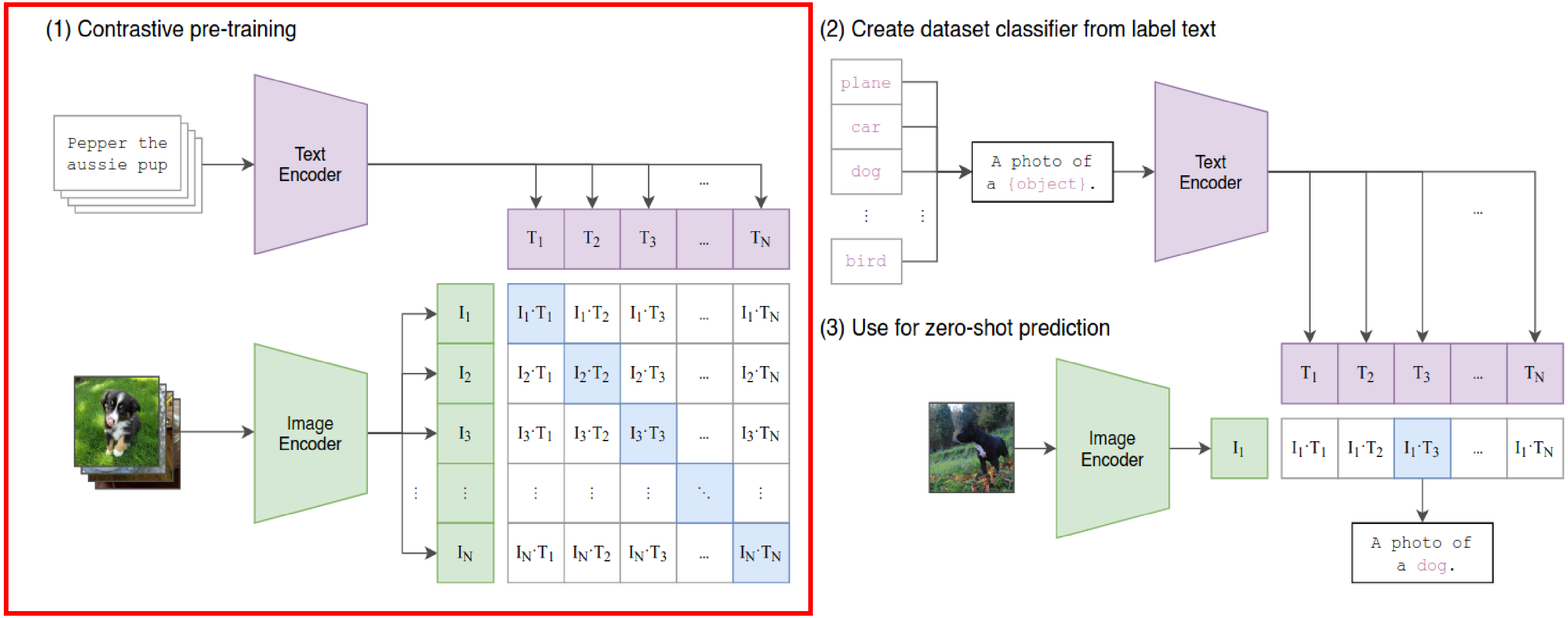
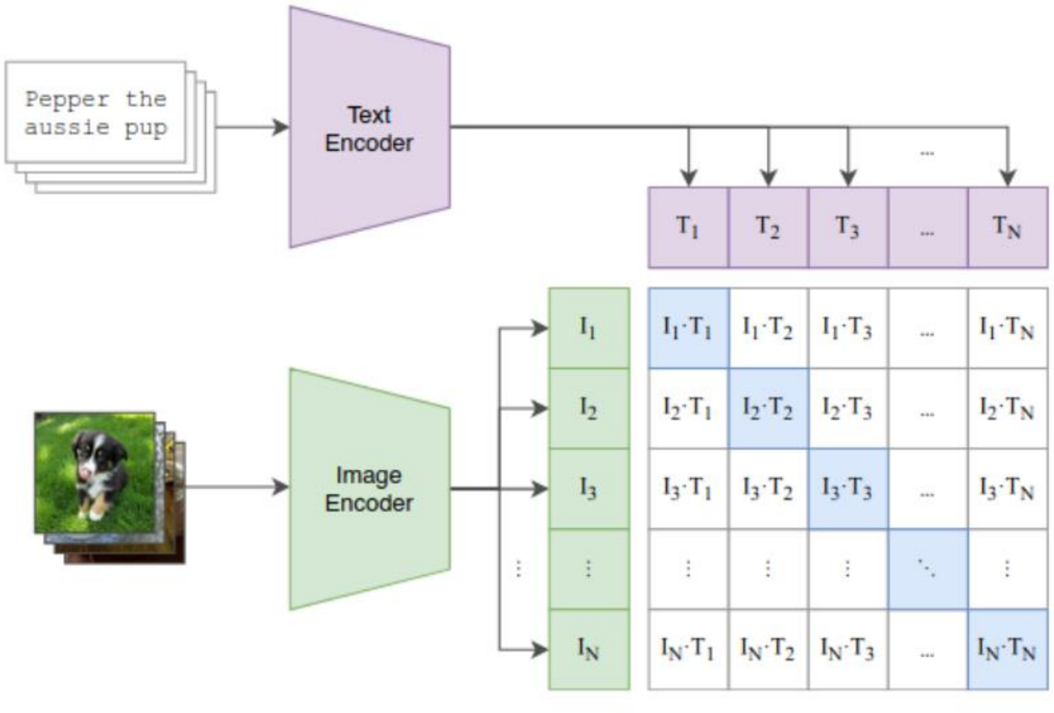



Figure 6: CLIP jointly trains an image encoder and a text encoder [image idea taken from here]



A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763

# Contrastive Learning Based Pretraining




**Cosine similarity loss (from image → text)**

$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle v_i, u_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle v_i, u_k \rangle / \tau)}$$

**Cosine similarity loss (from text → image)**

$$\ell_i^{(u \to v)} = -\log \frac{\exp(\langle u_i, v_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle u_i, v_k \rangle / \tau)}$$

**Final Loss Function**

$$L = \frac{1}{N} \sum_{i=1}^{N} \left( \lambda \ell_i^{(v \to u)} + (1 - \lambda) \ell_i^{(u \to v)} \right)$$

Figure 7: Contrastive Language-Image Pre-training

- Excels in **zero-shot learning**, enabling strong generalization to unseen tasks
- **Additional methods using multimodal contrastive learning**
  - → **ALIGN [1]** enhances CLIP by using large scale noisy web data
  - → **BLIP's [2]** uses encoder-decoder architecture, solve tasks like captioning and text generation

# Literature Review on MLLMs

## MLLMs Using CLIP's image encoder

Liu, Haotian, et al (2023) **[LLaVa-By Meta AI]**

**Task:** Visual question answering, image captioning, multimodal chat

**Method**
- Uses GPT-4 to generate multimodal instruction data
- **Vision Encoder**: CLIP **ViT-L/14**,
- **Language Model**: Vicuna (LLaMA-based)

**Training Steps**:
- Feature Alignment Pretraining (aligns CLIP with LLM)
- Fine-tuning on GPT-4 multimodal instructions.

