ChatGPT

Aligning Visual Models with LLM-Style Alignment Paradigms (2023–2025)

Introduction

Recent research is exploring how large language model (LLM) alignment techniques – such as **instruction tuning** and **reinforcement learning from human feedback (RLHF)** – can be applied to **vision models**. The goal is to train visual models (for tasks like detection, segmentation, recognition, etc.) to follow high-level instructions and align their behavior with human intentions, similar to how LLMs are aligned. These approaches often introduce **new post-training stages** beyond standard supervised learning. Instead of purely minimizing pixel-level or label losses, the model is tuned with human feedback, preference rewards, or instruction-following data. This promises stronger task alignment and robustness, potentially benefiting specialized applications (e.g. infant pose estimation in healthcare, where robust generalization from limited data is needed). Below, we survey representative works from 2023–2025 – spanning CVPR, ICCV, ECCV, NeurIPS, ICLR, etc. – that exemplify this emerging paradigm, and discuss their methods, results, and implications.

RLHF for Vision: Reward-Optimized Visual Models

In the language domain, RLHF has proven effective for aligning models to human preferences. Recent works show a similar trend in vision-language and vision-only models: using reinforcement learning with reward feedback to fine-tune model behavior. A notable example is **RLHF-V (CVPR 2024)** by Yu et al., which applies RLHF to **multimodal large language models (MLLMs)** to reduce visual hallucinations [1] [2]. Instead of instruction-tuning alone, RLHF-V collects fine-grained human feedback on model outputs (segment-level corrections of hallucinated descriptions) and directly optimizes the model via preference-driven policy updates. The method significantly improved factuality: with only ~1.4K human-annotated examples, RLHF-V cut the model’ s image-grounded hallucination rate by **34.8%** [3] [4], outperforming a baseline that used 10K feedback points. Figure 1 illustrates the RLHF-V process – the model generates responses about an image, human annotators mark erroneous segments, and the model is optimized via **direct preference optimization** to prefer corrected responses [1] [2]. This human-in-the-loop RL training (done in one hour on 8 GPUs) produced a more trustworthy vision-language model without sacrificing helpfulness [5] [6].

Figure 1: RLHF-V framework [7] [8]. A vision-language model’ s response is refined via human feedback: annotators provide fine-grained corrections to hallucinated segments, and the model is updated (using a reward model or direct optimization) so that its outputs align with these human preferences. This reinforcement fine-tuning stage significantly reduces image-grounded hallucinations [3] [4], improving the model’ s trustworthiness on multimodal tasks.

Beyond reducing hallucinations, reinforcement fine-tuning has been used to boost visual task performance and generalization with limited data. **Visual-RFT (ICCV 2025)** by Liu et al. introduced Visual Reinforcement Fine-Tuning, adapting OpenAI’ s RFT strategy to **vision-language tasks** [9] [10]. Their system first uses a pre-trained LVLM (Large Vision-Language Model) to generate multiple candidate outputs (with reasoning steps) for each input, then defines **verifiable reward functions** (based on measurable task criteria) to score these outputs [11] [10]. For example, Visual-RFT uses an **IoU-based**

reward for object detection and a classification accuracy reward for fine-grained recognition [11]. The model (built on Qwen-VL 7B) is then optimized with a policy algorithm (Group-PPO, following the "DeepSeek-R1" strategy) to maximize these rewards [11] [10]. This procedure led to **better accuracy and generalization than standard supervised fine-tuning**. For instance, on one-shot fine-grained image classification, Visual-RFT achieved substantial accuracy gains over a supervised baseline, despite using only ~100 training samples [12]. In few-shot object detection, Visual-RFT similarly outperformed supervised fine-tuning on COCO and LVIS benchmarks [12]. These results suggest that RL-based fine-tuning can **train vision models more data-efficiently**, by learning through trial-and-error reward feedback instead of purely imitating labeled answers [13] [14]. The authors note this approach represents a "paradigm shift" in fine-tuning LVLMs, yielding models that reason and adapt better in domain-specific tasks [15] [16].

