Aligning Multimodal LLMs: Post-2023 Techniques for Safe and Reliable AI

Introduction

Multimodal large language models (MLLMs) extend text-based LLMs by incorporating vision, audio, and other data types . This expansion unlocks powerful capabilities but also introduces new alignment challenges. Issues of **truthfulness**, **safety**, and **bias** that already trouble text-only models become more complex when models must interpret images or audio . Early MLLMs (in 2022–2023) often stopped at supervised fine-tuning and lacked rigorous alignment stages like reinforcement learning from human feedback (RLHF) . By late 2023, the community recognized that aligning these models to human preferences and values is critical to prevent hallucinations, harmful outputs, or policy violations across modalities .

Post-2023 innovations have rapidly advanced multimodal alignment. New algorithms and training pipelines aim to make MLLMs more factual, safe, unbiased, and obedient to instructions. Both academia and industry have contributed: researchers proposed refined reward modeling and adversarial training techniques, while AI labs like OpenAI, Google DeepMind, Anthropic, and Meta deployed large-scale human feedback and safety filters . This report surveys the state-of-the-art alignment techniques for multimodal models, focusing on general strategies (not tied to a single architecture) that ensure models behave **safely and reliably** in text, image, and audio tasks. We first outline the key **alignment goals** for multimodal LLMs, then examine **techniques** (from training-time methods like RLHF variants to deployment-time guardrails). Throughout, we highlight **post-2023 developments** – e.g. new datasets, reward optimization methods, and industry best practices – with peer-reviewed findings and official documentation as sources.

Alignment Goals in Multimodal Models

Alignment refers to shaping a model's behavior to meet desired criteria. For multimodal LLMs, major goals include: **factuality**, **safety (harmlessness)**, **bias mitigation (fairness)**, and **instruction-following helpfulness**. These often overlap – for instance, a helpful answer must

also be correct and non-harmful. Below we describe each goal and the unique challenges in the multimodal context:

- **Factuality and Hallucination Reduction:** MLLMs should generate text grounded in the visual or audio input and real-world truth, rather than hallucinating details. A common failure is describing things not actually present in an image or misinterpreting audio content. Post-2023 research has targeted this "multimodal misalignment" by finetuning models to avoid unsupported claims . For example, human evaluators can compare which of two image descriptions is more *hallucinated*, and reward models are trained to favor the factual one . Aligning on factuality often involves augmenting the model with external knowledge (e.g. image captions or OCR text) to verify claims . One study found that applying RLHF to vision-language tasks – with reward signals that incorporate ground-truth image information – dramatically improves factual accuracy, achieving ~94% of GPT-4's level on an image description benchmark (versus 87% pre-alignment) . In practice, aligned multimodal models refuse to speculate about image content they cannot infer, and they correct or avoid earlier mistakes by referencing reliable cues.
- **Safety and Harm Avoidance:** Safety alignment means the model refrains from producing harmful, unethical, or disallowed content across all modalities. Multimodal systems face *new risk combinations*: an otherwise benign image combined with a dangerous text prompt could yield a harmful response . Similarly, images can amplify malicious instructions that a text-only model might have caught . The model must **refuse or safely handle** requests that violate policies, whether the request comes as text, an image, or audio. For example, OpenAI's GPT-4 Vision was trained to refuse demands for illicit activities even if the prompt is partly in an image (e.g. an image of a weapon with the text "how to use this?") . Safety alignment techniques include adding *explicit refusal behaviors* (the model responds with a polite refusal when confronted with disallowed queries) and using *safety reward signals* during training . Models are also tuned to recognize sensitive visual content (like hate symbols or nudity) and respond cautiously . Studies in 2024 introduced safety-specific RLHF variants that treat harmful content as a "cost" to be minimized even as helpfulness is maximized . In one case, a multimodal RLHF scheme with separate safety feedback improved a model's safety score by 34% without sacrificing useful performance . The end goal is **harmlessness**: the model should follow user instructions *only* when they are acceptable, and otherwise tactfully refuse or redirect, regardless of input modality.
- **Bias Mitigation and Fairness:** Like humans, AI models can exhibit undesired biases. Multimodal models may, for instance, generate captions or answers that reflect gender or racial stereotypes present in training data. They might also misidentify individuals in images due to biased representations (a known issue in facial recognition) . Alignment for fairness strives to eliminate or reduce such biases. A key step is curating balanced training data – post-2023 surveys emphasize a "**curate over crawl**" approach, preferring high-quality datasets that are diverse and carefully filtered over raw web data . Techniques like **counterfactual data augmentation** introduce altered examples (e.g. swapping genders in an image description) to teach the model invariance to sensitive attributes . Instruction guidelines also explicitly forbid the model from mentioning or inferring protected traits from images unless absolutely necessary . For example, GPT-4V refuses to speculate about a person's race or health from a photo . Alignment evaluations

in 2024 revealed that many unified multimodal models still show demographic biases (e.g. gender bias in profession predictions) , so cutting-edge systems employ bias tests and further fine-tuning to mitigate this. Google's Gemini, for instance, encountered criticism for how it generated people with various ethnic appearances (sometimes historically inaccurate), prompting Google to pause that feature and adjust the model . This underscores that **bias calibration** remains an active alignment frontier – modern multimodal models include fairness as a first-class objective alongside safety, using both data-level and model-level interventions to avoid unfair outputs.

