Novelty Assessment Report

Paper: RL makes MLLMs see better than SFT
PDF URL: https://openreview.net/pdf?id=3gM6HwHvnc
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

A dominant assumption in Multimodal Language Model (MLLM) research is that its performance is largely inherited from the LLM backbone, given its immense parameter scale and remarkable capabilities. This has created a void in the understanding of the vision encoder, which determines 'how MLLMs perceive images'. The recent shift in MLLM training paradigms, from Supervised Finetuning (SFT) to Reinforcement Learning (RL), magnifies this oversight—namely, the significant lack of analysis on how such training reshapes the vision encoder as well as the MLLM. To address this, we first investigate the impact of training strategies on MLLMs, where RL shows a clear advantage in strongly vision-related VQA benchmarks than SFT. Motivated by this, we conduct a critical yet under-explored analysis of the vision encoder of MLLMs through diverse and in-depth experiments, ranging from ImageNet classification and segmentation to gradient visualization. Our results demonstrate that MLLM's post-training strategy 'i.e, SFT or RL' not only leads to disctinct outcomes on MLLM downstream tasks, but also fundamentally reshapes MLLM's underlying visual representations. Specifically, our main finding is that RL produces stronger and more localized visual representations compared to SFT, boosting the ability of the vision encoder for MLLM. We then reframe our findings into a simple recipe for building strong vision encoders for MLLMs, Preference-Instructed Vision OpTimization (PIVOT). When integrated into MLLMs, a PIVOT-trained vision encoder outperforms even larger and more heavily-trained counterparts, despite requiring less than 1\% of the computational cost of standard vision pretraining. This result opens an effective and efficient path for advancing the vision backbones of MLLMs.

Core Task Landscape

This paper addresses: **Vision Encoder Training in Multimodal Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Vision Encoder Architecture and Design**
- **Training Paradigms and Optimization**
- **Cross-Modal Alignment and Integration**
- **Multimodal Model Applications and Capabilities**
- **Evaluation, Benchmarking, and Analysis**
- **Efficient and Lightweight MLLM Design**

Complete Taxonomy Tree

- Vision Encoder Training in Multimodal Language Models Survey Taxonomy
- Vision Encoder Architecture and Design
  - Single Vision Encoder Architectures
  - CLIP-Based and Transformer Vision Encoders (3 papers)
    - [1] From clip to dino: Visual encoders shout in multi-modal large language models (Jiang Dongsheng, 2023) View paper
    - [29] Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning (Li, 2025) View paper
    - [38] Multimodal LLMs (Uday Kamath, 2024) View paper
  - Hierarchical and Convolutional Encoders (1 papers)
    - [16] Convllava: Hierarchical backbones as visual encoder for large multimodal models (Ge, 2024) View paper
  - Dynamic Resolution and High-Resolution Processing (3 papers)
    - [12] How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites (Zhe Chen, 2024) View paper
    - [28] Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution (Wang Peng, 2024) View paper
    - [47] InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD (Dong, 2024) View paper
  - Multi-Encoder Fusion Strategies
  - Mixture of Vision Experts (4 papers)
    - [2] Vcoder: Versatile vision encoders for multimodal large language models (Jitesh Jain, 2024) View paper
    - [17] Mome: Mixture of multimodal experts for generalist multimodal large language models (Chen Gongwei, 2024) View paper
    - [19] LEO: Boosting Mixture of Vision Encoders for Multimodal Large Language Models (Riddell, 2025) View paper
    - [20] Eagle: Exploring the design space for multimodal llms with mixture of encoders (Shi Min, 2024) View paper
  - Token Compression and Efficiency (3 papers)
    - [3] Tokenpacker: Efficient visual projector for multimodal llm (Wentong Li, 2025) View paper
    - [30] LaCo: Efficient Layer-wise Compression of Visual Tokens for Multimodal Large Language Models (Liu Jun-tao, 2025) View paper
    - [37] METEOR: Multi-Encoder Collaborative Token Pruning for Efficient Vision Language Models (Liu, 2025) View paper