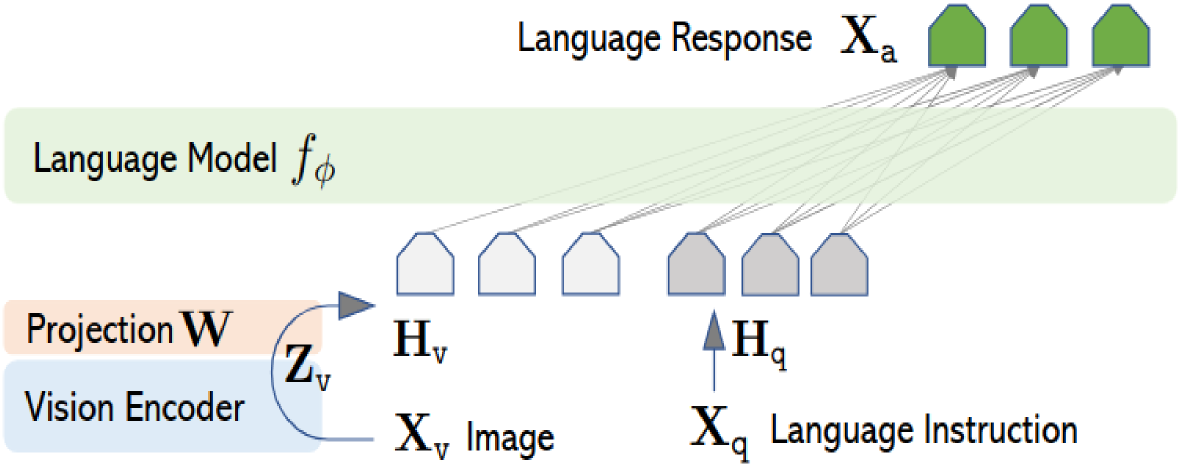



Figure 8: **LLaVA** network architecture [image taken from here ]


H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” Advances in neural information processing systems, vol. 36, 2024.

# Literature Review on MLLMs

## MLLMs Using CLIP's image encoder (LLaVA)

- **Key Advantages**
  - State-of-the-Art (SoTA) Performance: 92.53% on ScienceQA
  - Strong instruction-following (85.1% vs. GPT-4) in multimodal tasks

- **Limitations**
  - **Overfits to instruction-tuned data**, leading to **hallucinations** on unseen inputs

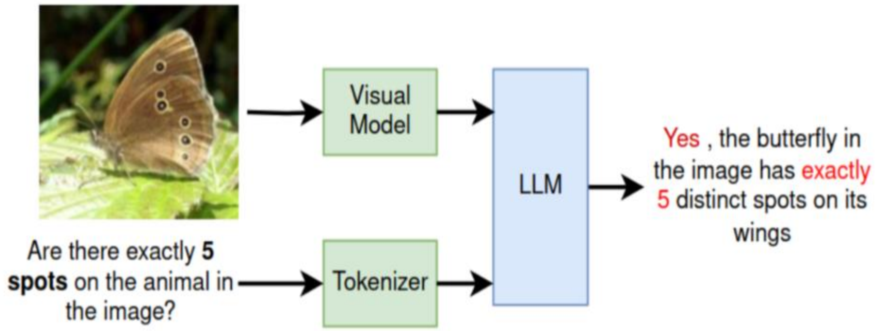



Figure 9: Example of hallucination

**Research Question: How to reduce this hallucinations from MLLMs?**



H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Advances in neural information processing systems, vol. 36, 2024

# Improved MLLMs to solve hallucinations

Jiang, Chaoya, et al.(2024) **[HACL]**

**Task:** Visual Question Answering (VQA), Image Captioning

**Motivation:** Reduce **hallucinations** producing **inaccurate image descriptions** due to **modality gap** and improve **cross-modal alignment** using **contrastive learning**