Another line of work, often termed **"R1"-style reinforcement**, avoids needing an explicit learned reward model by using rule-based rewards. **Vision-R1 (CAS IA, 2025)** by Zhan et al. exemplifies this "human-free" alignment: it fine-tunes a 7B vision-language model with **vision-guided rewards** derived from task logic [17] [18]. They use only curated instruction-following data (no separate preference dataset) and design a criterion-driven reward function that evaluates each generated answer on multiple facets (e.g. bounding box precision, answer correctness) using the ground-truth visual cues [19] [20]. A progressive rule refinement strategy is employed: the reward criteria are tightened as training progresses, like a curriculum, to prevent the model from gaming the reward ("reward hacking") and to continually push performance [21] [22]. Vision-R1 achieved notable gains on object localization tasks, improving both in-distribution and out-of-distribution performance – in some cases up to **50% relative improvement**, even surpassing the accuracy of a 10× larger model that was not RL-tuned [23] [24]. This underscores that properly designed rewards can drive behavior alignment effectively, producing smaller models with performance rivaling much larger ones.

A particularly interesting application of RL in vision is in **segmentation** tasks. Traditional segmentation models rely on per-pixel supervision; however, **Seg-R1 (2025)** by You and Wu explores training a segmentation-capable model purely via RL [25] [26]. Their setup connects a large vision-language model (Alibaba's Qwen-2.5-VL) with the pretrained Segment-Anything model (SAM2) in a loop [27] [28]. The vision-language model is trained (using a new Group RPO algorithm) to output prompts (points or bounding boxes) that guide SAM in producing the desired segmentation mask [27] [29]. The reward combines mask quality metrics (Intersection-over-Union and S-Measure) with format correctness, encouraging both high accuracy and correct reasoning format [30] [29]. Remarkably, Seg-R1 achieved strong results on challenging tasks like **camouflaged object detection** – e.g. an S-measure of 0.873 on COD10K, with no direct pixel-wise training loss [31] [32]. It also demonstrated impressive **open-world generalization**: after RL training solely on generic foreground segmentation, the model could zero-shot transfer to referring segmentation (segmenting by a given phrase) on RefCOCOg with 71.4% cIoU, outperforming some fully supervised models on that benchmark [33] [34]. The authors found that pure RL tuning gave the model a real **pixel-level understanding**, without needing expensive mask annotations or specialized decoder layers [35] [36]. This suggests RL optimization can sometimes replace or supplement supervised fine-tuning, yielding a more **generalizable segmentation model** that handles diverse inputs (points, boxes, text) and novel objects.

Finally, researchers have begun **transferring these methods to pose estimation and keypoint detection** tasks. One cutting-edge example is **Pose-RFT (ArXiv 2025)** by Li et al. [37] [38]. This work integrates reinforcement fine-tuning into a multimodal model that generates 3D human poses (SMPL parameters) from either images or text descriptions. Pose-RFT formulates pose generation as a **hybrid action space** problem: the policy (a vision-language model extended with a "pose head") outputs both discrete text tokens and continuous pose vector parameters [37] [39]. They develop a hybrid version of

Group RPO (called HyGRPO) that can optimize over mixed discrete–continuous outputs [40] [41]. Multiple task-specific reward functions guide the learning: e.g. a spatial reward (measuring keypoint location accuracy for image-to-pose), a semantic alignment reward (for text-to-pose, ensuring the pose matches the described action), a format correctness reward, and even an embedding-based similarity reward to encourage plausible pose structure [38] [42]. Training the model with these reinforced signals led to **significantly improved 3D pose generation** performance over the base pose model [43]. Notably, the RL-tuned model can produce poses that are both more spatially accurate and better aligned with textual intent, compared to a purely supervised approach that minimizes mean-squared joint error [44] [38]. Pose-RFT is the first to demonstrate that RLHF-style optimization is feasible in pose estimation, which involves continuous outputs; it had to address challenges like stabilizing continuous action updates and balancing multiple reward objectives [37] [41]. This advance hints that even specialized vision tasks like human pose estimation – including niche cases like **infant pose estimation** – could benefit from alignment techniques. A reward function could, for example, encode medical practitioner preferences or physical plausibility (penalizing anatomically impossible infant poses), and an RL-trained pose model might generalize better to infants’ unique poses than a model trained only on adult pose data.