- Training Paradigms and Optimization
  - Reinforcement Learning for Vision Encoder Training ★ (2 papers)
  - [0] RL makes MLLMs see better than SFT (Anon et al., 2026) View paper
  - [44] Vlm-r1: A stable and generalizable r1-style large vision-language model (Shen, 2025) View paper
  - Supervised and Self-Supervised Pre-Training (3 papers)
  - [5] Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs (Tong, 2024) View paper
  - [9] Mm1: methods, analysis and insights from multimodal llm pre-training (Brandon McKinzie, 2024) View paper
  - [27] MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training (McKinzie, 2024) View paper
  - Visual Instruction Tuning (3 papers)
  - [18] Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models (Chen Chi, 2023) View paper
  - [23] Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models (Maaz, 2023) View paper
  - [34] Visual Instruction Tuning (Liu Haotian, 2023) View paper
- Cross-Modal Alignment and Integration
  - Visual Projector and Connector Design (2 papers)
  - [8] An introduction to vision-language modeling (Bordes, 2024) View paper
  - [13] Multimodal few-shot learning with frozen language models (Maria Tsimpoukelli, 2021) View paper
  - Embedding Alignment and Distillation (3 papers)
  - [15] Visual Representation Alignment for Multimodal Large Language Models (Jung Jaewoo, 2025) View paper
  - [25] Ovis: Structural Embedding Alignment for Multimodal Large Language Model (Lu, 2024) View paper
  - [35] Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation (Jain, 2025) View paper
  - Visual Grounding and Spatial Reasoning (1 papers)
  - [6] Bubogpt: Enabling visual grounding in multi-modal llms (Zhao Yang, 2023) View paper
- Multimodal Model Applications and Capabilities
  - Perception-Focused Applications (4 papers)
  - [7] Visual cognition in multimodal large language models (Luca M. Schulze Buschoff, 2025) View paper
  - [11] Contextual Object Detection with Multimodal Large Language Models (Yuhang Zang, 2023) View paper
  - [24] Object detection with multimodal large vision-language models: An in-depth review (Ranjan Sapkota, 2025) View paper
  - [49] Incorporating visual experts to resolve the information loss in multimodal large language models (He, 2024) View paper
  - Embodied AI and Multi-Modal Extensions (4 papers)
  - [4] Palm-e: An embodied multimodal language model (Driess, 2023) View paper
  - [22] Vita: Towards open-source interactive omni multimodal llm (Fu, 2024) View paper
  - [36] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action (Jiasen LÃ¼, 2023) View paper
  - [42] OpenVLA: An Open-Source Vision-Language-Action Model (Kim, 2024) View paper
  - Text-Rich Visual Understanding (2 papers)
  - [43] Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond (Bai, 2023) View paper
  - [45] Bliva: A simple multimodal llm for better handling of text-rich visual questions (Hu Wenbo, 2024) View paper
  - Domain-Specific and Specialized Applications (3 papers)
  - [31] MolPROP: Molecular Property prediction with multimodal language and graph fusion (Zachary A. Rollins, 2024) View paper
  - [40] Automated context-aware navigation support for individuals with visual impairment using multimodal language models in urban environments (A Chao, 2025) View paper
  - [50] Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis (Ling Yan, 2022) View paper
- Evaluation, Benchmarking, and Analysis
  - Comprehensive Surveys and Reviews (4 papers)
  - [10] The revolution of multimodal large language models: a survey (Davide Caffagni, 2024) View paper
  - [21] A survey of multimodel large language models (Zijing Liang, 2024) View paper
  - [32] Multimodal large language models: A survey (Jiayang Wu, 2023) View paper
  - [41] From large language models to large multimodal models: A literature review (Dawei Huang, 2024) View paper
  - Benchmark Studies and Capability Evaluation (1 papers)
  - [14] A survey on benchmarks of multimodal large language models (Li Jian, 2024) View paper
  - Vision-Centric Analysis and Representation Studies (3 papers)
  - [33] Words over pixels? rethinking vision in multimodal large language models (Anubhooti Jain, 2025) View paper
  - [39] Imagined visual representations as multimodal embeddings (Guillem Collell, 2017) View paper
  - [48] Question Aware Vision Transformer for Multimodal Reasoning (Roy Ganz, 2024) View paper
- Efficient and Lightweight MLLM Design (2 papers)
  - [26] MobileVLM V2: Faster and Stronger Baseline for Vision Language Model (Chu, 2024) View paper
  - [46] Bridging Compressed Image Latents and Multimodal Large Language Models (Kao, 2024) View paper