**Method:** HACL – Hallucination Augmented Cross-modal Contrastive Learning

**Contrastive Learning Framework**:

- Uses **text with hallucination as negative examples**

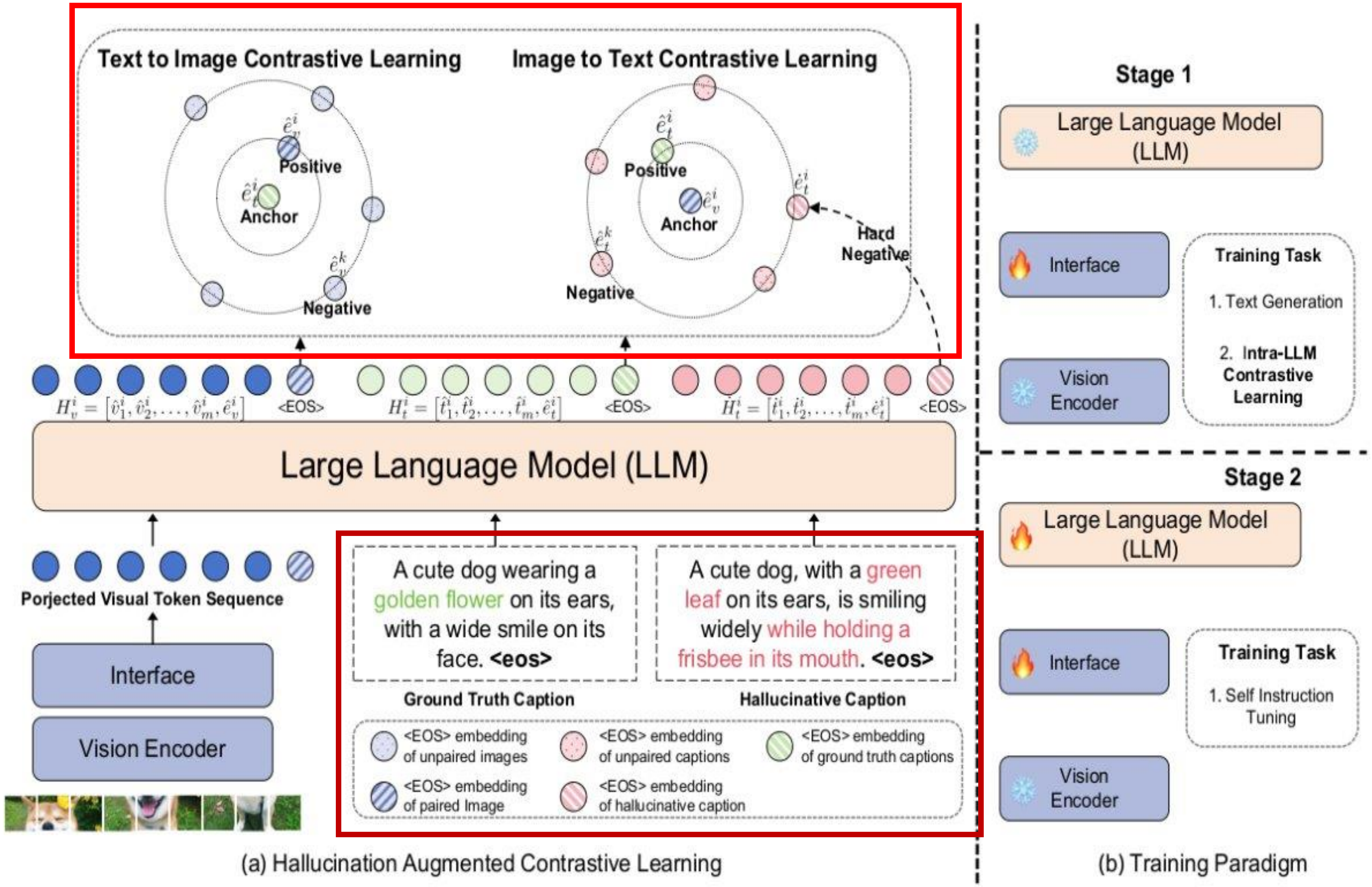



Figure 10: Hallucination Augmented Contrastive Learning[image taken from here]



C. Jiang, H. Xu, M. Dong, J. Chen, W. Ye, M. Yan, Q. Ye, J. Zhang, F. Huang, and S. Zhang, "Hallucination augmented contrastive learning for multimodal large language model," in Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27 036–27 046.

# Improved MLLMs to solve hallucinations

**Training Process**: **Extracts textual and visual** representations using:

- **LLM (Vicuna)**
- **vision encoder (CLIP)**

**Two-stage Training**:

- **Stage 1**: Pre-training with contrastive learning to **refine cross-modal alignment**.
- **Stage 2**: Instruction tuning with **hallucination-aware fine-tuning**

🔹 **Limitation:** Focuses on hallucination reduction but does not enhance complex reasoning

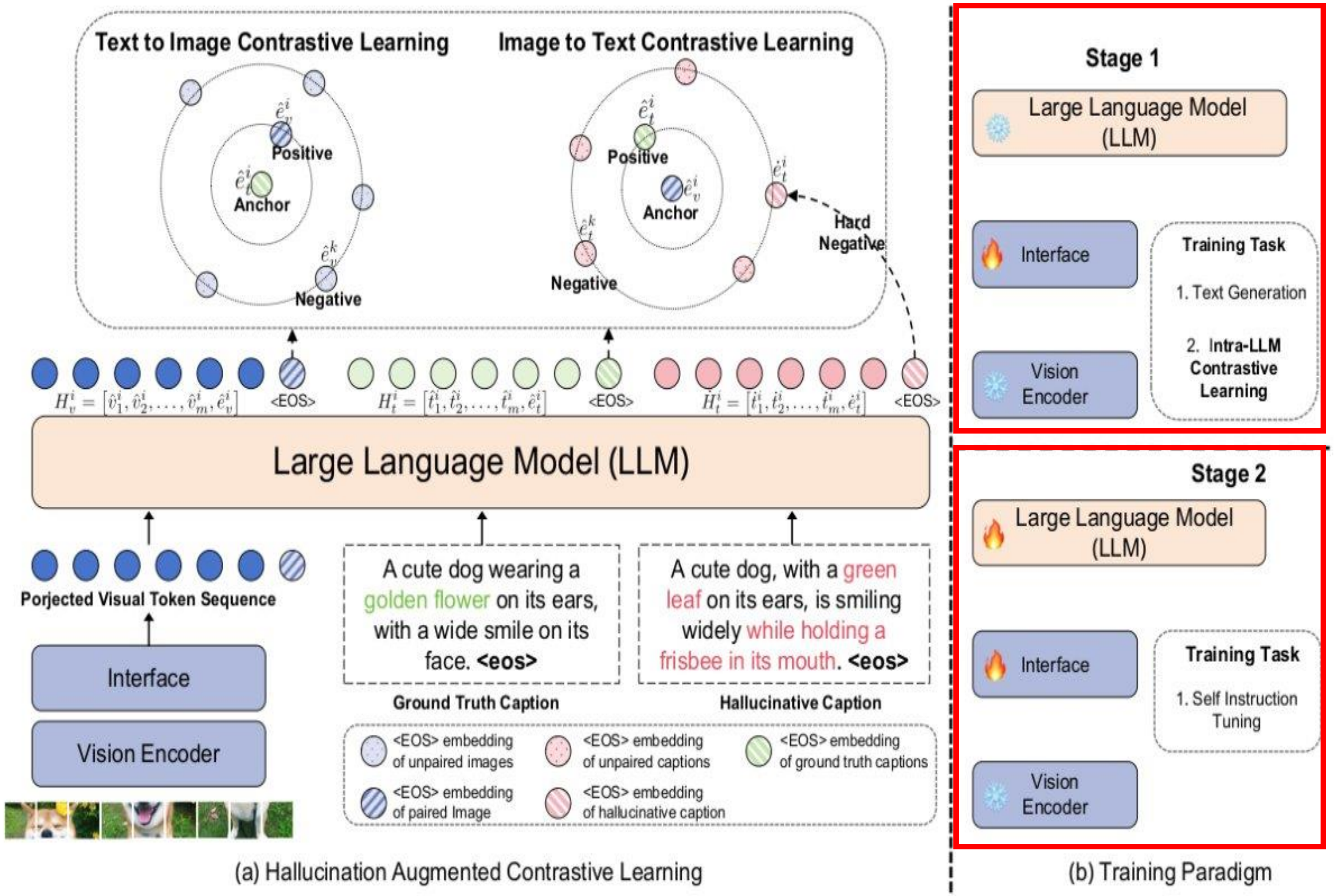



Figure 10: Hallucination Augmented Contrastive Learning[image taken from here]