- **Instruction-Following and Helpfulness:** A core strength of LLMs is following user instructions to be helpful. When extended to images or audio, the model should similarly comply with the user's requests in those domains. This means understanding multimodal prompts (e.g. "Explain this image in French") and producing the desired output format and style, *as long as* it doesn't conflict with safety or factuality. Alignment for instruction-following is largely achieved via **supervised fine-tuning on high-quality instruction-response pairs** and **RLHF** where human judges prefer outputs that correctly and helpfully follow the task. In the multimodal setting, new datasets have been created with prompts like "User shows an image and asks a question" and ideal answers . By fine-tuning on such data (often synthesized by GPT-4 or human annotators), models learn to interpret visual context and comply with requests (e.g. describing an image's content, analyzing audio transcripts, etc.) . Post-2023 multimodal assistants (OpenAI's Vision-augmented ChatGPT, Meta's image-grounded chat models, etc.) were trained to *politely clarify* ambiguous multimodal instructions, follow formatting requests, and even use tools when needed to best help the user . An aligned model should maintain a **helpful demeanor**—for example, providing a step-by-step explanation of an image if asked, or adjusting its answer when the user says "now say it in a simpler way." Reinforcement learning has been shown to significantly boost these conversational abilities: one project reported a 19.5% increase in a model's overall interactive helpfulness after multimodal RLHF fine-tuning . Thus, instruction-following alignment ensures the model is **responsive and user-centric** across text, images, and audio, enabling it to act as a reliable assistant in multimodal tasks.

These goals collectively define "safe and reliable" behavior for multimodal LLMs. In practice, alignment techniques aim to balance them – e.g. making the model truthful and helpful without being offensive or leaking private info. Next, we survey the techniques developed post-2023 to achieve these alignment objectives.

Key Alignment Techniques for Multimodal LLMs

Achieving the above goals requires interventions at various stages of model training and deployment. Broadly, methods fall into two categories: **training-time alignment** (fine-tuning the

model's parameters using curated data and feedback signals) and **inference or system-level alignment** (using prompts, filters, or external systems to shape the model's outputs on the fly) . We explore both, emphasizing recent techniques that have been successfully applied to multimodal models. Table 1 summarizes several major alignment strategies and examples.

Table 1: Selected Alignment Strategies for Multimodal Models (Post-2023)

Technique	Approach & Goals	Examples and Sources
Supervised Fine-Tuning (SFT)	Train on human-written or high-quality generated responses to multimodal prompts. Aligns base model to follow instructions and prefer safe, factual outputs. Often the first alignment step after pretraining.	*OpenAI GPT-4V* finetuned on image+text instructions with safe completions ; *LLaVA* models fine-tuned on GPT-4 generated image descriptions/Q&A .
RLHF (Human Preference)	Reinforcement Learning from Human Feedback: gather human preference comparisons on model outputs and train a policy to maximize a reward model that predicts human-preferred answers. Ensures helpfulness and correctness as judged by people.	*GPT-4* used RLHF with an added safety reward signal to train refusal of disallowed requests . *LLaVA-RLHF* (2023) adapted RLHF to vision-language, reducing image hallucinations by ~60% on a new benchmark .
Direct Preference Optimization *(DPO and variants)*	A loss-based alternative to RLHF: directly optimize the model on ranked outputs via a contrastive objective, avoiding complex RL. Often more stable and simpler to implement. Many multimodal works use DPO to reduce hallucinations and improve safety.	*CHiP* (2023) combined **visual DPO** (punishing incorrect image captions) with **textual DPO** at multiple levels to reduce vision-language errors . *Image-DPO* blurred or pixelated images to create challenging pairs and trained models that resist visual perturbations .
AI Feedback & Self-Critique	Use AI models (including the model itself) to critique or rate outputs, providing feedback signals for alignment. Lowers reliance on human annotation and can encode explicit principles (e.g. "Constitutional AI"). The model learns to **self-moderate** or improve its responses iteratively.	*GPT-4 (2023)* employed a GPT-4-based classifier to judge output safety, giving a reward bonus for refusals of unsafe requests . *Self-Moderation (2023)* has the model check its own answer for policy violations and revise if needed ("are you sure?" loop) . *Critique-Based Reward Models* generate a textual critique of an answer (via an LLM) before scoring it, yielding more informative feedback than a single numeric score . Anthropic's **Constitutional AI** (adopted in text