Summary: RLHF and related reinforcement fine-tuning approaches represent a new post-training paradigm for vision models. Instead of one-step fine-tuning on labeled data, the model’s outputs are iteratively refined by optimizing a reward signal – whether derived from human preference rankings, rule-based metrics, or simulated feedback. Empirical results across tasks (detection, segmentation, VQA, pose, etc.) consistently show **improved alignment and robustness**: models make fewer obvious mistakes (e.g. hallucinations [3]), require far fewer supervised examples to achieve strong results [12], and handle novel scenarios or instructions more gracefully (often matching or beating fully supervised models on zero-shot tests [33] [34]). These benefits are particularly relevant for domains like medical or infant pose estimation, where obtaining large annotated datasets is hard and model errors carry high costs. An RLHF-tuned infant pose model, for instance, could be aligned to prioritize medically relevant accuracy (through a reward) and might transfer knowledge from adult pose data more safely by following high-level feedback (like an expert’s preferences) rather than overfitting to scarce infant data.

Instruction Fine-Tuning of Vision Models (Flamingo, Kosmos, SEEM, etc.)

In parallel to RLHF, another major trend is **instruction fine-tuning for vision models** – teaching models to follow natural-language **commands or queries** about images. Inspired by the success of instruction-tuned LLMs (e.g. GPT-4, which was tuned to follow human instructions), researchers have begun to create **multimodal models** that can accept textual instructions + images and produce useful outputs (e.g. answers, descriptions, or even images). Several high-profile systems from 2022–2024 have established this paradigm:

- **DeepMind’s Flamingo (NeurIPS 2022)** was an early breakthrough in vision-language instruction following. Flamingo is a Visual Language Model that combines a frozen pretrained language model (e.g. Chinchilla 70B) with a visual encoder via special cross-attention layers [45]. It was trained on large-scale multimodal sequences of interleaved images and text (web data) to handle arbitrary image-text inputs [46]. The key outcome is that Flamingo can perform **few-shot learning on novel vision tasks** simply by being prompted with a task description and a few examples, without additional fine-tuning [47] [48]. For example, given an image and the instruction “What is this a picture of?”, Flamingo will output a caption; if prompted with a dialogue history, it will engage in visual QA, etc. Flamingo achieved **state-of-the-art few-shot results** on a wide spectrum of tasks – from open-ended VQA and image captioning to multiple-choice visual reasoning – often outperforming models that were fine-tuned on thousands of task-specific examples [48]. This demonstrated that a single large model can be taught to follow vision-language instructions and rapidly adapt to new tasks via prompting [47]. Flamingo’s design introduced architectural innovations to enable this flexibility, such as novel masking schemes for cross-attention that let the model ingest images or videos at arbitrary points in the sequence [45]. In essence, Flamingo behaves like an LLM that “speaks about images” – an early exemplar of aligning perception with the instruction-following abilities of language models.

- **Kosmos-1 (Microsoft, 2023)** pushed the idea further by training a multimodal large language model from scratch with cross-modal instructions. Described in “Language Is Not All You Need: Aligning Perception with Language Models”, Kosmos-1 is a **1.6B-parameter** model that was fed a web-scale corpus mixing text and images (including image–caption pairs and OCR tasks) [49]. Unlike Flamingo (which used a frozen LM), Kosmos-1’s entire network was learned to jointly perceive modalities and follow instructions. The resulting model can handle a remarkably broad set of tasks zero-shot, without gradient updates [50]. It not only performs standard language tasks but also vision-and-language tasks like VQA, image captioning, and even pure vision tasks specified by text (e.g. classification by description) [51]. For example, Kosmos-1 can be given an image and the prompt “Describe what is unusual in this image” and produce a sensible answer, or it can do OCR-free reading of an image with text [51]. On evaluation, Kosmos-1 showed **impressive zero-shot and few-shot performance** across NLP benchmarks, multimodal reasoning tasks, and image understanding tasks [51] [52]. This work validated that a single transformer can be aligned to both vision and language modalities simultaneously, learning in-context reasoning that transfers from language to multimodal domains [52]. In summary, Kosmos-1 treats vision as an extension of language – aligning perceptual inputs with the instruction-following paradigm, enabling multimodal chain-of-thought and cross-modal knowledge transfer [53] [52].