# Improved MLLMs to solve hallucinations

Sarkar, Pritam, et al.(2024) **[HALVA(LLaVa+DPO)]**

**Task:** Visual Question Answering (VQA), Image Captioning

**Motivation:** Address **object hallucinations** in MLLMs

## Why Does Object Hallucination Occur?

- Spurious Correlations in Training Data
- Positive Instruction Bias

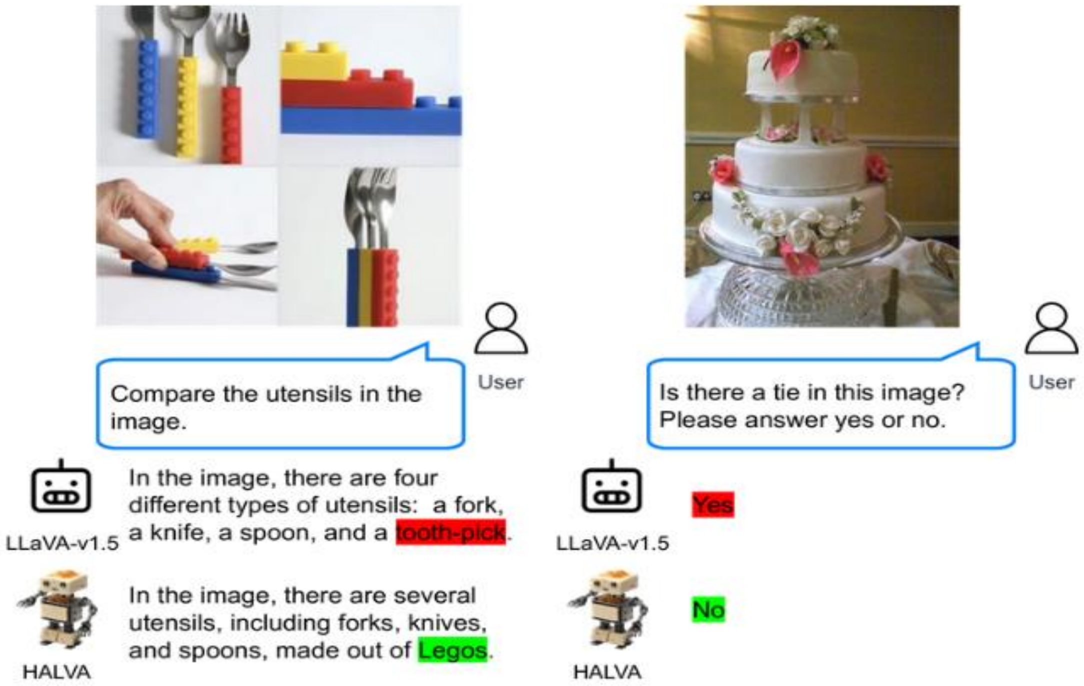



Figure 11: Examples of object hallucinations



Sarkar P, Ebrahimi S, Etemad A, Beirami A, Arık SÖ, Pfister T. Mitigating object hallucination via data augmented contrastive tuning. arXiv e-prints. 2024 May:arXiv-2405.

# Improved MLLMs to solve hallucinations

**Method:** Utilizes **Direct Preference Optimization (DPO),** a version of **Contrastive Learning**,

- **Data Augmentation** – Generates hallucinated responses by modifying **objects, attributes, actions, or locations** in correct responses

- **Phrase-Level Alignment** – Penalizes the model for assigning higher probabilities to hallucinated tokens

- **KL-Divergence Regularization** (keeps model stable)

🔹 **Limitations:** The method mainly targets object hallucinations, while other types of hallucinations remain an open challenge

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N} \ell(x_i, y_i^+, y_i^-; \pi_\theta)$$

$$\mathcal{L}_d = \sum_{t=1}^{T} \pi_{\text{ref}}(y_t^r|x^r) \cdot \left(\log\left(\pi_{\text{ref}}(y_t^r|x^r)\right) - \log\left(\pi_\theta(y_t^r|x^r)\right)\right)$$