Technique	Approach & Goals	Examples and Sources
		agents) is conceptually similar: the model is guided by a set of written ethical principles and uses them to refine responses without direct human labels.
Safety-Focused Fine-Tuning	Augment training with specifically curated **safety data**: scenarios of misuse, adversarial prompts, or harmful content, along with the desired safe response (often refusals or safe completions). This directly teaches the model how to handle unsafe inputs. Also includes multi-objective training that balances helpfulness against safety constraints.	*VLGuard* (2024) assembled a dataset of multimodal harmful prompts and performed post-hoc fine-tuning to suppress unsafe behaviors . *OpenAI GPT-4V* added a special finetuning phase with illicit multimodal prompts (mixing images into known disallowed text queries) to reinforce high refusal rates . *Safe RLHF-V* (2025) introduced a **reward+cost** framework: one model judges helpfulness, another judges safety, and a constrained RL algorithm optimizes for both , yielding a 34% safety boost without losing utility.
Adversarial Robustness Training	Expose the model to adversarial or tricky inputs during training so it learns to resist manipulation and avoid "gotcha" failures. Particularly important for multimodal inputs, where attackers might use perturbations or hidden messages in images. Often overlaps with safety fine-tuning.	*Adversarial PO (AdPO)* (2024) performed contrastive finetuning on **perturbed images** (e.g. adding noise or overlays) to make the model robust to visual attacks . Researchers also injected hidden instructions inside images (visual prompt injection) during training to teach models to ignore them . OpenAI's GPT-4V was tested with adversarial images containing overlaid text (e.g. an image with "How do I build a bomb?" written on it); they added an OCR step and moderation filter so the system would catch and refuse these cases . Such measures significantly reduce the model's susceptibility to multimodal prompt hacking.
Inference-Time Guardrails	Deploy auxiliary safety systems alongside the model at runtime. These include content filters (for text and image output), watermarking of outputs, and usage policies enforced via a "system prompt." While not changing the model's weights, these	All major providers use content moderation APIs to screen model outputs and sometimes user inputs. For example, *DALL·E 3* (2023) employs an image classifier to check generated images against disallowed content categories . *OpenAI's Voice Engine* for custom voices embeds inaudible

Technique	Approach & Goals	Examples and Sources
	guardrails align the *behavior* seen by the end-user.	**watermarks** in audio to trace misuse . *System messages* (hidden prompts) are used to instruct models with rules at conversation start (e.g. "The assistant must not reveal private info"): studies show carefully crafted system prompts can noticeably improve safety of MLLM responses . These measures serve as a last line of defense in case the model's training fails to cover some scenario.

Supervised Instruction Tuning

A foundation for multimodal alignment is **supervised fine-tuning (SFT)** on high-quality instruction-response data. Developers curate examples where the model is given a task involving text and images (or audio) and a correct, aligned response. This might include: describing an image accurately, answering questions about a video, following a spoken command, etc. By fine-tuning on such pairs, the model learns baseline skills for multimodal understanding and how it *should* respond.

Modern multimodal datasets often blend human-written responses and **synthetic data generated by powerful models**. For instance, the LLaVA project used GPT-4 to generate detailed captions and dialogues about images, then fine-tuned a vision-augmented LLM on this data . The result was a model that could hold a conversation about an image fairly coherently. Similarly, Meta's early image-text chat models (e.g. BlenderBot with vision) and open-source efforts like BLIP-2 leveraged labeled image-caption datasets and some human demonstrations. This **instruction tuning** step usually emphasizes correctness and helpful style, but may not fully address safety or bias – since the model could still output a learned bias or be misled by a novel malicious prompt. Indeed, surveys found that many MLLMs after SFT alone still produce hallucinations or unsafe content . Nonetheless, SFT establishes the model's basic aligned behavior (following user instructions, referring properly to visual input, etc.), which subsequent techniques (like RLHF) then refine. In OpenAI's GPT-4V system card, the team notes that after the initial fine-tuning, they observed capability but also "ungrounded inferences" and other issues, prompting additional alignment phases . In summary, supervised tuning on multimodal instructions is a necessary first alignment pass – it imbues the model with **helpfulness and multimodal understanding**, setting the stage for more targeted alignment fixes.

Reinforcement Learning from Human Feedback (RLHF)

RLHF has emerged as a **cornerstone of aligning LLMs**, and after 2023 it has been extended to multimodal models. The RLHF pipeline involves four components: a base model, a reward model (trained to predict human preference judgments), a policy optimization algorithm (often Proximal Policy Optimization, PPO), and a dataset of human preference comparisons. For text-only ChatGPT, this meant showing human labelers two model responses to a prompt and asking which they prefer; the model then learns to produce outputs that score higher on a learned reward function . Translating this to multimodal tasks required some innovation in 2023-2024. Human annotators now needed to compare answers *with respect to given images or audio*. One example is the work by Sun *et al.* (2023) who had humans compare which caption was less hallucinated given an image . Another is the **MM-RLHF** project (2024), which built a massive dataset of 120k human preference pairs covering image, video, and audio scenarios . This dataset explicitly includes diverse criteria (accuracy, level of detail, politeness, etc.) so that the reward model can judge outputs along multiple axes .