Narrative

Core task: vision encoder training in multimodal language models. The field has evolved around several interconnected branches that address different facets of building effective multimodal systems. Vision Encoder Architecture and Design explores foundational choices such as selecting between CLIP-style and DINO-style encoders (CLIP to DINO[1]) or designing specialized architectures like Vcoder[2] and TokenPacker[3]. Training Paradigms and Optimization investigates how to effectively learn visual representations, spanning supervised fine-tuning, reinforcement learning approaches, and pretraining strategies exemplified by MM1 Pretraining[9] and PaLM-E[4]. Cross-Modal Alignment and Integration focuses on bridging vision and language modalities through projection layers and alignment mechanisms, while Multimodal Model Applications and Capabilities demonstrates the breadth of tasks these systems can handle, from visual question answering to embodied AI (OpenVLA[42]). Evaluation, Benchmarking, and Analysis provides the measurement frameworks needed to assess progress (Benchmark Survey[14]), and Efficient and Lightweight MLLM Design addresses deployment constraints through compression and distillation techniques like MobileVLM V2[26].

A particularly active tension exists between different training paradigms: while many works rely on supervised fine-tuning with large-scale instruction datasets (Visual Instruction Tuning[34]), recent efforts explore whether reinforcement learning can yield better alignment and reasoning capabilities. RL Better Than SFT[0] sits squarely within this emerging direction, investigating reinforcement learning for vision encoder training alongside related work like VLM-R1[44], which also examines RL-based optimization for multimodal

models. This contrasts with the dominant supervised paradigm seen in systems like Cambrian[5] and Qwen2-VL[28], which achieve strong performance through careful data curation and architectural choices. The central question is whether RL's ability to optimize for task-specific rewards can overcome the sample efficiency and stability challenges that have historically favored supervised approaches in vision-language pretraining.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Vlm-r1: A stable and generalizable r1-style large vision-language model

Authors: Shen, Haozhan, Liu Peng, Haozhan Shen, Li Jingcheng, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recently DeepSeek R1 has shown that reinforcement learning (RL) can substantially improve the reasoning capabilities of Large Language Models (LLMs) through a simple yet effective design. The core of R1 lies in its rule-based reward formulation, which leverages tasks with deterministic ground-truth answers to enable precise and stable reward computation. In the visual domain, we similarly observe that a wide range of visual understanding tasks are inherently equipped with well-defined ground-tru...

Relationship Analysis

Both papers belong to the same taxonomy category of applying reinforcement learning to reshape vision encoders in multimodal language models, specifically investigating how RL training affects visual representations compared to supervised fine-tuning. The original paper (RL makes MLLMs see better than SFT) conducts a comprehensive analysis of how DPO versus SFT reshapes vision encoders across diverse benchmarks, demonstrating that RL produces stronger localized visual representations and proposing PIVOT as a recipe for vision encoder training. The candidate paper (VLM-R1) focuses on developing a practical framework for applying R1-style RL (GRPO) to VLMs on specific visual understanding tasks (REC and OVD), emphasizing reward engineering, framework implementation, and task-specific performance improvements rather than the fundamental analysis of how RL reshapes visual representations across the vision encoder.

Contributions Analysis

Overall novelty summary. The paper investigates how reinforcement learning (specifically DPO) reshapes vision encoders in multimodal language models, contrasting this with supervised fine-tuning. It resides in the 'Reinforcement Learning for Vision Encoder Training' leaf under 'Training Paradigms and Optimization', which contains only two papers including this one. This is a notably sparse research direction within the broader taxonomy of 50 papers across 19 leaf nodes, suggesting the work addresses an emerging and under-explored area. The sibling paper (VLM-R1) also examines RL-based optimization for multimodal models, indicating nascent interest in this training paradigm.

