Hallucination-Free Multimodal LLMs: A Unified Training Approach

WAYNE STATE
College of Engineering
Computer Science

Aditi Sarker, Nasim Tohidi, Prashant Khanduri

Optimization for Large Scale Machine Learning Lab (OptLML)
Department of Computer Science

Motivation

- **Multimodal Large Language Model (MLLM)**: LLMs that processes and understands **multiple modalities of data, such as text, images, video, etc.** instead of just text like traditional LLMs.
- MLLMs often generate **hallucinations**—fabricated or incorrect information that does not match visual input.
- There is a **modality gap** between textual and visual representations, leading to misalignment.
- Hallucinative and non-hallucinative texts are **entangled**, making it difficult to differentiate them.
- **Existing methods fail** to effectively bridge the vision-language gap, increasing the occurrence of hallucinations.

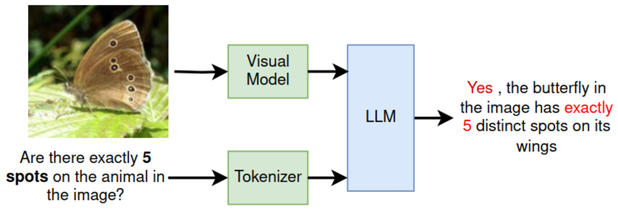
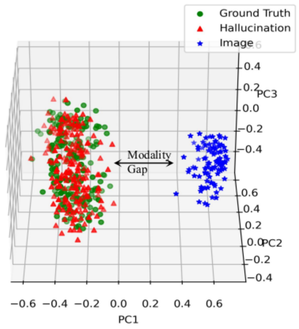


Figure 1. Example of Hallucination

Why Hallucinations Happen in MLLMs?

- ➜ **Representation Misalignment:** Visual features are not well mapped to the textual space.
- ➜ **Lack of Semantic Differentiation:** Models struggle to separate correct vs. incorrect textual representations.
- ➜ **Training Data Issues:** Models are trained on noisy, biased, or incomplete multimodal datasets.
- ➜ **Over-Reliance on Language Priors:** The model prioritizes textual context over actual visual content, leading to hallucinated descriptions.
- ➜ **Inefficient Cross-Modal Learning:** Current projection-based methods fail to fully integrate visual cues into the language model.

Problem Statement



- In Figure 2, the **modality gap** between image representations (blue stars) and text representations (green circles for ground truth, red triangles for hallucinations) shows **poor alignment** in MLLMs.
- This gap makes it hard for the model to correctly map images with correct text, leading to **hallucinations**.
- Additionally, ground truth and hallucinated text **overlap**, making it difficult to differentiate correct descriptions from incorrect ones.

Figure 2. Token Representations.

How to address this problem ?

- ✓ **Contrastive Learning (CL)[1]:** improves alignment by treating hallucinated text as hard negatives and correct text as positives, pulling accurate pairs closer while pushing incorrect ones away.
- ✓ **Direct Preference Optimization (DPO)[2]:** trains models to prefer correct responses over hallucinated ones using a pairwise learning approach. It reinforces accurate descriptions but struggles with aligning image-text representations, requiring refinements.
- ✓ **Cross-Modal Hierarchical DPO (CHiP)[3]**: extends DPO by incorporating visual and textual preferences at multiple levels. It optimizes at the response, segment, and token levels, ensuring better alignment and significantly reducing hallucinations in MLLMs.

Contributions

Challenges

- ✗ **Joint** embedding and parameter learning.
- ✗ Pre-training (L_{CL}) aligns vision-text representations but lacks task supervision.
- ✗ Learns suboptimal embeddings and requires a large amount of (pre)training data.
- ✗ Fine-tuning (L_T) applies task learning but does not update embeddings, preventing adaptation.