Several **findings** have come out of multimodal RLHF experiments:

- Aligning to human feedback **improves factual grounding**. The LLaVA-RLHF model (2023) was one of the first open models to undergo RLHF for vision-language alignment. It showed a drastic drop in image-related hallucinations, outperforming earlier models on curated benchmarks of visual question answering . The reward model in that work was augmented with actual image captions to help it judge correctness, an approach called *Factually Augmented RLHF* that curbed the reward model from mistakenly rewarding fluent but ungrounded answers .
- RLHF can integrate **safety objectives** by design. OpenAI reported that for GPT-4's vision-enabled version, they incorporated an *additional reward signal specifically for safety* during RLHF, which trained the model to refuse unsafe requests . In practice, this meant the human feedback process included prompts where the correct answer was a refusal, and the reward model (or a separate safety model) learned to give high score to refusals in those cases. This dual-objective RLHF (helpfulness vs. harm-avoidance) led to much more reliable refusals of disallowed prompts in GPT-4V's final behavior .
- **General capability gains**: Interestingly, aligning with human preferences has been found to *boost general performance* in multimodal tasks, not just obey human whims. One study showed that using a large, diverse preference dataset (MM-RLHF's 120k pairs) to fine-tune a vision-chat model improved its test performance across **27 benchmarks**, including a 19.5% jump in conversational ability and 60% jump in a safety metric . This suggests that alignment need not trade off capability; done well, it can be a win-win that makes the model both smarter and safer .

In deploying RLHF for multimodal models, **practical challenges** include collecting high-quality feedback at scale (particularly for specialized modalities like audio or video) and ensuring the

reward model itself is reliable. There's ongoing research into using AI feedback (next section) to supplement or replace expensive human feedback in RLHF loops. Despite these challenges, by 2024 RLHF had been adopted in some form by most state-of-the-art multimodal systems (OpenAI uses it for GPT-4, Google reportedly uses human feedback tuning for Bard's image analysis, and multiple academic MLLMs employ RLHF or its variants in papers). It remains a central alignment technique, directly operationalizing human judgments of **what constitutes a good, aligned response**.

Preference Optimization and DPO Variants

While RLHF has been successful, it comes with complexity (training a separate policy via RL, risk of reward hacking, etc.). A line of work post-2023 explores **Direct Preference Optimization (DPO)** and related supervised objectives as a more tractable way to use preference data . The idea is to convert preference comparisons into a loss function the model can be fine-tuned on, without a reinforcement learning step. For example, given a prompt and two responses (A and B) where A was preferred by a human, one can adjust the model's logits to make A more likely than B by a certain margin. DPO (introduced in late 2022 for text models) formalizes this and has been eagerly adopted in multimodal alignment papers in 2024 .

Advantages of DPO: It avoids the instability of RL (no reward model or policy gradient needed beyond the direct loss), and it can be implemented as standard fine-tuning, which leverages existing infrastructure. Many multimodal alignment methods report using DPO as the final fine-tuning stage to inject human (or AI) preferences. For instance, the **CHiP** method (2023) – which stands for *Contrastive Human-aligned Instruction-tuning for Multimodal Preference* – used a hierarchical DPO loss: it gathered human preferences not just on the final answer, but also on segments and even tokens within the answer . CHiP then combined a **visual preference loss** (to handle cases where an image might be better explained by cropping or other processing) with multiple text-level losses. This multi-level DPO fine-tuning led to notable reductions in hallucinated details and alignment errors . Similarly, **HDPO** (Hallucination-specific DPO) constructed targeted comparison pairs to tackle different types of multimodal hallucinations (e.g. an image that tempts the model to describe something not present) and applied DPO on those, successfully reducing those failure modes .

Beyond DPO, there are related strategies such as **Direct Reward Fine-Tuning** and others where the distinction blurs between "RL" and supervised learning. Some works blend supervised loss with preference losses. For example, **MPO** (Multimodal Preference Optimization) combined standard instruction tuning loss with multiple preference-based losses to improve reasoning performance . *Dynamic Reward Scaling* (noted in the MM-RLHF paper) can also be seen in a

DPO context: the idea there was to weight each training sample's loss by how large the reward model's margin was, so that the most informative preference comparisons (the ones with a strong human preference) have greater effect . This yielded better sample efficiency – essentially focusing learning on the clearest preference signals.

Overall, these approaches demonstrate a **trend towards simpler, more direct alignment training**. By late 2024, many researchers preferred formulating alignment as a supervised learning problem (using comparisons or even AI-generated feedback as training data) rather than a complex RL control problem . This has enabled rapid experimentation: numerous papers introduced a new flavor of "X-DPO" (where X is some modifier like visual, adversarial, etc.) to tackle a specific alignment issue. The take-home message is that there is now a **toolkit of loss functions** beyond vanilla cross-entropy that one can use to fine-tune multimodal models for alignment goals – these achieve many of the same benefits as RLHF with potentially less overhead.

AI Feedback and Self-Alignment

Human feedback can be expensive and slow; thus a promising avenue is to leverage AI itself in the alignment loop. Two prominent ideas are **Reinforcement Learning from AI Feedback (RLAIF)** and **self-critiquing models**. In RLAIF, instead of (or in addition to) human judges, another AI system evaluates the candidate outputs. This could be a separate model or the same model in a different mode. The evaluating AI might be an expert for a specific domain or a prior version of the model that's been instructed to be critical. Anthropopic's *Constitutional AI* (introduced in 2022 for text) is a form of this: the model generates an answer, then it generates a critique of that answer based on a list of principles (the "constitution"), and finally revises the answer accordingly. In the multimodal setting, such approaches only started gaining traction post-2023.