- **LLaVA: Large Language and Vision Assistant (NeurIPS 2023)** is another representative effort in visual instruction tuning [54]. Instead of manually curating multimodal instructions, LLaVA cleverly used **GPT-4** to generate them. Liu et al. had GPT-4 (with vision disabled) imagine being a “vision assistant” and produce dialogues about images – yielding a synthetic training set of image–instruction–response triples [54] [55]. Using around 158K such GPT-4-generated examples, they fine-tuned a CLIP ViT-L image encoder fused to a language model (Vicuna) to create LLaVA [56]. The result is an open-source chat assistant that can discuss images in detail, follow user commands about an image, etc. Impressively, LLaVA’s GPT-4-inspired training gave it chat capabilities close to multimodal GPT-4’s level on a test set of complex vision-language tasks – achieving about **85% of GPT-4’s score** on a benchmark of generated instructions [55]. For example, LLaVA can answer questions like “What might this person be feeling, given the picture?” or “List the objects on the table,” demonstrating reasoning about the image content. LLaVA thus pioneered instruction-following fine-tuning using machine-generated multimodal data, providing a recipe to align vision models to instructions without requiring expensive human annotation [54] [55]. Many subsequent works (e.g. **mPLUG-Owl, Otter, InstructBLIP,** and others in 2023) adopted similar strategies – often starting from a strong vision-language foundation model (like BLIP-2 or CLIP-ViT plus LLM) and **fine-tuning it on various image-text instruction datasets** (some human-curated, some GPT-generated). The common outcome is a new breed of **general-purpose multimodal assistants** that can take images plus natural-language prompts and produce coherent, contextually aligned outputs.

- Beyond vision-language dialogue, instruction tuning has also appeared in pure vision tasks through promptable models. **SEEM: Segment Everything Everywhere All at Once (NeurIPS 2023)** is a prime example of bringing an instruction-like interface to segmentation [57] [58]. Zou et al. designed SEEM as a **universal segmentation model** that accepts diverse prompts – text queries, clicked points, boxes, or even another reference image – and outputs the corresponding segmentation masks [57] [59]. The aim was to build a “universal interface” for segmentation that behaves more like an LLM [60]. SEEM’s architecture (built on the powerful pre-trained X-Decoder) includes a joint visual-semantic embedding space so that textual labels and visual prompts (points, scribbles) can be handled together, plus a memory mechanism to allow iterative, interactive prompting [61] [58]. During training, SEEM was fed a mixture of segmentation tasks (interactive segmentation data, referring segmentation with text, video object segmentation, etc.), effectively instruction-tuning it to handle any segmentation query type [57] [58]. The resulting model can segment “everything everywhere” in an image: it can produce **multiple masks for all objects** with a single prompt, or focus on specific objects described by a text label (open-vocabulary segmentation) [62]. Notably, a single SEEM model achieved strong performance across 9 different segmentation benchmarks (interactive, referring, generic, and video segmentation) – often within reach of task-specific state-of-the-arts, despite using **100× less supervised data** by leveraging prompts [63]. More impressively, SEEM shows an ability to **generalize to new combinations of prompts** that it never saw during training (e.g. a text query + a point together) [64]. This compositional generalization is analogous to LLMs handling novel instructions: SEEM effectively aligned the segmentation task with an instruction-following paradigm, allowing it to adapt to arbitrary user inputs about an image. In practical terms, one can now “ask” SEEM to segment things in an image via language (e.g. “segment the baby and the crib”) or via clicks, and it will respond accordingly – a big step beyond traditional fixed-function segmentation models.

- “CommandR” and related developments: The question also mentions **CommandR**, which likely refers to vision models that can be commanded or controlled via text instructions. (While specific details of "CommandR" are scarce in published literature, the name suggests an approach to **“command” a model’s behavior** in a vision context.) In general, 2023 saw the emergence of methods that integrate textual commands to control vision models’ outputs. For example, researchers have explored **prompt-based editing or refinement** of an image output: one can give a segmentation model a command like “now refine the mask to include the left arm” or give an object detector an instruction “focus on smaller objects”. Early works in this vein include **InstructPix2Pix (CVPR 2023)** for image editing, where a generative model is fine-tuned on instruction-image pairs to perform described edits, and **Edit Everything** which accepts language instructions to modify image regions. In the context of recognition/detection, a “commandable” model might allow a user to specify the categories or the style of output (e.g. “only detect cars and pedestrians” or “explain why you think this object is a cat”). This trend of **language-directed vision control** is still emerging, but aligns with the overall goal: making vision models **interactive and aligned to user intent** rather than static classifiers. We can expect to see more formalized “Command-&-Response” vision models in the near future that combine an LLM’s flexibility with a vision model’s perception – essentially treating vision tasks as conditional instruction-following problems.

