Multimodal LLM Alignment: Recent Techniques for Safely Aligning Text-Image-Audio Language Models

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in processing and generating content across text, images, and audio modalities. However, these powerful models present unique alignment challenges beyond those faced by text-only systems. This report surveys post-2023 techniques for safely aligning MLLMs with human preferences, focusing on methods that address hallucination, safety concerns, and cross-modal coherence.

Fundamentals of Multimodal Alignment

Multimodal alignment refers to ensuring that MLLMs generate outputs that are consistent, factual, and safe across different modalities. As defined in recent literature, alignment involves "aligning Large Multimodal Models (LMMs) with human values" to address issues related to "truthfulness, safety, o1-like reasoning, and alignment with human preference"[1]. The goal is to create models that accurately interpret multimodal inputs and generate appropriate, factual responses that align with human expectations and ethical considerations.

Unlike text-only models, MLLMs face unique challenges such as cross-modal consistency, where the model must ensure that text outputs accurately reflect visual or audio inputs. This has led to the development of specialized alignment techniques that extend beyond traditional text-only approaches.

RLHF-Based Approaches for Multimodal Alignment

Factually Augmented RLHF

One of the most significant advancements in multimodal alignment is Factually Augmented RLHF, which addresses the "hallucination" problem where models generate

text outputs not grounded in the visual context. This approach "augments the reward model with additional factual information such as image captions and ground-truth multi-choice options"[9], which helps mitigate reward hacking and improves factuality.

Researchers demonstrated that this approach achieved "remarkable improvement on the LLaVA-Bench dataset with the 94% performance level of the text-only GPT-4 (while previous best methods can only achieve the 87% level), and an improvement by 60% on MMHAL-BENCH over other baselines"[9]. This represents a substantial advancement in reducing multimodal hallucinations.

MM-RLHF: Large-Scale Preference Data

Building on traditional RLHF approaches, researchers introduced MM-RLHF in early 2025, "a dataset containing 120k fine-grained, human-annotated preference comparison pairs"[14]. This dataset represents a substantial advancement in terms of "size, diversity, annotation granularity, and quality"[14].

The MM-RLHF approach introduces two key innovations:

1. **Critique-Based Reward Model**: This model "generates critiques of model outputs before assigning scores, offering enhanced interpretability and more informative feedback compared to traditional scalar reward mechanisms"[14].

2. **Dynamic Reward Scaling**: This method "adjusts the loss weight of each sample according to the reward signal, thereby optimizing the use of high-quality comparison pairs"[14].

When applied to the LLaVA-ov-7B model, MM-RLHF produced a "19.5% increase in conversational abilities and a 60% improvement in safety"[14], demonstrating significant improvements across multiple dimensions of model performance.

Direct Preference Optimization Approaches

mDPO: Addressing Unconditional Preference

Traditional Direct Preference Optimization (DPO), while effective for text-only models, often struggles with multimodal inputs due to what researchers call "unconditional preference" where "the image condition is overlooked"[8]. To address this issue, researchers developed mDPO, "a multimodal DPO objective that prevents the over-prioritization of language-only preferences"[8].

The mDPO approach works by "jointly optimizing text and image preferences" and incorporates a "reward anchor to prevent a decrease in the likelihood of the chosen instance"[8]. This ensures that the model considers both textual and visual information when generating responses.

Combining Offline and Online Alignment Methods

Recent research has categorized alignment algorithms into two primary groups:

1. **Offline methods** (such as DPO): These utilize preference pairs collected prior to training.
2. **Online methods** (such as online-DPO): These involve sampling from the model during policy optimization.

Studies have shown that "combining offline and online methods can improve the performance of the model in certain scenarios"[15]. This hybrid approach offers the benefits of both pre-collected human preference data and dynamic, model-generated comparative examples.

Modality-Specific Alignment Techniques

Audio-Visual Alignment

For models that process both audio and visual inputs, specialized alignment techniques have emerged. The Dolphin audio-visual LLM introduces "a multi-scale adapter for spatial alignment and an interleaved merging module"[5] to ensure proper integration of both modalities.

This approach emphasizes the importance of both spatial and temporal alignment between audio and visual modalities. The "audio-visual multi-scale adapter" achieves spatial alignment, while the "audio-visual interleaved merging" ensures temporal alignment[5]. These mechanisms enable more coherent and accurate processing of multimodal inputs that contain both visual and audio components.