One example is a **self-critique mechanism** described as *RLAIF-V* in a 2024 work . The model would produce multiple responses to a visual prompt, break each response into sentences, then ask an open-source LLM to score each sentence's trustworthiness, summing those scores to decide which response was best . That scoring model essentially provides AI feedback, which is then used to construct a DPO training signal. Another method called **LLaVA-Critic** trained a dedicated vision-language critic model: they used an existing aligned model (LLaVA-OneVision) to generate answers, had an expert model (GPT-4V) rate those answers, and trained the critic on that data . The critic then could be used to score new outputs and fine-tune the original model via iterative DPO, yielding improvements in multiple aspects (fewer hallucinations, better understanding) .

OpenAI implicitly used AI feedback in GPT-4's training: as mentioned, a GPT-4 zero-shot classifier judged safety of outputs to provide an extra reward signal . This is essentially the model learning from its own judgment (since the classifier was GPT-4 itself). The **self-moderation** approach by Chen et al. (2023) goes further – at inference time, the model detects if its answer might violate privacy or other safety issues, and if so, it modifies its answer and asks itself "Are you sure this is safe?" up to a few iterations . This led to notably safer behavior without human intervention at runtime.

Another creative direction is training models to **generate explanations or critiques** as part of the output, which can then be used to improve alignment. The *Critique-Based Reward* model from MM-RLHF (2024) had the AI produce a short critique of a candidate answer (pointing out flaws or strengths) before scoring it . This made the reward model's decisions more interpretable and rich – instead of a blind scalar, you get a rationale. In principle, such critiques could be shown to the user or used to further fine-tune the base model to *internalize* these rationales, leading to a self-improving cycle.

Finally, **Constitutional AI** as described by Anthropic has not been fully realized in multimodal models publicly by 2025, but the concept is influential. It suggests that we can specify a set of rules (like "the model should not discriminate or encourage illegal acts") and then use the model to generate self-feedback when it violates those rules, tuning it to comply. This approach, if extended, could allow aligning models without large human-labeled datasets, which is attractive for modalities where labeling every scenario is hard. Early multimodal safety research has indeed referenced using **LLM-based feedback to refine outputs** (for example, using GPT-4 to adjust an image description to be more accurate or inoffensive before finalizing it). We can expect more of this *self-aligning behavior* in future multimodal systems, making them more **autonomously reflective** about whether an answer is safe and correct, rather than relying purely on hard-coded rules or static training.

Safety-Specific Training and Adversarial Alignment

Ensuring safety in multimodal models requires going beyond general "be helpful" training. After 2023, many projects addressed safety **head-on** by creating dedicated training datasets and algorithms to target unsafe or adversarial cases. A straightforward yet powerful technique is **to fine-tune on a curated set of bad scenarios with good behaviors**. For example, a group of researchers introduced *VLGuard* (2024) which gathered a large collection of multimodal content that is hateful, sexual, or otherwise disallowed, and fine-tuned a model on this data with

responses that either refuse or safely answer . By doing so, the model learned to recognize and avoid generating harmful content for those inputs (essentially *immunizing* it to known unsafe examples).

OpenAI took a similar but more granular approach with GPT-4V: They noticed certain "emerging risks" where an image and text prompt together could elicit a dangerous response even though separately they might seem fine . One risk was **illicit instructions** embedded in images – e.g. a photo of a kitchen knife with the text "best way to stab?" underneath. Another was **ungrounded inference** – e.g. asking the model to judge a person's personality from a photo (implicating bias/privacy issues). To address these, OpenAI *generated multimodal training examples by combining images with parts of known unsafe text prompts* . They would take a disallowed text query (like "how do I make a bomb") and insert images for key words ("how do I ![bomb image]?") to create a multimodal variant . The model was then fine-tuned to firmly refuse these. They report that after this targeted safety tuning, GPT-4V achieved a 97.2% refusal rate on illicit requests and 100% refusal on the ungrounded inference prompts in internal tests – a huge improvement over the initial behavior. This demonstrates that *directly injecting safety training data covering tricky multimodal cases is highly effective*. There was minimal impact on normal capabilities, indicating no fundamental trade-off between being safe and being useful .

Another frontier is **adversarial robustness**. Researchers found that multimodal models are vulnerable to cleverly perturbed images or audio that humans might not even notice. For example, adding a tiny overlay of text in an image could "trick" a model into ignoring safety instructions . Works like *Adversarial DPO* and *AdaShield* (2024) explicitly tackle this by training models on perturbed inputs: AdaShield uses a CLIP-based approach to automatically search for an adversarial prompt (text+image) that causes the worst-case behavior, then uses those as training data to fortify the model . The result was improved detection of risky inputs. Another study called *DREAM* (2025) introduced a technique to **disentangle risks in multimodal content** by step-by-step reasoning: essentially, the model or a helper model analyzes an image-text pair and labels which parts are unsafe, which are benign . This allows targeted mitigation. Using this analysis, DREAM then fine-tuned the model with a mixture of supervised and AI-feedback reinforcement learning, achieving a **16% higher safe-response rate** compared to GPT-4V on challenging inputs . An important aspect was avoiding *oversafety* – they ensured the model doesn't refuse far too much. Indeed, DREAM claims to boost safety "without compromising performance on normal tasks" , which they validate by showing the model still answers innocuous prompts correctly.