Figure 2: Instruction-followable vision model. Here we see a reinforcement-trained segmentation model (Seg-R1) handling diverse prompts: in the left example, a text instruction (“segment the doll”) yields a mask for the doll; on the right, a visual prompt (red point on an object) yields that object’s mask [65] [66]. Models like SEEM and Seg-R1 can take points, boxes, text, or regions as commands and segment accordingly. This showcases a new paradigm of vision models that behave like interactive assistants, aligning their output to the user’s query rather than a fixed label set.

Benefits of instruction-tuned vision models: They introduce a unified interface to many tasks and can generalize to tasks not explicitly seen in training. For instance, Flamingo and Kosmos-1 exhibit few-shot learning, meaning they can be prompted to perform a new vision task (e.g. count objects, describe an emotion from an image) with just a couple of examples, whereas prior models would need retraining for each new task [47] [48]. Moreover, instruction tuning often uses multi-task and multimodal data, which tends to produce more **robust and generalizable models**. SEEM’s ability to handle nine segmentation datasets with one model [67], and Flamingo’s state-of-the-art results across widely different benchmarks [48], both hint at the power of this approach. By aligning training objectives with high-level instructions, these models learn a form of abstract reasoning about images (e.g. understanding “what” to do from a prompt), rather than just mapping pixels to labels. This could be especially fruitful for domains like **“baby pose estimation”** mentioned by the question: Instead of training a separate pose estimator for infants, one could prompt a general vision model with an instruction like “Locate the baby’s body keypoints” or “Is the baby’s pose correct for its age?” and get a reasonable output. In fact, a large instruction-tuned model that has seen diverse human poses and read about them might transfer knowledge about typical poses to infant scenarios with minimal data (via prompting or light tuning). There is early evidence that multimodal models can learn from descriptions: e.g. a model could be taught what a correct infant sleeping posture is through annotated images plus text explanations, thereby aligning it to safety-oriented criteria in a way a normal pose model would not capture. While we are not fully there yet, the trend suggests that **task alignment via instructions** gives models a flexible grounding in human concepts (like body parts, activities) that can be harnessed for better adaptation and interpretability in specialized tasks.

Learning Visual Preferences: Reward Modeling and Feedback

Alongside RLHF and instruction tuning, a complementary idea is **reward modeling for vision** – i.e. learning a function that judges the quality of a visual model’s output, which can then be used as a training signal. In the LLM world, this is common (training a reward model on human preferences to guide RLHF). In vision, reward modeling is nascent but starting to appear, especially for open-ended generative tasks and safety. For instance, to align image generation models with user aesthetics or content preferences, researchers have trained reward models that score images by human-labeled criteria (e.g. realism, prompt alignment) [68] [69], and then used RL (PPO or similar) to adjust the generator. A recent survey on preference alignment in diffusion models notes that RLHF and **Direct Preference Optimization (DPO)** have been applied to text-to-image models to reduce problems like misalignment and toxicity [70] [71]. For example, OpenAI’s image model DALL-E was fine-tuned with a reward model to better follow user prompts while avoiding disallowed content. These efforts mirror in vision the pipelines used in text: first gather human preference data (e.g. users compare two image outputs or mark elements they like/dislike), train a reward model $R(image, prompt)$, then optimize the image model with RL to maximize R. The survey highlights that such **preference alignment** has been tried in domains like autonomous driving, medical imaging, and robotics as well [72] [71], where safety or user-specific criteria are crucial.