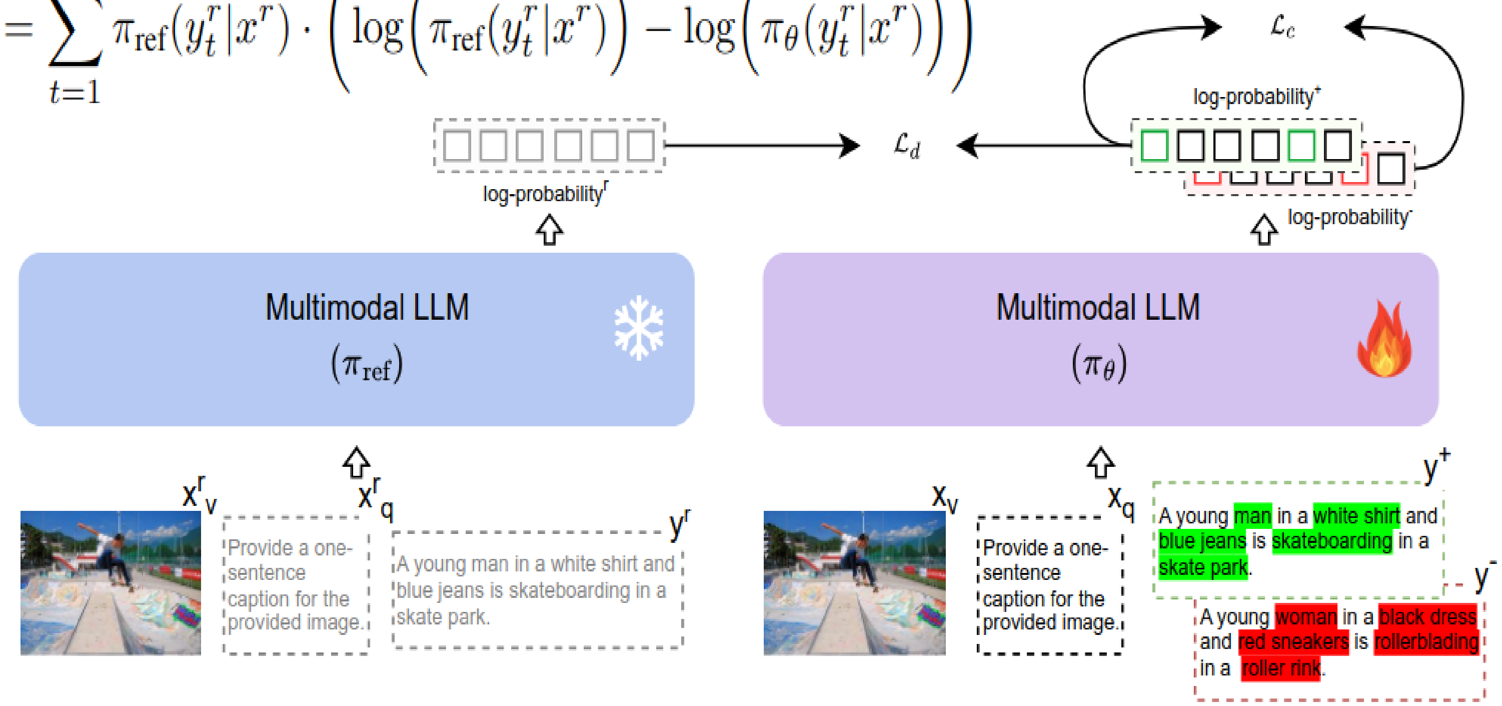



Figure 12: Overview of Hallucination Attenuated Language and Vision Assistant (HALVA) method [image taken from here]

# Improved MLLMs to solve hallucinations

Fu, Jinlan, et al.(2025) **[CHiP]**

**Task:** Visual Question Answering (VQA), Image Captioning

**Motivation:** Improve **hallucination mitigation** by introducing **fine-grained preference optimization** for both **text** and **visual** data

**Method:** CHiP: Cross-Modal Hierarchical Direct Preference Optimization (DPO)

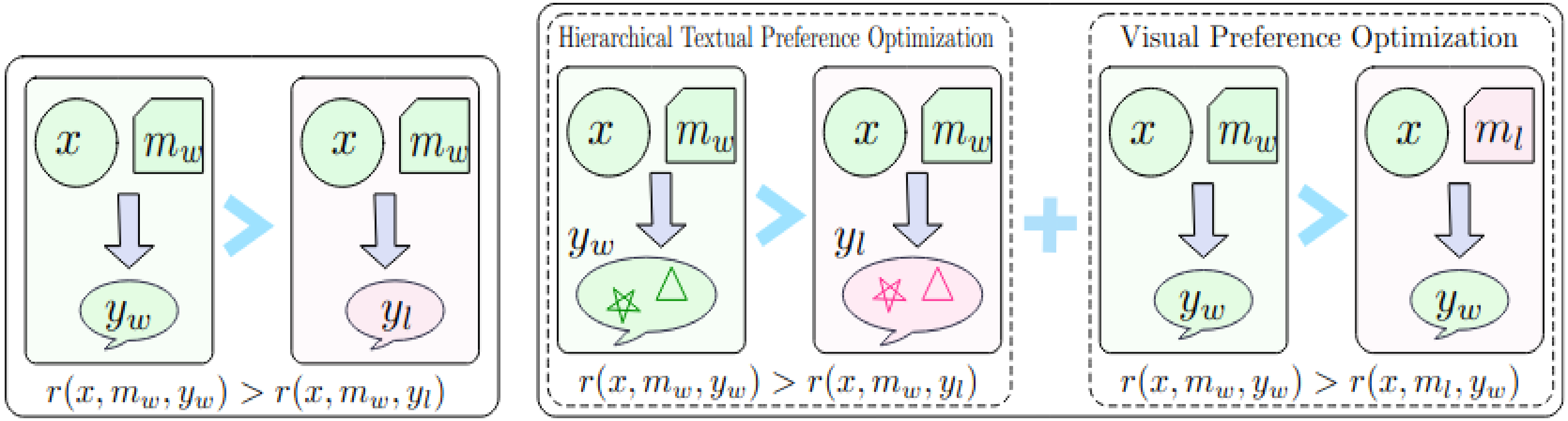
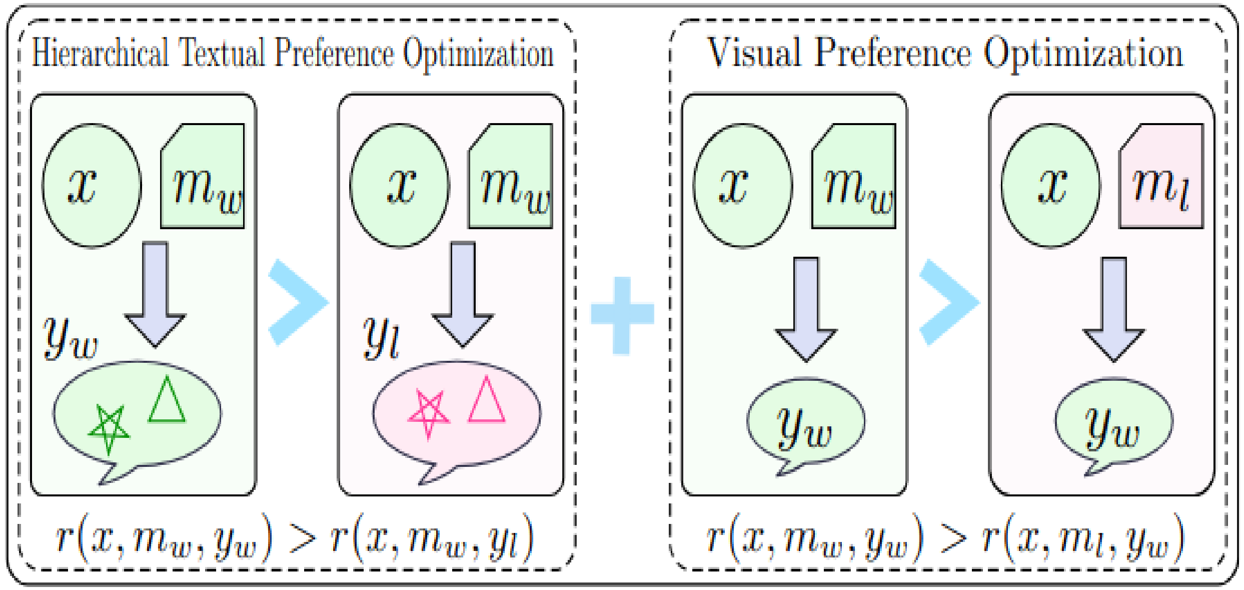



Figure 13: Comparison of preference optimization in different scenarios: Multimodal DPO and CHiP [image taken from here]



Fu J, Huangfu S, Fei H, Shen X, Hooi B, Qiu X, Ng SK. CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs. arXiv preprint arXiv:2501.16629. 2025 Jan 28.