Speech-to-Text Alignment

For speech-focused MLLMs, researchers have developed lightweight alignment modules that can efficiently transform speech inputs to match the expected format of text-trained LLMs. One approach uses "one layer module and hundred hours of speech-text multitask corpus"[16] to achieve modal alignment. This enables models to "generate the same response when being fed spoken (audio) input instead of its text version"[20].

The AudioChatLlama project demonstrates this approach by extending "the capabilities of [Llama-2-chat] to the speech domain without compromising on the LLM's original capabilities"[20]. This is achieved by keeping the LLM completely frozen and only training the audio encoder component.

Text-to-Audio Alignment

For text-to-audio generation, the BATON system integrates "three modules: (1) An audio generation unit using LLM-augmented prompts, with human-scored annotations; (2) A

reward model trained on" these human preferences[18]. This approach has demonstrated "gains of +2.3% and +6.0% in CLAP scores for integrity and temporal relationship tasks, respectively"[18].

The reward model in this system uses both text and audio encoders to evaluate the alignment between the provided text and generated audio, providing a scalar reward that "signifies human preference"[18]. This enables better alignment between textual descriptions and the audio content they should generate.

Safety-Focused Alignment Techniques

SafeText for Safe Image Generation

For text-to-image models, researchers have developed SafeText, "a novel alignment method that fine-tunes the text encoder rather than the diffusion module"[2]. This approach is particularly focused on preventing harmful image generation from unsafe prompts.

By adjusting the text encoder, "SafeText significantly alters the embedding vectors for unsafe prompts, while minimally affecting those for safe prompts"[2]. This ensures that "the diffusion module generates non-harmful images for unsafe prompts while preserving the quality of images for safe prompts"[2]. This approach has proven more effective than methods that modify the diffusion module directly.

Evaluating Then Aligning (ETA) Framework

The ETA framework introduces a two-stage approach for vision-language model safety:

1. **Visual Safety Evaluation**: Assesses the safety of multimodal inputs.
2. **Shallow and Deep Alignment**: Ensures outputs remain both safe and useful.

This framework is designed to operate at inference time, "thus enabling real-time alignment of VLM outputs"[4]. A key advantage of this approach is that it performs "safety checks dynamically without additional data or training, making the approach efficient and flexible"[4].

Multimodal Situational Safety

A novel safety challenge termed "Multimodal Situational Safety" explores "how safety considerations vary based on the specific situation in which the user or agent is engaged"[11]. This recognizes that "for an MLLM to respond safely-whether through language or action-it often needs to assess the safety implications of a language query within its corresponding visual context"[11].

To evaluate this capability, researchers developed the Multimodal Situational Safety benchmark (MSSBench), comprising "1,820 language query-image pairs, half of which the image context is safe, and the other half is unsafe"[11]. Current findings reveal that "MLLMs struggle with this nuanced safety problem in the instruction-following setting"[11], highlighting an important area for future research.

Challenges and Open Problems

Hallucination Bias

One of the most significant challenges in multimodal alignment is hallucination bias, where "AI language models generate outputs that are not grounded in reality or are based on incomplete or biased data sets"[7]. This issue becomes particularly problematic in multimodal contexts, where models might "hallucinate" features in one modality based on expectations from another.

The consequences of such hallucinations can be severe: "In healthcare, an AI diagnosing tool might hallucinate symptoms that aren't present, leading to misdiagnoses. In autonomous vehicles, a bias-induced hallucination could cause a car to perceive a non-existent obstacle, resulting in an accident"[7].

Modality Imbalance

Another challenge is modality imbalance during alignment, where models may "prioritize language-only preferences and overlook the image condition"[8]. This leads to suboptimal performance and increased hallucination, as the model fails to properly ground its responses in visual or audio inputs.

Research has found that "multimodal LLMs can achieve similar performance even when all images are removed from the multimodal preference data during DPO"[8], indicating that many current alignment approaches may not be effectively utilizing visual information.

Cross-Modal Coherence

Ensuring coherence across modalities remains a significant challenge. Unlike text-only models, MLLMs must maintain consistency between what they "see" or "hear" and what they generate as output. This requires specialized alignment techniques that explicitly account for relationships between modalities.

Evaluation and Benchmarking

Several specialized benchmarks have emerged to evaluate multimodal alignment:

1. **MMHAL-BENCH**: Focuses specifically on penalizing hallucinations in multimodal models[9].
2. **MSSBench**: Evaluates situational safety across 1,820 language-image pairs[11].