In the context of **vision-language models**, one interesting direction is using a language model as a reward model or feedback provider. Some works (e.g. **MM-Reinforcement Learning from AI Feedback, 2024**) have explored replacing human feedback with an LLM that critiques the vision model’s output (this is analogous to “AI feedback” used in pure NLP). For example, an LLM with vision (like GPT-4V) could inspect a model’s image caption and indicate if it’s accurate or not, forming a pseudo-reward. The listing by Zhang et al. (2025) on “Aligning Large Multimodal Models with Factually Augmented RLHF” suggests using critiques and factual feedback to improve multimodal alignment [73]. In their approach, a critique-based reward model is trained – essentially an LLM that was augmented with factual image captions so it can judge the factuality of a vision model’s answer [74]. By optimizing against this reward, the vision model learns to avoid factual errors (a similar goal to RLHF-V). The combination of language feedback and vision-specific rewards is powerful: language can flexibly express complex preferences (e.g. “the description should mention the background objects too”), which can complement hard metrics like IoU or accuracy. Indeed, Ji et al. (2023) proposed **Align Anything** – a framework to train all-modality models (image, audio, language) with **language feedback** as the unifying signal [75] . The idea is to have a human or AI provide a natural language critique of the model’ s output (for any modality), and train the model to improve based on that. An example given is instructing a model to refine an image segmentation by saying “the mask missed the person’ s left foot” – the model should then correct it. This approach moves beyond scalar rewards to using the richness of language as feedback.

While still in early stages, **learning from language/image feedback** opens the door to aligning vision models on criteria that are hard to encode in loss functions. For instance, for infant pose estimation, one might train a reward model that takes an image of a baby with a predicted pose and outputs a score reflecting how **clinically useful** or **safe** that pose annotation is (perhaps penalizing if key joints were missed or if the pose is physiologically implausible). Such a reward model could be informed by pediatric experts’ preferences or by an LLM that knows anatomy. By optimizing a pose model against this reward (through RL or iterative refinement), the model could learn to prioritize the aspects of pose estimation that matter for infants (e.g. correctly identifying limb positions even under blankets, ignoring irrelevant objects like toys). This goes beyond traditional MSE loss on keypoints – it imbues the model with a notion of “what good looks like” as defined by humans. Early evidence that reward tuning improves robustness includes Seg-R1’ s findings that RL-trained segmentation had far better open-set performance than supervised ones [32] [33] . Similarly, Visual-RFT and Vision-R1 both report improved **out-of-distribution generalization** after reinforcement alignment [23] [12] . This suggests that feedback-aligned models are less overfit to the idiosyncrasies of the training data and more tuned to the underlying task goals – a valuable property when adapting to new domains like infant images.

Implications for “Baby Pose Estimation” and Other Adapting Tasks

The question specifically asks if these new paradigms can inspire adaptation in domains like **infant pose estimation**, and whether they offer stronger task alignment and generalization than traditional fine-tuning. Based on the research surveyed:

- **New training processes:** Indeed, works like RLHF-V, Visual-RFT, etc., propose post-training alignment stages that are **novel to vision models**. Instead of the conventional pipeline (pre-train on ImageNet, fine-tune on target data with supervised loss), we now see **multistage pipelines**: e.g. (1) Supervised pre-training or base model, (2) Instruction fine-tuning on broad data (to make a generalist, instruction-following model), then (3) Reward-based fine-tuning (RLHF or RFT) to specialize the behavior. This could be directly applicable to an infant pose model: one could start with a general human-pose model like ViTPose or OpenPose, then instruction-tune it with a few examples of “how to interpret infant poses” (possibly pairing images with textual descriptions of the pose or differences from adult poses), and further RL-tune it with a reward that captures pediatric experts’ feedback (e.g. reward high recall of subtle limb movements). This multi-step alignment might yield a model more **in tune with the task’ s needs** than simply fine-tuning on a small infant pose dataset. Notably, each stage adds alignment: instruction tuning broadens the model’ s ability to take **task directives**, and RLHF refines its **behavior to match human preferences** (like prioritizing certain keypoints or being cautious with uncertain predictions).