# Improved MLLMs to solve hallucinations




✅ **Hierarchical Textual Preference Optimization**

- **Response-Level** → Optimizes full responses
- **Segment-Level** → Gives **higher rewards** to corrected segments
- **Token-Level** → Uses **KL-divergence loss** to refine hallucination-prone words

$$\mathcal{L}_{HDPO} = \mathcal{L}_{DPO_r} + \lambda \mathcal{L}_{DPO_s} + \gamma \mathcal{L}_{PO_k}$$

✅ **Visual Preference Optimization**

- Compares **preferred vs. modified images** to improve text-image consistency

$$\mathcal{L}_{DPO_v} = -\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|m_w, x)}{\pi_{\text{ref}}(y_w|m_w, x)} - \beta \log \frac{\pi_\theta(y_w|m_l, x)}{\pi_{\text{ref}}(y_w|m_l, x)}\right)$$

Objective of CHiP (Visual+text):

$$\mathcal{L}_{CH\rangle P} = \underbrace{\mathcal{L}_{DPO_v}}_{\text{Visual}} + \underbrace{\mathcal{L}_{DPO_r} + \lambda \mathcal{L}_{DPO_s} + \gamma \mathcal{L}_{PO_k}}_{\text{Text}} .$$

🔹 **Limitations**: **Higher training complexity** due to complex loss function

# Challenges with Current MLLMs

# Challanges in MLLMs

- **Modality Misalignment**
  - Difficulty in mapping representations from different modalities into a unified space
- **Computational Overhead**
  - High training and inference costs due to complex architectures
- **Hallucination in Multimodal Outputs**
  - Generates misleading associations between modalities

**Goal: To design optimization techniques for MLLMs to mitigate hallucination and reduce computational overhead**

# Current and Future Work

# Problem Statement

## Why Hallucinations Happen in MLLMs?

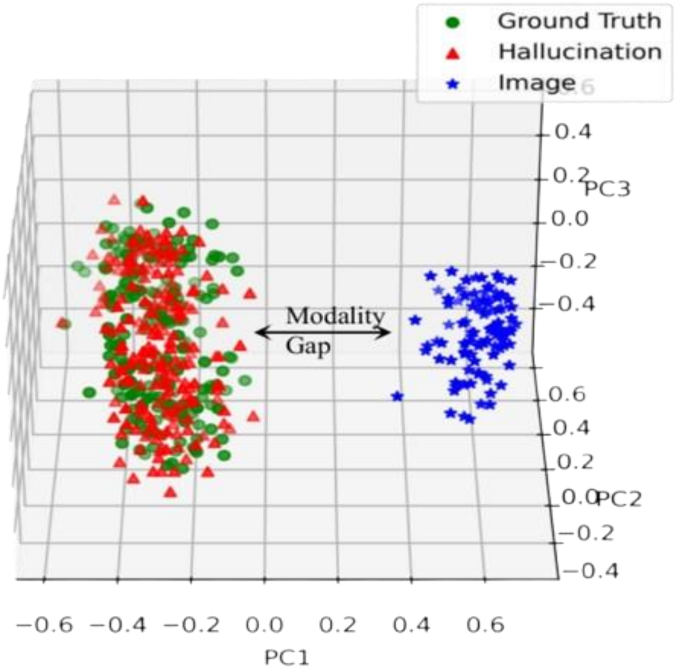



Figure 14: Token Representations w/o Contrastive Learning

**Reasons:**

- Representation Misalignment
- Lack of Semantic Differentiation
- Inefficient Cross-Modal Learning