The taxonomy reveals that most vision encoder training research concentrates on supervised and self-supervised pre-training (three papers in that leaf) or visual instruction tuning (three papers). The paper's leaf sits alongside these more populated directions within 'Training Paradigms and Optimization', which collectively address how to learn effective visual representations. Neighboring branches like 'Cross-Modal Alignment and Integration' and 'Vision Encoder Architecture and Design' focus on complementary aspects—bridging modalities and architectural choices—rather than the training strategy itself. The scope notes clarify that this leaf specifically excludes supervised methods and instruction tuning, positioning the work as an alternative training paradigm.

Among 30 candidates examined across three contributions, none were found to clearly refute any claim. The first contribution (systematic SFT vs. RL comparison) examined 10 candidates with zero refutable matches. The second contribution (RL producing stronger localized representations) and third contribution (PIVOT recipe) each examined 10 candidates, also with zero refutable matches. This suggests that within the limited search scope, the specific angle of analyzing how RL fundamentally reshapes vision encoder representations—through gradient visualization, ImageNet classification, and segmentation—appears relatively unexplored. However, the small candidate pool means this assessment reflects top-30 semantic matches rather than exhaustive coverage.

Given the sparse leaf occupancy and absence of refuting prior work among examined candidates, the contributions appear to occupy a relatively novel position within the limited search scope. The systematic comparison of training paradigms' effects on vision encoders, rather than just downstream task performance, represents a distinct analytical focus. However, the analysis is constrained by examining only 30 candidates, and the broader literature on RL for multimodal models may contain relevant work not captured in this semantic search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Systematic comparison of SFT and RL (DPO) effects on MLLMs and vision encoders

Description: The authors perform a comprehensive analysis comparing supervised finetuning (SFT) and reinforcement learning (DPO) in multimodal language models, examining not only downstream MLLM performance but also the impact on the vision encoder itself through vision-only tasks and gradient visualizations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Thinking With Videos: Multimodal Tool-Augmented Reinforcement Learning for Long Video Reasoning

URL: View paper

Brief Assessment

Thinking With Videos[72] focuses on video reasoning with tool-augmented learning and multi-task RL (DGRPO), not on comparing SFT vs DPO effects on vision encoders or analyzing visual representations.

---

2. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models

URL: View paper

Brief Assessment

Reason-RFT[51] focuses on visual reasoning tasks using GRPO-based RL, not DPO. The original paper specifically compares SFT vs. DPO effects on vision encoders through vision-only tasks and gradient analysis, which is not addressed in this candidate.

---

3. MindOmni: Unleashing Reasoning Generation in Vision Language Models with RGPO

URL: View paper

Brief Assessment

MindOmni[73] focuses on reasoning generation in vision-language models using RGPO (a different RL algorithm), not on comparing SFT vs DPO effects on vision encoders or analyzing visual representations through vision-only tasks.

---

4. Kimi-VL Technical Report

URL: View paper

Brief Assessment

Kimi-VL[74] focuses on building an efficient MoE vision-language model with long-context capabilities and agent performance. It does not provide a systematic comparative analysis of SFT versus RL/DPO effects on vision encoders through vision-only tasks or gradient visualizations.

---

5. Video-LMM Post-Training: A Deep Dive into Video Reasoning with Large Multimodal Models

URL: View paper

Brief Assessment

Video-LMM Post-Training[77] focuses on post-training methodologies for video-large multimodal models, not on comparing SFT and RL effects on vision encoders through vision-only tasks and gradient visualizations as in the original paper.

---

6. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering

URL: View paper

Brief Assessment

RL Outperforms SFT[78] focuses on audio question answering tasks using GRPO algorithm with large audio language models, not on vision encoders or multimodal vision-language models. The candidate does not address vision encoder analysis or DPO-based training for visual representations.

---

7. Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback

URL: View paper

Brief Assessment

Video RLAF[75] focuses on video-language alignment using RLAIF for video large multimodal models, not on systematic comparison of SFT vs. DPO effects on vision encoders through vision-only tasks and gradient analysis.

---

8. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding

URL: View paper

Brief Assessment

DocThinker[79] focuses on rule-based RL (GRPO) for document understanding with explainability, not on comparing SFT vs. DPO effects on vision encoders through vision-only tasks and gradient analysis.

---

9. Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models.