By 2025, safety-focused alignment has become more *proactive*. Rather than waiting for users to find a new exploit, developers engage in **red-teaming** (hiring experts or using adversarial generation tools to find model weaknesses) and then incorporate those findings into training. OpenAI's system cards detail extensive red team exercises for vision and voice models, with

mitigations implemented prior to release . For instance, testers tried to get GPT-4V to identify people in images or reveal private info; in response, OpenAI hardened the model's refusal in those areas . Google's handling of Gemini's bias issue (pausing image-of-people generation) is another example of swiftly adjusting deployment when a safety issue is discovered . Furthermore, companies increasingly rely on *system-level defenses* like monitoring for certain patterns. In the audio domain, the new custom voice models can be misused for impersonation, so OpenAI built **watermarking** and monitoring to trace generated audio . These don't change the model's weights but are vital to overall safe usage.

In summary, the alignment community has embraced a **multi-pronged safety strategy**: adversarial training data, specialized algorithms (like constrained RLHF for safety), ongoing red-team feedback, and external safeguards all work in tandem. This is what it takes to align powerful multimodal models with the nuanced norms of human society – covering not just blatant harms but also subtle issues of privacy, fairness, and misuse.

Bias and Fairness Mitigation Strategies

Mitigating bias in multimodal models overlaps with safety but deserves special note. Bias can be harder to measure – it's not a single forbidden output, but a skew in model behavior. Post-2023, researchers have started systematically evaluating **bias in image-and-text models**. An ACL 2024 paper by Luo *et al.* found significant gender and racial bias in many vision-language models when captioning images of people or associating professions . Industry too has faced embarrassment: e.g. image generators creating people might systematically produce lighter skin tones for prompts or mix ethnic features incorrectly, as seen with Google Gemini's mishap .

Data diversification is a primary mitigation: ensure the training set has diverse representation so the model doesn't learn one dominant association. The survey *Fairness and Bias in Multimodal AI* (2024) highlights efforts like the **FaceAware** and **FairFace** datasets which balance race and gender in face images . Models fine-tuned on these more balanced datasets exhibit less biased predictions when identifying demographic attributes. Another approach is **bias-focused loss or constraints** during alignment. For example, one could add a term in the reward model to penalize outputs that use gendered terms when not needed, or that switch dialect when the speaker's identity changes. Some experimental multimodal RLHF setups have likely included such signals implicitly by instructing human labelers to prefer "more fair" outputs (though this is hard to do consistently).

OpenAI's alignment of GPT-4V specifically **disallows certain inferences about people from images** to prevent bias and privacy issues: it won't say someone's race or guess their occupation from appearance . This blanket refusal avoids a whole class of biased or stereotype-laden outputs (even at the cost of sometimes being under-informative). They note plans to refine this – currently GPT-4V errs on the side of caution ("broad but imperfect refusals") and they aim to handle sensitive attributes in a more precise way going forward . We also see mention of tackling **representational harms** – for instance, if a vision model always describes women's appearances but men by their roles, that's a representational bias. Future alignment may involve fine-tuning models to produce more equivalent descriptions for different groups, or using **counterfactual augmentation** (generate pairs of images differing only in one attribute and ensure the model's descriptions don't shift inappropriately). Early studies like Zhao et al. (2024) demonstrated that such counterfactual data augmentation can reduce gender bias in vision-language models .

Lastly, evaluation is key: new benchmarks are being developed to quantify multimodal bias. These might show the model a series of images of individuals across demographics and ask it to generate descriptions or answers, then analyze for differences. As alignment is an ongoing process, **transparency reports and system cards now routinely include bias analysis**. OpenAI's system card discussed questions like "Should the model identify public figures in images? Should it infer gender or not?" and references how these are intertwined with fairness and societal values . By making these dilemmas explicit, researchers can better design alignment criteria. In summary, bias mitigation in multimodal models uses **data, policy (rules), and objective tuning**. It strives to ensure the model's performance is consistent and fair across subgroups – an essential aspect of reliability.

Industry Practices and Case Studies

Leading AI labs have converged on broadly similar *alignment pipelines* for multimodal models, often combining many of the techniques above. A typical recipe in 2024–2025 looks like: **Pretrain on vast data → Supervised fine-tune on curated multimodal instructions → RLHF (or similar) with human and/or AI feedback → Red-team testing and safety fine-tuning → Deploy with monitoring and guardrails.** Within this outline, each organization has developed its own tools and lessons:

- **OpenAI (GPT-4 and successors):** OpenAI's GPT-4 (text and vision) was at the forefront of multimodal alignment efforts. According to their technical report and system cards, GPT-4's vision model was trained with a two-tier RLHF: one reward model to ensure helpfulness and another safety classifier to penalize unsafe outputs . They also performed *iterative deployment* – rolling out GPT-4V to a small group of beta testers (e.g. visually impaired users via Be My Eyes) and gathering feedback on its hallucinations and errors . This surfaced issues like the model sometimes being

overconfident about image details that were wrong ("hallucinating matter-of-fact confidence" as one tester noted) – but over the beta period they adjusted and saw the error rates drop . OpenAI heavily used **expert red teaming**: they hired outside experts to attack GPT-4V, which led to discovering exploits (like the OCR loophole for disallowed text in images) that they then patched with system-level fixes . OpenAI's public system cards explicitly list mitigations: e.g. *model-level*: additional training on combined image+text harm data; *system-level*: blocking face recognition, adding OCR filters, etc. . For DALL·E 3 (image generation), OpenAI similarly used alignment techniques: they mention using "**preference modeling**" to tune DALL·E 3 to follow user intent better and a host of safety measures to stop it from producing disallowed images . This included prompt filtering (the model refuses certain prompts) and output filtering (detecting and removing unsafe images). The trend with OpenAI is a **belt-and-suspenders approach** – multiple layers of alignment and safety, thoroughly documented in system cards . They have also open-sourced evaluation frameworks (like their "Preparedness" tests for extreme risks) which, while aimed at frontier AI, also improve everyday alignment.