- **Stronger task alignment:** The surveyed works indicate that alignment methods can make models do what we want, not just what the raw data says. For example, RLHF-V explicitly aligned an MLLM’ s behavior (reducing hallucinations to make it more factual) [1] [2] – an alignment objective not captured by standard loss. In infant pose estimation, “what we want” might be consistent and safe pose tracking even under occlusions or atypical poses. A reward model could encode that (perhaps learned from doctors’ ratings of pose estimates), thus RLHF would optimize the pose model to meet those criteria. This goes beyond minimizing average error; it could, for instance, emphasize never missing a limb (because missing a limb might mean a dangerous positioning undetected) even if that means occasionally false-detecting a limb – something a normal training regime wouldn’t consider. Thus, RLHF provides a mechanism to bake in such **preferences or safety constraints** explicitly via the reward.

- **Generalization and robustness:** A striking theme is that many RL or instruction-aligned models generalize better to new tasks or domains. Visual-RFT’s model handled new fine-grained categories with very little data [12]; Seg-R1 could zero-shot to referring segmentation [33]; Flamingo handled tasks it was never specifically trained on by just prompting [47]. This bodes well for adaptation-heavy scenarios. In infant pose estimation, models often struggle because infant images differ from adult ones (babies have different proportions, often lie on backs, etc.). An instruction-tuned model that has seen diverse contexts (textual and visual) might be less rigid – e.g. it might have the concept “baby” and “lying down” in its knowledge and could adjust its predictions accordingly. And an RLHF-tuned model might have learned a more **exploratory, error-tolerant strategy** (since RL encourages exploring output space to get a better reward) – potentially making it more adaptable when the distribution shifts. There is evidence from RL-based LLMs that they are more robust to distribution shift than purely supervised ones [76], and similar claims are made in vision RL works (e.g., “RL…suggests a promising direction for efficient model training” with fewer steps and broader capability [77]).

In summary, aligning visual models with LLM-style techniques has opened a path to **vision models that are more flexible, controllable, and aligned with end-task goals**. For cutting-edge niche applications like baby pose estimation, this means we could move from training a narrow model that outputs 2D keypoints, to training a “baby pose assistant” that understands context (e.g. “the baby is crawling vs. supine”), takes instructions (e.g. “alert me if the baby’s face is covered”), and optimizes its behavior for what human users care about (safety, comprehensiveness, low false-negatives). While more research is needed to fully realize this vision, the works from 2023–2025 provide strong proof-of-concept that **LLM alignment paradigms can substantially improve visual models** – yielding better alignment with human intent and stronger generalization than traditional fine-tuning alone [15] [32]. The convergence of vision and language training may ultimately produce unified models that can see and act according to high-level human guidance, which is an exciting development for the field.

Representative Works and Key Results:

- Flamingo (NeurIPS 2022) – 80B-parameter VLM, cross-attention architecture; few-shot vision-language learning by prompting; outperformed task-specific models on captioning, VQA with just prompt examples [47] [48].

- Kosmos-1 (Microsoft, 2023) – Multimodal LLM trained on image-text data; zero-shot follows instructions for OCR, VQA, classification via description; showed cross-modal transfer learning [51] [52].

- LLaVA (NeurIPS 2023) – GPT-4-generated visual instruction data to fine-tune a vision-chat model; achieved 85% of GPT-4’s performance on a multimodal benchmark [55]; open-source and spurred many follow-ups in multimodal chat.

- SEEM (NeurIPS 2023) – Universal segmentation with text/point prompts; single model on 9 datasets with 1/100th supervision, interactive segmentation memory [67]. New interface for vision tasks (promptable segmentation).

- Visual Instruction Tuning Survey (2023) – Liu et al. survey many emerging "instruction-following vision models", naming this field "Visual Instruction Tuning". They note that combining GPT-4 or other LLMs to generate training data has been a key enabler, and report that dozens of models (InstructBLIP, BLIP-2 OPT-2.7B, PaLM-E, etc.) are converging on the idea of a **general-purpose multimodal agent** that can be tuned with instruction/answer pairs [78] [55].

- RLHF-V (CVPR 2024) – Applied RLHF to vision-language (13B model); human feedback on hallucinations -> 34.8% hallucination reduction [3]; achieved SOTA open-source trustworthiness, even outperforming GPT-4V on some robustness tests [79] [80]. Introduced dense correction feedback for vision.