## Goal

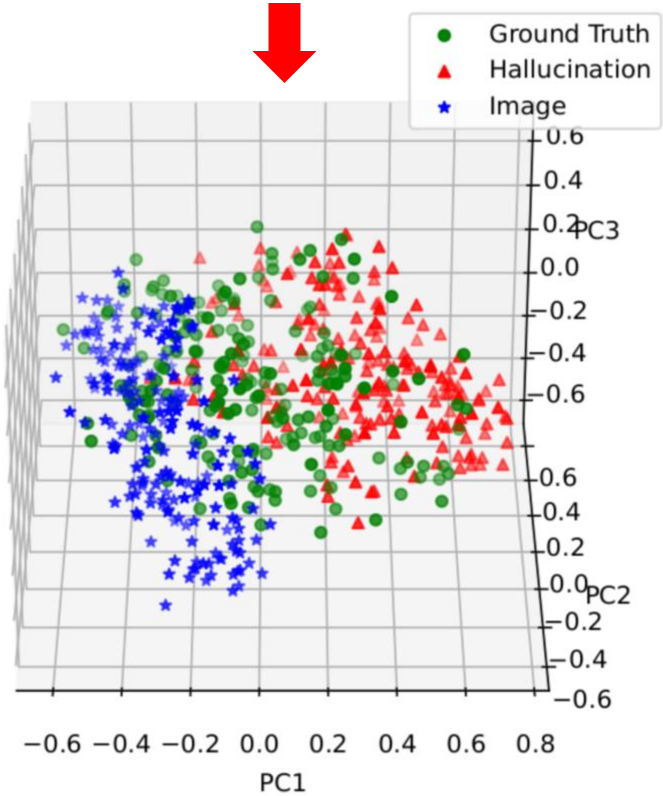


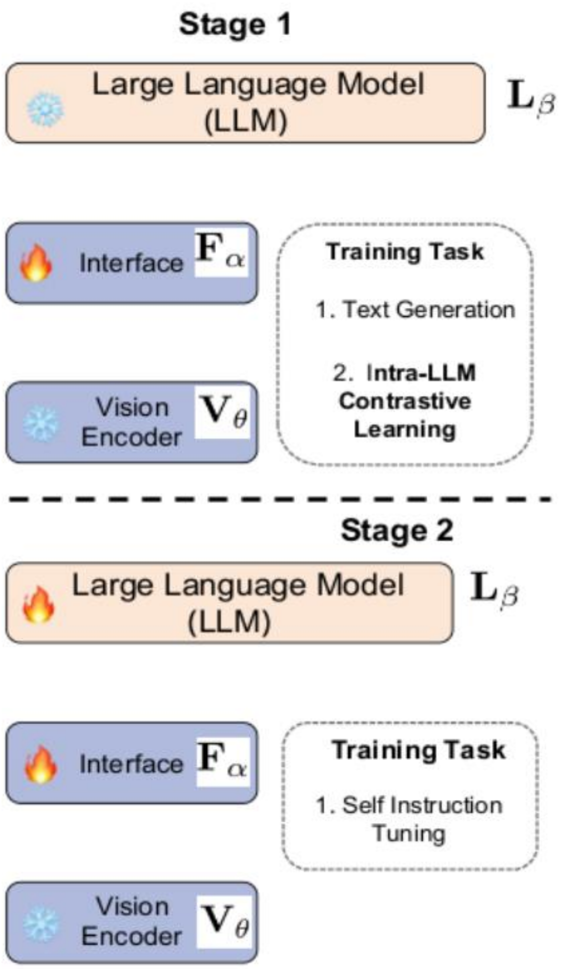

Figure 15: Token Representations w Contrastive Learning

# Issues with earlier approaches

- **Utilize a Loss Function of the form:**

$$L = \min_{\alpha,\beta} \left[ L_T(\beta, \alpha) + \lambda L_{CL}(E) \right]$$

where $L_{CL}$ ensures cross-modal alignment using contrastive learning
$L_T$ fine-tunes the model for task specific purposes




This previous approaches (like the above) introduce a fundamental **trade-off**:

- Attempt to optimize both learned embeddings and task-specific adaptation using a **single objective**
- Leads to compromising **either embedding quality or task performance**

**Research Question: How to address this problem ?**



C. Jiang, H. Xu, M. Dong, J. Chen, W. Ye, M. Yan, Q. Ye, J. Zhang, F. Huang, and S. Zhang, "Hallucination augmented contrastive learning for multimodal large language model," in Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27 036–27 046.

# Our Approach (Current work)

## Problem Formulation

The objective should take a **hierarchical optimization formulation:**

- **Level 1:** Learn the optimal embeddings
- **Level 2:** From the set of optimal embeddings chose the best for the specific task

$$\min_{E} L_T(E) \quad \text{subject to} \quad E \in \arg\min_{\alpha,\beta} L_{CL}(E(\alpha,\beta))$$

where E is the learned multimodal embedding space and α, β are model parameters shown in the Figure

**Advantage:**

- No compromise over quality of learned embeddings and the task performance

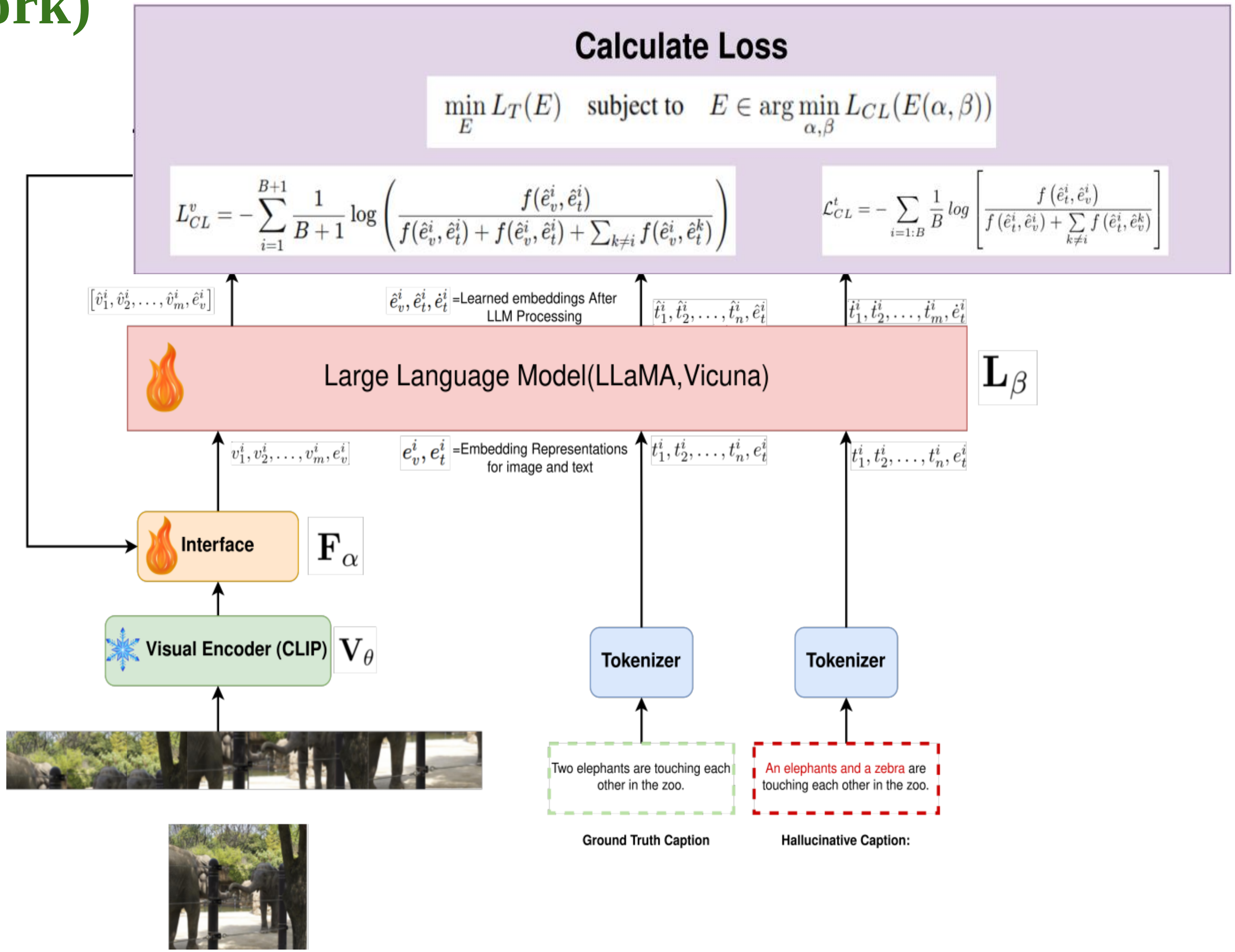



Figure 16: Overview of Proposed Hierarchical Fine-Tuning Approach

# Future Work

- Design a model capable of processing **more modalities** for enhanced multimodal understanding
- Enhance our proposed architecture by incorporating **improved optimization**
- Instead of processing visual and text embeddings Separately
  - Integrate both embeddings directly into the LLM
  - **Reduce complexity and improve multimodal** representation learning