URL: View paper

Brief Assessment

Foundation Models Evolution[76] is a broad survey paper covering the evolution from LLMs to LMMs. It does not present empirical comparisons of SFT vs. DPO training strategies on MLLMs or analyze their differential effects on vision encoders through vision-only tasks and gradient visualizations.

---

10. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models

URL: View paper

Brief Assessment

SFT or RL[71] focuses on comparing SFT and RL (GRPO) for reasoning capabilities in LVLMs, not on analyzing their effects on vision encoders or vision-only tasks like the original paper does.

Contribution 2: Finding that RL produces stronger and more localized visual representations than SFT

Description: The paper demonstrates that reinforcement learning (specifically DPO) fundamentally reshapes visual representations in MLLMs, yielding stronger and more fine-grained localization capabilities compared to supervised finetuning, as evidenced by ImageNet classification, segmentation probing, and gradient analysis.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Simplevla-rl: Scaling vla training via reinforcement learning

URL: View paper

Brief Assessment

SimpleVLA-RL[60] focuses on vision-language-action models for robotic manipulation tasks, not on visual representation learning in multimodal language models. The candidate addresses action planning in embodied AI rather than visual encoder analysis in MLLMs.

---

2. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning

URL: View paper

Brief Assessment

R1-Omni[59] focuses on emotion recognition in omni-multimodal models (video+audio), not on visual representation quality in vision encoders for MLLMs. The paper does not analyze visual representations through ImageNet classification, segmentation probing, or gradient visualization as the original paper does.

---

3. Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning

URL: View paper

Brief Assessment

BigCharts-R1[52] focuses on chart reasoning tasks using GRPO-based RL with chart-specific rewards, not on analyzing how RL reshapes visual representations in general MLLMs compared to SFT.

---

4. Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback

URL: View paper

Brief Assessment

Visual RL[58] focuses on training MLLMs using only raw images through a reasoning-rendering-visual-feedback loop for image-to-code generation tasks. It does not compare RL versus SFT effects on visual encoder representations or analyze localization capabilities through ImageNet classification, segmentation probing, or gradient visualization as the original paper does.

5. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models

URL: View paper

Brief Assessment

Reason-RFT[51] does not analyze visual representations at the vision encoder level. It focuses on end-to-end visual reasoning performance without examining how RL reshapes visual features through ImageNet classification, segmentation probing, or gradient visualization as the original paper does.

6. VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning

URL: View paper

Brief Assessment

VideoRFT[54] focuses on video reasoning in MLLMs using reinforcement fine-tuning, not on comparing visual representation quality between RL and SFT methods as the original paper does.

7. UniRL: Self-Improving Unified Multimodal Models via Supervised and Reinforcement Learning

URL: View paper

Brief Assessment

UniRL[56] focuses on unified multimodal models (generation + understanding) using GRPO for post-training, not on analyzing how RL reshapes vision encoder representations compared to SFT in MLLMs.

8. RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning

URL: View paper

Brief Assessment

RewardMap[57] focuses on multi-stage RL for fine-grained visual reasoning on transit maps, not on comparing visual representation quality between RL and SFT in MLLMs.

9. Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner

URL: View paper

Brief Assessment

Patho-R1[53] focuses on pathology-specific multimodal models using RL (GRPO and DAPO) for diagnostic reasoning tasks, not on comparing visual representation quality between RL and SFT in general MLLMs as the original paper does.

10. Q-insight: Understanding image quality via visual reinforcement learning

URL: View paper

Brief Assessment

Q-Insight[55] focuses on image quality assessment using GRPO for score regression and degradation perception tasks, not on comparing RL versus SFT effects on visual representations in general multimodal models.

Contribution 3: PIVOT: a simple recipe for building strong vision encoders for MLLMs

Description: The authors propose PIVOT, a training method that applies RL-based preference optimization to evolve vision encoders for MLLMs. PIVOT-trained encoders outperform larger and more heavily pretrained counterparts while requiring less than 1% of the computational cost of standard vision pretraining.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Yi: Open Foundation Models by 01.AI

URL: View paper

Brief Assessment

Yi[63] focuses on building a multimodal model family through data engineering and vision-language alignment, not on RL-based preference optimization methods for evolving vision encoders. The candidate does not discuss preference optimization or RL training strategies for vision components.