3. **LLaVA-Bench**: Measures general multimodal capabilities and alignment[9].

These benchmarks provide standardized ways to measure alignment quality across different dimensions, including factuality, safety, and coherence between modalities.

Future Directions

Unified Cross-Modal Alignment

Future research is likely to focus on developing unified frameworks that can simultaneously align across text, image, audio, and potentially other modalities. This would enable more coherent and consistent multimodal experiences.

Real-Time Alignment

Approaches like the ETA framework demonstrate the potential for real-time alignment at inference time, rather than relying solely on pre-training or fine-tuning. This direction appears promising for addressing emerging safety concerns without requiring complete model retraining.

Multi-Agent Alignment Systems

The MSSBench research suggests that "multi-agent pipelines to coordinately solve safety challenges" show "consistent improvement in safety over the original MLLM response"[11]. This collaborative approach to alignment may enable more robust safety mechanisms than single-agent systems.

Conclusion

Multimodal LLM alignment has seen significant advances since 2023, with researchers developing specialized techniques to address the unique challenges of aligning models across text, image, and audio modalities. RLHF-based approaches like Factually Augmented RLHF and MM-RLHF have demonstrated significant improvements in reducing hallucinations and improving alignment with human preferences.

DPO variants such as mDPO have addressed the challenge of modality imbalance, while modality-specific techniques have improved alignment for audio-visual, speech-to-text, and text-to-audio applications. Safety-focused techniques like SafeText and the ETA framework provide mechanisms to ensure that multimodal models generate appropriate and non-harmful content.

Despite these advances, significant challenges remain, including hallucination bias, modality imbalance, and cross-modal coherence. Future research will likely focus on unified cross-modal alignment approaches, real-time alignment techniques, and multi-agent systems for more robust safety guarantees.

As multimodal models continue to grow in capability and deployment, effective alignment techniques will be essential for ensuring these powerful systems remain safe, helpful, and aligned with human values across all modalities they can process and generate.

Sources

[1] Aligning Multimodal LLM with Human Preference: A Survey - arXiv https://arxiv.org/html/2503.14504v1

[2] SafeText: Safe Text-to-image Models via Aligning the Text Encoder https://openreview.net/forum?id=T7kThJhl02

[3] [PDF] Aligning Large Multimodal Models with Factually Augmented RLHF https://aclanthology.org/2024.findings-acl.775.pdf

[4] Advancing Safety in Vision-Language Models: Semantic-Based ... https://blog.aim-intelligence.com/advancing-safety-in-vision-language-models-semantic-based-interpretation-and-real-time-alignment-296f5e71415c

[5] Aligned Better, Listen Better for Audio-Visual Large Language Models
https://openreview.net/forum?id=1SYUKPeM12

[6] [PDF] A Hybrid Approach to Audio-to-Score Alignment
http://eecs.qmul.ac.uk/~simond/pub/2019/Agrawal-Dixon-ML4MD-2019.pdf

[7] The Risks of Hallucination Bias in AI Language Models
https://www.expresscomputer.in/artificial-intelligence-ai/the-risks-of-hallucination-bias-in-ai-language-models/104013/

[8] mDPO - Fei Wang https://feiwang96.github.io/mDPO/

[9] Aligning Large Multimodal Models with Factually Augmented RLHF
https://arxiv.org/abs/2309.14525

[10] [2411.17040] Multimodal Alignment and Fusion: A Survey - arXiv
https://arxiv.org/abs/2411.17040

[11] Multimodal Situational Safety https://mssbench.github.io

[12] [PDF] Aligning Multimodal LLM with Human Preference: A Survey - arXiv
https://arxiv.org/pdf/2503.14504.pdf

[13] AVicuna: Audio-Visual LLM with Interleaver and Context-Boundary ...
https://arxiv.org/html/2403.16276v1

[14] MM-RLHF: The Next Step Forward in Multimodal LLM Alignment
https://arxiv.org/abs/2502.10391

[15] [PDF] UNDERSTANDING ALIGNMENT IN MULTIMODAL LLMS ...
https://openreview.net/pdf/dfb3ff433f662041508bf2dc184f9f07e933bc53.pdf

[16] Transferable speech-to-text large language model alignment module
https://arxiv.org/html/2406.13357v1