- Visual-RFT (ICCV 2025) – Rule-based RL fine-tuning on LVLMs (Qwen2-VL); improved one-shot classification and few-shot detection significantly over supervised fine-tune [12]; data-efficient (tens of samples) and improved reasoning and adaptability [14] [15]. No reward model needed (direct, verifiable rewards).

- Vision-R1 (ArXiv 2025) – Human-free preference optimization; criterion-driven multi-dimensional rewards (e.g. box precision) + progressive rule tightening; boosted a 7B model to match/outperform a 70B model on localization [23] [24]; consistent gains in and out of distribution.

- Seg-R1 (ArXiv 2025) – Pure RL for segmentation (no supervised masks); LMM generates prompts for SAM; used IoU-based reward; achieved SOTA on camouflaged object detection and excellent zero-shot transfer to referring segmentation [29] [33]. Demonstrated that RL-trained segmentation has strong open-world generalization [33].

- Pose-RFT (ArXiv 2025) – First RL fine-tuning for 3D pose in multimodal models; hybrid (text+continuous) action space; multiple rewards (spatial, semantic) [38] [42]; improved pose accuracy and alignment with text descriptions over SFT baselines [43]. Suggests RLHF can handle structured continuous outputs.

Each of these works contributes pieces toward a future where vision models can be tuned like dialogue agents – you can tell them what you want, they can learn from feedback on their output, and they aim to satisfy user-defined objectives rather than just maximizing a likelihood on fixed labels. This is a promising direction for making vision AI more useful, safe, and adaptable in real-world scenarios.

References: The information above is synthesized from recent publications and preprints, including arXiv papers and conference proceedings from CVPR, ICCV, NeurIPS, etc. Key sources include the RLHF-V paper [1] [2], the Visual-RFT paper [14] [12], the Seg-R1 paper [29] [33], the Flamingo paper [47] [48], the Kosmos-1 paper [51] [52], the SEEM paper [57] [58], and others as cited inline. These provide detailed descriptions of model architectures, training processes, and empirical results that back up the points discussed. Each represents a step toward aligning vision models with techniques originally developed for language model alignment.

[1] [2] [3] [4] [5] [6] [79] [80] RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback
https://openaccess.thecvf.com/content/CVPR2024/papers/Yu_RLHF-V_Towards_Trustworthy_MLLMs_via_Behavior_Alignment_from_Fine-grained_Correctional_CVPR_2024_paper.pdf

[7] [8] RLHF-V
https://rlhf-v.github.io/

[9] [10] [11] [12] [13] [14] [15] [16] Visual-RFT: Visual Reinforcement Fine-Tuning
https://arxiv.org/html/2503.01785v1

[17] [18] [19] [20] [21] [22] [23] [24] Vision-R1: Evolving Human-Free Alignment in Large Vision-Language Models via Vision-Guided Reinforcement Learning
https://arxiv.org/html/2503.18013v1

[25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [76] [77] Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning
https://arxiv.org/html/2506.22624v1

[37] [38] [39] [40] [41] [42] [43] [44] Pose-RFT: Enhancing MLLMs for 3D Pose Generation via Hybrid Action Reinforcement Fine-Tuning
https://arxiv.org/pdf/2508.07804

[45] [46] [47] [48] [2204.14198] Flamingo: a Visual Language Model for Few-Shot Learning
https://arxiv.org/abs/2204.14198

[49] [50] [51] [52] [53] [2302.14045] Language Is Not All You Need: Aligning Perception with Language Models
https://arxiv.org/abs/2302.14045

[54] [55] [56] [78] [2304.08485] Visual Instruction Tuning
https://arxiv.org/abs/2304.08485

[57] [58] [59] [60] [61] [62] [63] [64] [67] [2304.06718] Segment Everything Everywhere All at Once
https://arxiv.org/abs/2304.06718

[65] [66] Seg-R1: Segmentation Can Be Surprisingly Simple with Reinforcement Learning
https://geshang777.github.io/seg-r1.github.io/

[68] [69] [70] [71] [72] Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing
https://arxiv.org/html/2502.07829v1

[73] [74] [75] GitHub - opendilab/awesome-RLHF: A curated list of reinforcement learning with human feedback resources (continually updated)
https://github.com/opendilab/awesome-RLHF