2. Font-Agent: Enhancing Font Understanding with Large Language Models

URL: View paper

Brief Assessment

Font-Agent[69] focuses on font quality assessment and understanding using vision-language models with specialized modules for font analysis. It does not address RL-based preference optimization for training vision encoders in general multimodal language models, which is the core contribution of PIVOT.

3. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation

URL: View paper

Brief Assessment

Popen[67] focuses on preference-based optimization for LVLM-based reasoning segmentation tasks, not on training vision encoders for general multimodal language models. The candidate applies preference optimization to the entire LVLM for segmentation-specific objectives, whereas PIVOT specifically targets vision encoder evolution through RL-based preference optimization as an auxiliary training process.

4. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models

URL: View paper

Brief Assessment

Multi-Modal Preference Alignment[64] focuses on addressing modality degradation in MLLMs through preference alignment methods (DPO, SteerlM) applied to the entire MLLM system, not specifically on training vision encoders. The candidate's goal is to restore language capabilities after visual instruction tuning, whereas PIVOT specifically targets vision encoder evolution through RL-based preference optimization.

5. mdpo: Conditional preference optimization for multimodal large language models

URL: View paper

Brief Assessment

mDPO[62] focuses on multimodal preference optimization to address hallucination issues in MLLMs, not on training vision encoders. The candidate addresses unconditional preference problems in DPO for multimodal scenarios, while PIVOT uses RL-based preference optimization specifically to evolve vision encoders.

---

6. Aligning modalities in vision large language models via preference fine-tuning

URL: View paper

Brief Assessment

Aligning Modalities[68] focuses on preference optimization for reducing hallucinations in VLLMs through AI-generated dispreferences, not on training vision encoders as standalone components for MLLMs.

---

7. LPOI: Listwise Preference Optimization for Vision Language Models

URL: View paper

Brief Assessment

LPOI[70] focuses on listwise preference optimization for reducing hallucinations in vision-language models through object-aware masking and interpolation. This is fundamentally different from PIVOT's approach of using RL-based preference optimization to evolve vision encoders for improved visual representations in MLLMs.

---

8. Calibrated self-rewarding vision language models

URL: View paper

Brief Assessment

Calibrated Self-Rewarding[65] focuses on calibrating self-rewarding mechanisms for vision-language models through preference optimization to address hallucination, not on training vision encoders specifically for MLLMs through RL-based preference optimization as PIVOT does.

---

9. Direct preference optimization of video large multimodal models from language model reward

URL: View paper

Brief Assessment

DPO Video[61] focuses on video-based multimodal models using language model rewards for preference optimization in video question answering tasks, not on evolving vision encoders for general MLLMs through RL-based preference optimization as PIVOT does.

---

10. Modality-Balancing Preference Optimization of Large Multimodal Models by Adversarial Negative Mining

URL: View paper

Brief Assessment

Modality Balancing[66] focuses on preference optimization to address modality imbalance in multimodal models through adversarial negative mining, not on training vision encoders specifically. The candidate's approach targets balanced multimodal reasoning rather than vision encoder evolution.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] RL makes MLLMs see better than SFT View paper
- [1] From clip to dino: Visual encoders shout in multi-modal large language models View paper
- [2] Vcoder: Versatile vision encoders for multimodal large language models View paper
- [3] Tokenpacker: Efficient visual projector for multimodal llm View paper
- [4] Palm-e: An embodied multimodal language model View paper
- [5] Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs View paper
- [6] Bubogpt: Enabling visual grounding in multi-modal llms View paper
- [7] Visual cognition in multimodal large language models View paper
- [8] An introduction to vision-language modeling View paper
- [9] Mm1: methods, analysis and insights from multimodal llm pre-training View paper
- [10] The revolution of multimodal large language models: a survey View paper
- [11] Contextual Object Detection with Multimodal Large Language Models View paper
- [12] How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites View paper
- [13] Multimodal few-shot learning with frozen language models View paper
- [14] A survey on benchmarks of multimodal large language models View paper
- [15] Visual Representation Alignment for Multimodal Large Language Models View paper
- [16] Convllava: Hierarchical backbones as visual encoder for large multimodal models View paper
- [17] Mome: Mixture of multimodal experts for generalist multimodal large language models View paper
- [18] Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models View paper
- [19] LEO: Boosting Mixture of Vision Encoders for Multimodal Large Language Models View paper
- [20] Eagle: Exploring the design space for multimodal llms with mixture of encoders View paper
- [21] A survey of multimodel large language models View paper
- [22] Vita: Towards open-source interactive omni multimodal llm View paper
- [23] Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models View paper
- [24] Object detection with multimodal large vision-language models: An in-depth review View paper
- [25] Ovis: Structural Embedding Alignment for Multimodal Large Language Model View paper
- [26] MobileVLM V2: Faster and Stronger Baseline for Vision Language Model View paper
- [27] MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training View paper
- [28] Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution View paper
- [29] Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning View paper
- [30] LaCo: Efficient Layer-wise Compression of Visual Tokens for Multimodal Large Language Models View paper
- [31] MolPROP: Molecular Property prediction with multimodal language and graph fusion View paper