[17] [PDF] MM-LLMs: Recent Advances in MultiModal Large Language Models
https://aclanthology.org/2024.findings-acl.738.pdf

[18] [PDF] Aligning Text-to-Audio Model Using Human Preference Feedback
https://www.ijcai.org/proceedings/2024/0502.pdf

[19] Understanding Multimodal LLMs - by Sebastian Raschka, PhD
https://magazine.sebastianraschka.com/p/understanding-multimodal-llms

[20] [PDF] AudioChatLlama: Towards General-Purpose Speech Abilities for LLMs
https://aclanthology.org/2024.naacl-long.309.pdf

[21] Multimodal Situational Safety - arXiv https://arxiv.org/html/2410.06172v1

[22] Semantic Alignment for Multimodal Large Language Models
https://dl.acm.org/doi/10.1145/3664647.3681014

[23] SEA: Low-Resource Safety Alignment for Multimodal Large ... - arXiv
https://arxiv.org/html/2502.12562v1

[24] Aligning Large Multimodal Models with Factually Augmented RLHF
https://huggingface.co/papers/2309.14525

[25] Paper page - SafeVLA: Towards Safety Alignment of Vision ...
https://huggingface.co/papers/2503.03480

[26] MMDT: Decoding the Trustworthiness and Safety of Multimodal...
https://openreview.net/forum?id=qIbbBSzH6n

[27] BradyFU/Awesome-Multimodal-Large-Language-Models - GitHub
https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

[28] Multimodal AI: Breaking Down Barriers Between Text, Image, Audio ...
https://www.ayadata.ai/multimodal-ai-breaking-down-barriers-between-text-image-audio-and-video/

[29] Aligning Large Multimodal Models with Factually Augmented RLHF
https://aclanthology.org/2024.findings-acl.775/

[30] Cross-Modal Safety Mechanism Transfer in Large Vision-Language ...
https://openreview.net/forum?id=45rvZkJbuX

[31] [PDF] Safety of Multimodal Large Language Models on Images and Text
https://www.ijcai.org/proceedings/2024/0901.pdf

[32] [PDF] Multi-modal Preference Alignment Remedies Degradation of Visual ...
https://aclanthology.org/2024.acl-long.765.pdf

[33] Virtual DPO Service - Online Virtual Data Protection Officer - DPO India
https://www.dpo-india.com/Services/virtual-dpo-service/

[34] Soundwave: Less is More for Speech-Text Alignment in LLMs
https://huggingface.co/papers/2502.12900

[35] Audio Alignment - Steinberg Help
https://www.steinberg.help/r/nuendo/13.0/en/cubase_nuendo/topics/parts_events/parts_and_events_tempo_matching_audio_alignment_c.html

[36] [PDF] Addressing Bias and Hallucination in Large Language Models
https://aclanthology.org/2024.lrec-tutorials.12.pdf

[37] mDPO: Conditional Preference Optimization for Multimodal Large ...
https://aclanthology.org/2024.emnlp-main.460/

[38] trl/docs/source/online_dpo_trainer.md at main - GitHub
https://github.com/huggingface/trl/blob/main/docs/source/online_dpo_trainer.md

[39] run-audio-alignment-mistral.sh - mesolitica/multimodal-LLM - GitHub
https://github.com/mesolitica/multimodal-LLM/blob/master/audio-only/run-audio-alignment-mistral.sh

[40] [PDF] Standards for Safe Listening – how they align and how some differ
https://www.efhoh.org/wp-content/uploads/2021/04/Article-Laureyns-Masahito-Best-safe-listening-standards-ent-may20.pdf

[41] Hallucination, Inconsistency, and Bias: The Essential Guide
https://www.nightfall.ai/ai-security-101/hallucination-inconsistency-and-bias

[42] [PDF] MDPO: Conditional Preference Optimization for - ACL Anthology
https://aclanthology.org/2024.emnlp-main.460.pdf

[43] Online DPO Trainer - Hugging Face
https://huggingface.co/docs/trl/main/en/online_dpo_trainer

[44] [PDF] OneLLM: One Framework to Align All Modalities with Language
https://openaccess.thecvf.com/content/CVPR2024/papers/Han_OneLLM_One_Framework_to_Align_All_Modalities_with_Language_CVPR_2024_paper.pdf

[45] AlignBench Dataset - Papers With Code
https://paperswithcode.com/dataset/alignbench