The **overall loss** is defined as:

$$L = \min_{\alpha,\beta} [L_T(\beta, \alpha) + \lambda L_{CL}(E)] \tag{1}$$

Our approach (Hierarchical Fine-Tuning)

- ✓ **Decoupled** embedding and parameter learning.
- ✓ Embeddings are refined within fine-tuning, rather than being learned separately.
- ✓ Pre-trained embeddings are not frozen, they continuously adapt during fine-tuning.
- ✓ Contrastive learning is integrated, ensuring task-aware representation learning.
- ✓ Prevents hallucinations by aligning contrastive learning with task objectives.
- ✓ Optimizes embeddings dynamically, improving generalization across multimodal tasks.

Problem Formulation

The **overall loss** in the proposed **hierarchical optimization** formulation:

$$\min_{E} L_T(E) \quad \text{subject to} \quad E \in \arg\min_{\alpha,\beta} L_{CL}(E(\alpha, \beta)) \tag{2}$$

- **Inner Optimization (L_{CL}):** Dynamically refines embeddings during fine-tuning through contrastive learning, ensuring better vision-text alignment and reducing modality misalignment.
- **Outer Optimization (L_T):** Optimizes task-specific loss using the continuously updated embeddings, allowing for better adaptation to downstream tasks and improved generalization.

System Architecture

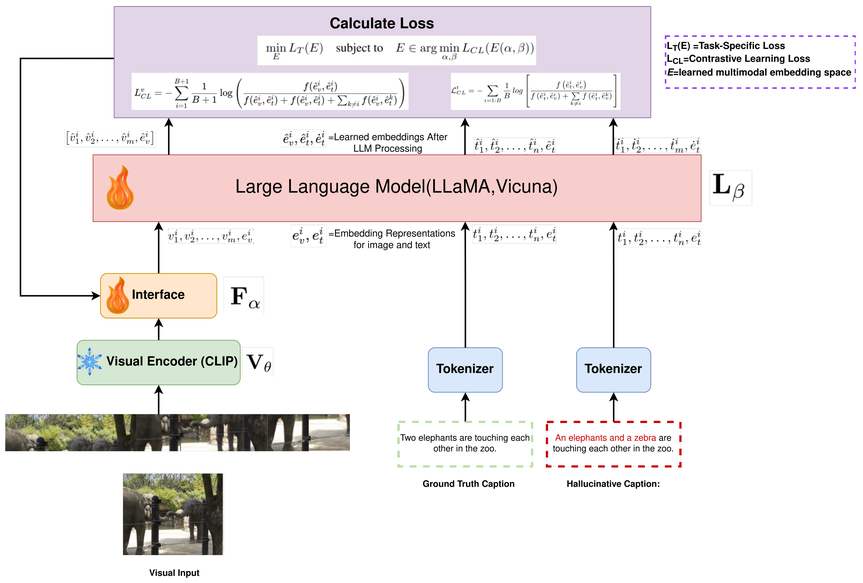


Figure 3. Overview of Proposed Hierarchical Fine-Tuning Approach

Algorithm Description

Algorithm 1 Hierarchical Fine-Tuning Algorithm (HFA)

Initialization:
  Input: V_θ, F_α, L_β, and dataset $D = \{(I_i, T_i)\}$, where $i \in \{1, 2, ..., N\}$.
  Output: Optimized parameters α, β and multimodal embeddings E.
for each image I_i and text T_i in D **do**
  For image: $S_{vi} \leftarrow V_\theta(I_i)$ and extract $\hat{e}_i^v \leftarrow F_\alpha(I_i)$
  For text: Extract $\hat{e}_i^t$ and e_i^t from actual and hallucinated text
end for
for each batch of B samples **do**
  Compute contrastive loss:

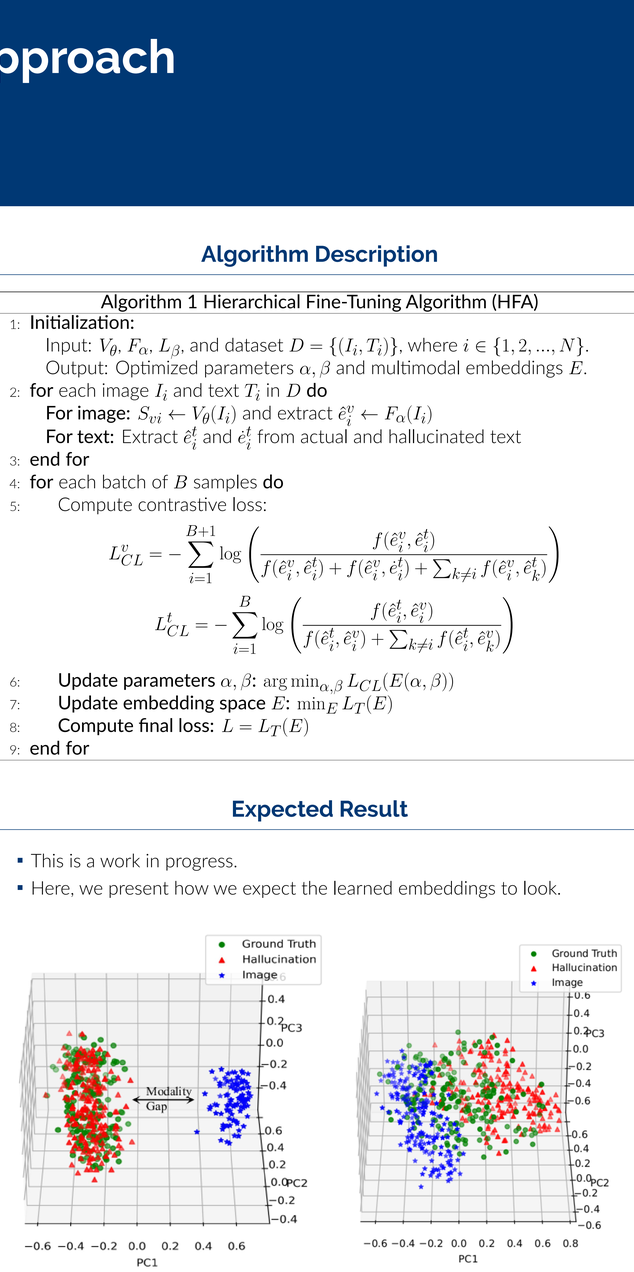
$$L^v_{CL} = -\sum_{i=1}^{B+1} \log\left(\frac{f(\hat{e}_i^v, \hat{e}_i^t)}{f(\hat{e}_i^v, \hat{e}_i^t) + f(\hat{e}_i^v, \hat{e}_i^t) + \sum_{k\neq i} f(\hat{e}_i^v, \hat{e}_k^t)}\right)$$

$$L^t_{CL} = -\sum_{i=1}^{B} \log\left(\frac{f(\hat{e}_i^t, \hat{e}_i^v)}{f(\hat{e}_i^t, \hat{e}_i^v) + \sum_{k\neq i} f(\hat{e}_i^t, \hat{e}_k^v)}\right)$$

  Update parameters α, β: $\arg\min_{\alpha,\beta} L_{CL}(E(\alpha, \beta))$
  Update embedding space E: $\min_E L_T(E)$
  Compute final loss: $L = L_T(E)$
end for

Expected Result

- This is a work in progress.
- Here, we present how we expect the learned embeddings to look.



(a) Token Representations w/o Contrastive Learning (b) Token Representations w Contrastive Learning

Figure 4. Comparison of Expected Outputs.

Conclusion

✓ The contrastive loss moves the **image embeddings closer to the ground truth** while **separates the ground truth from the hallucinated embeddings**.

References

[1] 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*, pp. 27036–27046, 2024.

[2] P. Sarkar, S. Ebrahimi, A. Etemad, A. Beirami, S. Ö. Arık, and T. Pfister, "Mitigating object hallucination via data augmented contrastive tuning," *arXiv preprint arXiv:2405.18654*, 2024.

[3] J. Fu, S. Huangfu, H. Fei, X. Shen, B. Hooi, X. Qiu, and S.-K. Ng, "Chip: Cross-modal hierarchical direct preference optimization for multimodal llms," *arXiv preprint arXiv:2501.16629*, 2025.

hq1351@wayne.edu