- **Google DeepMind (Gemini, Bard, etc.):** Google's recent models (e.g. PaLM 2 and Gemini) are multimodal and Google has emphasized **responsible AI principles**. In practice, Google employs extensive human feedback tuning as well. For instance, Bard's image understanding capabilities (introduced 2023) were evaluated and improved using an internal human rating system ensuring its responses are accurate and not inappropriate . A case study is **Gemini's image generation bias**: as reported by *The Guardian*, Gemini's image module would generate historical figures in randomized ethnicities/genders to be inclusive, but sometimes in a nonsensical way (e.g. a Viking with East Asian features, or a World War II German soldier portrayed inaccurately) . Users on social media pointed this out as a bias or calibration issue. Google's response was swift – they paused the model's ability to generate images of people and promised to "re-release an improved version soon" after fixing the issue . This highlights how industry teams monitor public feedback and are willing to rollback features that aren't aligned. Google has also integrated **watermarking** for AI-generated images from their models to help identify AI outputs , aligning with their commitment to transparency (so that users can tell when an image is AI-made, addressing misinformation concerns). Another example is Google's **Universal Translator** (an experimental speech-to-speech translator that clones voice) – they reportedly built in constraints to avoid misuse such as not allowing certain content and indicating the output is synthesized (likely via audio watermark or verbal markers). Though details are sparse, Google's AI Principles (2018) explicitly forbid deploying models that likely cause harm or bias, so their alignment process involves rigorous review stages. In summary, Google's practice leans on internal policy compliance: if a model output violates a principle (like fairness), engineering teams intervene to adjust either the model or the allowed usage.
- **Anthropic (Claude):** Anthropic's models like Claude 2 are text-only publicly, but their research heavily influences alignment approaches. They championed **Constitutional AI**, which they found can achieve a high level of harmlessness with far fewer human labels by letting the model self-correct using a set of rules . While we haven't seen a public multimodal Claude, one can surmise that Anthropic would apply similar ideas – e.g. a vision-augmented Claude would check outputs against a constitution (no identifying people in images, no hate content, etc.) and revise accordingly. Anthropic also puts a big

emphasis on **evaluations for honesty and harmlessness** (like TruthfulQA and bias benchmarks). Their 2024 research on "Collective Constitutional AI" even explores getting diverse public input on what the AI's principles should be , which could be very relevant for models that see the real world through images and might need culturally diverse norms. In industry practice, Anthropic distinguishes itself by *transparency about failures*: they discuss where Claude can still be tricked or where it refuses too much. This kind of openness is gradually spreading – even OpenAI and others now acknowledge "our model can still be jailbroken" and solicit feedback . So Anthropic's influence is in pushing alignment beyond just technical fixes towards involving broad values and being clear about limitations.

- **Meta (LLaMA and others):** Meta's LLaMA series (through LLaMA-2 in 2023) were text-only but with a strong emphasis on open and reproducible alignment. LLaMA-2-Chat, for example, was released with a fine-tuning recipe using supervised instruction data and over a million human annotations for RLHF (on helpfulness and safety) . They also released a **responsible use guide** and a built-in safety classifier. For multimodal models, Meta's divisions have created things like **AudioGen** (audio generation model) and **ImageBind** (a multimodal embedding model), though not a single large multimodal chat model as of 2024. However, Meta has invested in **content filtering technologies** – e.g. their open-source Segment Anything model can detect objects, which could be used to screen image inputs for disallowed objects (weapons, etc.) before feeding to a model. And for generative image models, they had introduced a**"Make-A-Scene"** and later **Emu** (2023) with likely similar safety filters as DALL·E. Being largely open-source, Meta often relies on the community to help red-team and find issues. They released **LLaMA-2** with known caveats (it may produce toxic text if prompted; they provided a separate safety model to mitigate this). For vision-language, open models like **LLaVA, BLIP-2, MiniGPT-4** (some developed with Meta collaborations) followed the same pattern: release the model and a recommended alignment method (usually fine-tuning on some instruction data), and encourage users to fine-tune further for their needs. This open approach accelerates research – indeed many of the academic alignment papers we cited use LLaMA or BLIP as base models to test new alignment algorithms . The industry trend here is **"open model, open alignment"** – sharing not just weights but the data and codes for alignment steps, which is crucial for safety in the long run because it allows independent scrutiny.