[46] How's everyone enjoying Multimodal Rlhf? : r/outlier_ai - Reddit
https://www.reddit.com/r/outlier_ai/comments/1gibn6f/hows_everyone_enjoying_multimodal_rlhf/

[47] Guide to Reinforcement Learning from Human Feedback (RLHF)
https://encord.com/blog/guide-to-rlhf/

[48] mDPO: Conditional Preference Optimization for Multimodal Large ...
https://huggingface.co/papers/2406.11839

[49] Quality measures for protein alignment benchmarks - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC2853116/

[50] [PDF] RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from ...
https://openaccess.thecvf.com/content/CVPR2024/papers/Yu_RLHF-V_Towards_Trustworthy_MLLMs_via_Behavior_Alignment_from_Fine-grained_Correctional_CVPR_2024_paper.pdf

[51] EchoInk-R1: Exploring Audio-Visual Reasoning in Multimodal LLMs ...
https://arxiv.org/html/2505.04623

[52] Multimodal Recommendation Dialog with Subjective Preference
https://aclanthology.org/2023.findings-acl.217/

[53] LLM-Assisted Real-Time Anomaly Detection for Safe Visual ... - arXiv
https://arxiv.org/abs/2403.12415

[54] A benchmark of multiple sequence alignment programs upon ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC1087786/

[55] Kwai-YuanQi/MM-RLHF: The Next Step Forward in Multimodal LLM ...
https://github.com/Kwai-YuanQi/MM-RLHF

[56] Aligning Multimodal LLM with Human Preference: A Survey
https://huggingface.co/papers/2503.14504

[57] Multimodal Situational Safety - OpenReview
https://openreview.net/forum?id=I9bEi6LNgt

[58] [PDF] A Survey on Multimodal Large Language Models - arXiv
https://arxiv.org/pdf/2306.13549.pdf

[59] Understanding Alignment in Multimodal LLMs: A Comprehensive ...
https://openreview.net/forum?id=49qqV4NTdy&noteId=BmpGFgu040

[60] Multimodal Alignment and Fusion: A Survey - alphaXiv
https://www.alphaxiv.org/overview/2411.17040

[61] MM-SafetyBench Dataset | Papers With Code
https://paperswithcode.com/dataset/mm-safetybench

[62] A Survey on Multimodal Large Language Models - Hugging Face
https://huggingface.co/papers/2306.13549

[63] [PDF] Benchmarking Multi-Modal Entity Alignment - ACL Anthology
https://aclanthology.org/2025.coling-main.582.pdf

[64] A Survey of Multimodal Learning: Methods, Applications, and Future https://dl.acm.org/doi/10.1145/3713070

[65] SafeBench: A Safety Evaluation Framework for Multimodal Large ... https://arxiv.org/abs/2410.18927

[66] A Survey on Multimodal Large Language Models - SciSpace https://scispace.com/papers/a-survey-on-multimodal-large-language-models-2lthj9rg

[67] Multi-modal Entity Alignment - Papers With Code https://paperswithcode.com/task/multi-modal-entity-alignment

[68] LLaVA-RLHF https://llava-rlhf.github.io

[69] Enhance speech synthesis and video generation models with RLHF ... https://aws.amazon.com/blogs/machine-learning/enhance-speech-synthesis-and-video-generation-models-with-rlhf-using-audio-and-video-segmentation-in-amazon-sagemaker/

[70] Paper page - Enhancing the Reasoning Ability of Multimodal Large ... https://huggingface.co/papers/2411.10442

[71] How Does Vision-Language Adaptation Impact the Safety of ... - arXiv https://arxiv.org/html/2410.07571

[72] Multimodal Alignment and Fusion: A Survey - arXiv https://arxiv.org/html/2411.17040v1

[73] Understanding Everything About Alignment in Multimodal Machine ... https://www.linkedin.com/pulse/understanding-everything-alignment-multimodal-machine-kahar-zjokf

[74] A survey on multimodal large language models - Oxford Academic https://academic.oup.com/nsr/article/11/12/nwae403/7896414

[75] [PDF] The Revolution of Multimodal Large Language Models: A Survey https://aclanthology.org/2024.findings-acl.807.pdf

[76] [2306.13549] A Survey on Multimodal Large Language Models - arXiv https://arxiv.org/abs/2306.13549

[77] AlignMMBench: Evaluating Chinese Multimodal Alignment in Large ... https://arxiv.org/abs/2406.09295