- [32] Multimodal large language models: A survey View paper
- [33] Words over pixels? rethinking vision in multimodal large language models View paper
- [34] Visual Instruction Tuning View paper
- [35] Elevating Visual Perception in Multimodal LLMs with Visual Embedding Distillation View paper
- [36] Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action View paper
- [37] METEOR: Multi-Encoder Collaborative Token Pruning for Efficient Vision Language Models View paper
- [38] Multimodal LLMs View paper
- [39] Imagined visual representations as multimodal embeddings View paper
- [40] Automated context-aware navigation support for individuals with visual impairment using multimodal language models in urban environments View paper
- [41] From large language models to large multimodal models: A literature review View paper
- [42] OpenVLA: An Open-Source Vision-Language-Action Model View paper
- [43] Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond View paper
- [44] Vlm-r1: A stable and generalizable r1-style large vision-language model View paper
- [45] Bliva: A simple multimodal llm for better handling of text-rich visual questions View paper
- [46] Bridging Compressed Image Latents and Multimodal Large Language Models View paper
- [47] InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD View paper
- [48] Question Aware Vision Transformer for Multimodal Reasoning View paper
- [49] Incorporating visual experts to resolve the information loss in multimodal large language models View paper
- [50] Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis View paper
- [51] Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models View paper
- [52] Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning View paper
- [53] Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner View paper
- [54] VideoRFT: Incentivizing Video Reasoning Capability in MLLMs via Reinforced Fine-Tuning View paper
- [55] Q-insight: Understanding image quality via visual reinforcement learning View paper
- [56] UniRL: Self-Improving Unified Multimodal Models via Supervised and Reinforcement Learning View paper
- [57] RewardMap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning View paper
- [58] Learning only with images: Visual reinforcement learning with reasoning, rendering, and visual feedback View paper
- [59] R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning View paper
- [60] Simplevla-rl: Scaling vla training via reinforcement learning View paper
- [61] Direct preference optimization of video large multimodal models from language model reward View paper
- [62] mdpo: Conditional preference optimization for multimodal large language models View paper
- [63] Yi: Open Foundation Models by 01.AI View paper
- [64] Multi-modal preference alignment remedies degradation of visual instruction tuning on language models View paper
- [65] Calibrated self-rewarding vision language models View paper
- [66] Modality-Balancing Preference Optimization of Large Multimodal Models by Adversarial Negative Mining View paper
- [67] Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation View paper
- [68] Aligning modalities in vision large language models via preference fine-tuning View paper
- [69] Font-Agent: Enhancing Font Understanding with Large Language Models View paper
- [70] LPOI: Listwise Preference Optimization for Vision Language Models View paper
- [71] SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models View paper
- [72] Thinking With Videos: Multimodal Tool-Augmented Reinforcement Learning for Long Video Reasoning View paper
- [73] MindOmni: Unleashing Reasoning Generation in Vision Language Models with RGPO View paper
- [74] Kimi-VL Technical Report View paper
- [75] Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback View paper
- [76] Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. View paper
- [77] Video-LMM Post-Training: A Deep Dive into Video Reasoning with Large Multimodal Models View paper
- [78] Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering View paper
- [79] Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding View paper