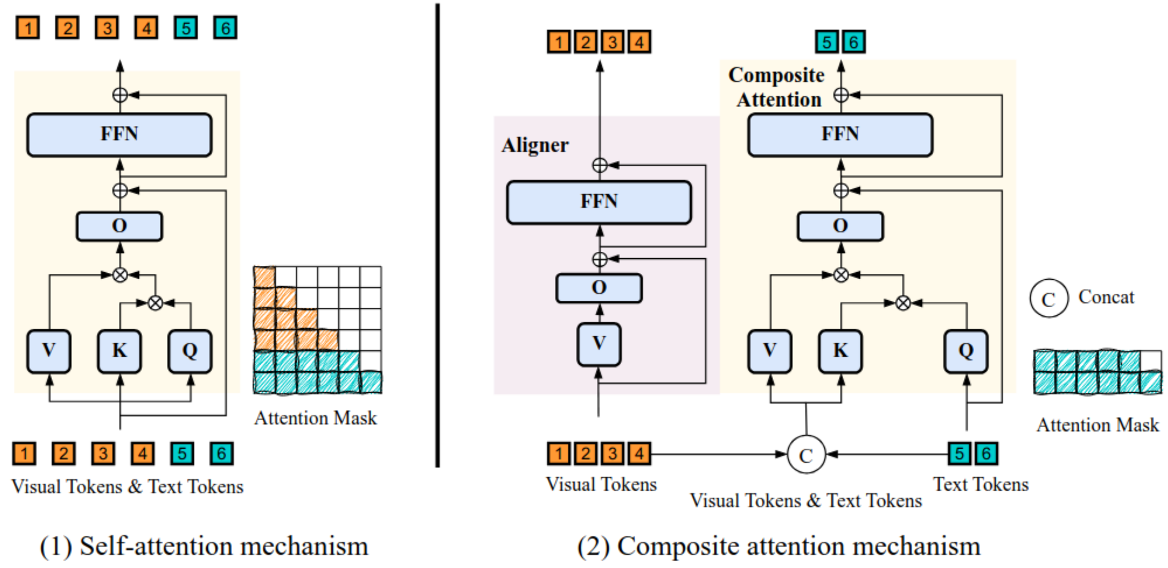



Figure 17: Composite attention mechanism [image taken from here here]

F. Ma, Y. Zhou, H. Li, Z. He, S. Wu, F. Rao, Y. Zhang, and X. Sun, "Ee-mllm: A data-efficient and compute-efficient multimodal large language model," arXiv preprint arXiv:2408.11795, 2024

# Conclusion

- **MLLMs have revolutionized AI** by integrating multiple modalities like text, images, and audio
- **Contrastive learning** enhances **alignment, zero-shot learning, and retrieval** in models like **CLIP, BLIP-2, and OpenFlamingo**
- **CLIP** is utilized as a **backbone** in many modern MLLMs
- **Modality imbalance, high computational costs, and hallucination**
- **Efficient contrastive learning can help** to reduce hallucination
- Overcoming these challenges will make MLLMs **more adaptive, interpretable, and human-aligned**

# Works and Plan

## Timeline

| Task | Year | 2025 | | | | | | | | | | 2026 | | | | | | | | | | | | After |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Months | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | |
| Implement Joint Optimization of α and β | | | | | | | | | | | | | | | | | | | | | | | | |
| Research and Implement Integration of DPO | | | | | | | | | | | | | | | | | | | | | | | | |
| Research on simplifying multimodal representation learning. | | | | | | | | | | | | | | | | | | | | | | | | |
| Attack on Multimodality | | | | | | | | | | | | | | | | | | | | | | | | |

# References


[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., “Learning transferable visual models from natural language supervision,” in International conference on machine learning. PMLR, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in International conference on machine learning. PMLR, 2021, pp. 4904–4916.

[3] Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. InInternational conference on machine learning 2022 Jun 28 (pp. 12888-12900). PMLR.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning. PMLR, 2020, pp.1597–1607.

[5] C. Jiang, H. Xu, M. Dong, J. Chen, W. Ye, M. Yan, Q. Ye, J. Zhang, F. Huang, and S. Zhang, “Hallucination augmented contrastive learning for multimodal large language model,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp.27 036–27 046.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,”arXiv preprint arXiv:1807.03748, 2018.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” Advances in neural information processing systems, vol. 36, 2024.

[8] Sarkar P, Ebrahimi S, Etemad A, Beirami A, Arık SÖ, Pfister T. Mitigating object hallucination via data augmented contrastive tuning. arXiv e-prints. 2024 May:arXiv-2405.

# References

[9] Fu J, Huangfu S, Fei H, Shen X, Hooi B, Qiu X, Ng SK. CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs. arXiv preprint arXiv:2501.16629. 2025 Jan 28.

[10] F. Ma, Y. Zhou, H. Li, Z. He, S. Wu, F. Rao, Y. Zhang, and X. Sun, "Ee-mllm: A data-efficient and compute-efficient multimodal large language model," arXiv preprint arXiv:2408.11795, 2024.

[11] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. Alwala, A. Joulin, and I. Misra, "Imagebind: one embedding space to bind them all. arxiv," Preprint posted online on May, vol. 9, 2023.

[12] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu et al., "Palm-e: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.

[13] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: a visual language model for few-shot learning," Advances in neural information processing systems, vol. 35, pp. 23 716–23 736, 2022.

[14] C. X. Liang, P. Tian, C. H. Yin, Y. Yua, W. An-Hou, L. Ming, T. Wang, Z. Bi, and M. Liu, "A comprehensive survey and guide to multimodal large language models in vision-language tasks," arXiv preprint arXiv:2411.06284, 2024.

[15] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," arXiv preprint arXiv:2401.13601, 2024.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in International conference on machine learning. PMLR, 2023, pp. 19 730–19 742.

# Acknowledgment


- Thanking my advisor Dr. **Prashant Khanduri** for giving me precious advices and directions on the literature review and my research.

- Thanking Dr. **Zichun Zhong**, Dr. **Dongxiao Zhu** and Dr. **Lihui Liu** for being the committee members and helping me in my qualifying exam.

- I also want to thank my lab mate Nasim in Optimization for Large Scale Machine Learning Lab (OptLML) and in **Trustworthy AI**, especially Rafi ,Hui, Xiangyu , Amin, Saleh who help and guide me everyday in various tasks.

# THANK YOU ☺

# Q/A