In all, industry practices in 2024-2025 show a maturation of alignment: from one-off fixes to integrated pipelines with **continuous improvement loops**. Companies treat alignment as an ongoing process, not a one-time training event. They publish **system cards and model cards** detailing how the model was aligned and what tests it passed . External audits and partnerships (OpenAI, for example, partnered with the *Partnership on AI* for red-teaming and consulted with fairness experts) are becoming common. This industry collaboration on safety is a notable trend post-2023, spurred in part by regulatory and public pressure. The result is that new multimodal models are **significantly safer and more aligned** than their predecessors, though not perfect. When issues arise, the response is faster and more transparent than before.

Conclusion and Future Trends

The fast-paced developments post-2023 have greatly advanced the safe alignment of multimodal LLMs. Models like GPT-4V, PaLM-2, and others represent significant leaps in capability, accompanied by sophisticated alignment to ensure those capabilities are used responsibly. We now have a diverse toolkit: from RLHF and DPO for instilling human preferences, to adversarial training and safety-specific reward models for guarding against misuse, to deployment-time tools like monitoring, OCR filtering, and watermarking. Empirical results show these techniques **work best in combination** – e.g. a model fine-tuned on curated data, then optimized with human feedback, and finally polished with targeted safety data can substantially close the alignment gaps that existed initially .

Importantly, many **post-2023 innovations** emphasize that alignment need not come at the expense of capability. By crafting training objectives carefully, researchers achieved models that are *both* more aligned *and* more competent . This is a hopeful trend, as earlier it was feared that making a model safer might dull its usefulness. Techniques like multi-objective RLHF, critique-based rewards, and iterative self-refinement contribute to this positive synergy between helpfulness and harmlessness.

Looking forward, several trends and open challenges stand out:

- **Scaling Human Feedback with AI Assistants:** As models become more complex (e.g. incorporating video, multiple images, and text in a single session), collecting human feedback at scale is challenging. Future alignment may rely on AI-assisted labeling – powerful models helping to evaluate or provide feedback on weaker models. Ensuring the AI feedback is itself unbiased and accurate will be an area of focus (to avoid *feedback loops* of error). The use of **GPT-4 to supervise GPT-4V** was an early example ; we expect more "model judges model" frameworks, potentially involving multiple models debating or checking each other.
- **Holistic Multimodal Understanding for Factuality:** To further reduce hallucinations, multimodal models might incorporate retrieval systems (e.g. searching the web or a database when unsure about an image) so that their answers are grounded in verified information. Aligning the use of such tools (when to trust an image vs. when to retrieve external facts) will be a new dimension of alignment. Some recent work already augments reward models with factual sources – a trend that blurs the line between pure learning and hybrid systems.
- **Continuous Learning and Updates:** As the world changes and as norms evolve, aligned behavior today might be insufficient tomorrow. We anticipate frameworks for *online*

alignment: models that can take in new alignment data or adjust their safety parameters on the fly. This could involve user-specific alignment (model adapts to a community's norms) or global updates (coordinated improvements pushed to all users when a new threat is discovered). Designing such systems safely (so they don't forget past alignment when learning new ones) is an open problem.

- **Evaluation and Benchmarks:** There is a need for standardized multimodal alignment benchmarks. Text has benchmarks like TruthfulQA, HHH (Harmlessness, Helpfulness, Honesty) evaluations, bias tests, etc. Multimodal analogues are emerging – e.g. SIUO (Safe & Interactive Understanding Benchmark) used in DREAM , or MMHAL-Bench for hallucinations . We expect more community efforts to evaluate models on safety across images and audio. Better evaluation will drive better alignment techniques by identifying specific weaknesses. Also, **calibrating "oversafety"** is important – ensuring a model doesn't become so cautious that it refuses legitimate requests (a user frustration). Fine-grained metrics that reward refusal of truly unsafe requests but penalize unnecessary refusals will guide future training.
- **Alignment in Specialized Domains:** When multimodal models are applied in high-stakes fields like medicine, law, or autonomous driving (robotics vision), alignment techniques must incorporate domain-specific constraints. For example, a medical image analysis model must follow different privacy rules (e.g. HIPAA) and factual standards than a general model. We may see **domain-specific constitutions or reward models** that ensure alignment not just with generic human preferences, but with professional ethical codes or regulations. This could involve collaborations between AI researchers and domain experts to encode those alignments. The survey of future directions by Fu *et al.* (2023) indeed suggests tailoring alignment frameworks to specific domains as a key opportunity .

In closing, safely aligning multimodal LLMs is a multi-faceted challenge, but the progress since 2023 is encouraging. Through innovations in training algorithms and a commitment to responsible deployment by industry leaders, we have seen a substantial reduction in toxic or incorrect outputs from these models . Yet, alignment is an ongoing journey – it requires vigilance as new capabilities (and new failure modes) arise. The convergence of ideas from academia (e.g. novel RLHF tricks, bias mitigation research) and industry (large-scale implementation and red-teaming) will continue to push the frontier. By keeping models **aligned with human values across all modalities**, we move closer to AI systems that are not only powerful, but also **trustworthy** and beneficial in our daily lives.

Sources: The information in this report is drawn from recent literature and official reports, including alignment surveys , technical system cards , research papers on RLHF and safety for multimodal models , and news on industry practices . These references reflect the cutting-edge developments up to 2025 in aligning multimodal AI with human preferences and ethical norms.