# RoVRM: A Robust Visual Reward Model Optimized via Auxiliary Textual Preference Data

**Chenglong Wang[1*], Yang Gan[1*], Yifu Huo[1*], Yongyu Mu[1], Murun Yang[1], Qiaozhi He[1], Tong Xiao[1,2†], Chunliang Zhang[1,2], Tongran Liu[3] and Jingbo Zhu[1,2]**

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] NiuTrans Research, Shenyang, China
[3] CAS Key Laboratory of Behavioral Science, Institute of Psychology, CAS, Beijing, China
{clwang1119, zzhu8250}@gmail.com, {xiaotong, zhujingbo}@mail.neu.edu.cn

## Abstract

Large vision-language models (LVLMs) often fail to align with human preferences, leading to issues like generating misleading content without proper visual context (also known as *hallucination*). A promising solution to this problem is using human-preference alignment techniques, such as best-of-$n$ sampling and reinforcement learning. However, these techniques face the difficulty arising from the scarcity of visual preference data, which is required to train a visual reward model (VRM). In this work, we continue the line of research. We present a **Ro**bust **V**isual **R**eward **M**odel (RoVRM) which improves human-preference alignment for LVLMs. RoVRM leverages auxiliary textual preference data through a three-phase progressive training and optimal transport-based preference data selection to effectively mitigate the scarcity of visual preference data. We experiment with RoVRM on the commonly used vision-language tasks based on the LLaVA-1.5-7B and -13B models. Experimental results demonstrate that RoVRM consistently outperforms traditional VRMs. Furthermore, our three-phase progressive training and preference data selection approaches can yield consistent performance gains over ranking-based alignment techniques, such as direct preference optimization.

## Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various natural language processing tasks (Stiennon et al. 2020; Ouyang et al. 2022). Recent works tend to fine-tune LLMs using specialized visual instruction tuning datasets, leading to the emergence of powerful large vision-language models (LVLMs) (Liu et al. 2024a; Lin et al. 2024; Huang et al. 2024b). Despite these advancements, current LVLMs are not well-aligned with human preferences. A glaring problem is that LVLMs sometimes generate misleading content without anchoring to the given visual context (also known as *hallucination*) (Leng et al. 2024). For instance, as illustrated in Figure 1, an LVLM incorrectly identifies a "pitaya" in an image of mangosteens due to their visual similarity.

Two predominant research approaches aim to address this problem. The first approach focuses on generating richer and higher-quality visual instruction data (Li et al. 2023b; Liu et al. 2023, 2024c), *i.e.*, annotating rich instruction samples on images of mangosteens to enable LVLMs to identify them more accurately. In contrast, a more sophisticated approach is applying human-preference alignment techniques, including best-of-$n$ sampling and reinforcement learning (RL), which can efficiently align models with human preferences on various tasks by optimizing against a reward model without instruction samples. However, applying these alignment techniques to LVLMs is not a low-hanging fruit. It typically faces the difficulty of training a visual reward model (VRM) due to the scarcity of high-quality visual preference data (Sun et al. 2023; Yu et al. 2024a; Zhou et al. 2024b).

This work is motivated by a simple idea: human preferences are well-captured by text and these preferences can be transferred across different modalities. In this way, we can make use of rich, high-quality textual preference data in training VRMs. Building on this idea, we present a **Ro**bust **V**isual **R**eward **M**odel (RoVRM), which can improve human-preference alignment for LVLMs in two ways. For one, we propose a three-phase progressive training approach to gradually bridge the task and modality gaps between textual and visual preference data, which can take full advantage of auxiliary textual preference data to improve the robustness of RoVRM. Furthermore, considering the conflict in preferences (Coste et al. 2023; Eisenstein et al. 2023), leveraging textual preference data poses a problem: not all data is beneficial for training the RoVRM. Addressing this problem, we propose an optimal transport-based preference data selection approach. This approach can select textual preference data that better aligns with the vision-language task preferences, thereby improving the efficacy of the RoVRM training process. To the best of our knowledge, we are the first to investigate the integration of preferences from different modalities.

Through experiments on commonly used vision-language tasks, we aim to evaluate RoVRM using two human-preference alignment techniques: best-of-$n$ sampling and RL. Our results demonstrate improved performance in each task when aligned with reward signals from RoVRM. Notably, when performing best-of-$n$ sampling on the LLaVA-

[*] Authors contributed equally.
[†] Corresponding author.

1.5-7B model, RoVRM outperforms a traditional VRM by 8.4 points on the LLaVA-Bench benchmark.

As another bonus, our three-phase progressive training and preference data selection can be seamlessly integrated with arbitrary ranking-based alignment techniques, such as direct preference optimization (DPO) (Rafailov et al. 2024), SimPO (Meng, Xia, and Chen 2024), and ORPO (Hong, Lee, and Thorne 2024). For instance, on the LLaVA-1.5-13B model, integrating with DPO results in an additional improvement of 17.82 points on the MM-Instruct benchmark compared to standard DPO.

Our code is publicly available*. This version summarizes the key experimental setup and results, with further details provided in our arXiv submission†.

## Related Work

In recent years, LVLMs have served as the primary backbone for vision-language tasks (Achiam et al. 2023; AI 2023). Aligning LVLMs with human preferences is effective in gaining more performance (Liu et al. 2023; Wang et al. 2024b). However, in this process, they only used visual preference data and never leveraged the textual preference data that exists in abundance.

**Large Vision-Language Models** Inspired by the success of LLMs such as GPTs (Brown et al. 2020; Ouyang et al. 2022) and LLaMA (Touvron et al. 2023), researchers have been aiming to develop LVLMs. The basic idea is to augment LLMs with visual inputs (*e.g.,* images) to provide an interface for vision-language tasks (Alayrac et al. 2022; Awadalla et al. 2023; Aiello et al. 2023). Recent works on LVLMs could be classified into two groups. The first group focused on integrating visual information into LLMs (Chen et al. 2023; Liu et al. 2024a; Wang et al. 2024c). For example, Liu et al. (2024b) constructed a large amount of visual instruction data to pre-train the visual projection layer. Lin et al. (2024) further investigated the effective pre-training design options to augment LVLMs. The second group that has attracted attention commonly aimed to improve the consistency of output text and visual content, particularly addressing the problem of hallucination (Zhou et al. 2023; Leng et al. 2024; Gunjal, Yin, and Bas 2024; Huang et al. 2024a; Favero et al. 2024). This work belongs to the latter, where our RoVRM can improve the consistency of output text and visual content.

**Human-Preference Alignment for LVLMs** Reinforcement learning with human feedback (RLHF) has been shown to effectively align LLM behaviors with human preferences (Stiennon et al. 2020; Ouyang et al. 2022). Several works have improved RLHF by using fine-grained reward models (Wu et al. 2024), reward model ensembles (Coste et al. 2023), and direct preference optimization objectives (Rafailov et al. 2024). Additionally, some works focused on generating large, high-quality textual preference datasets to further improve RLHF in LLMs (Cui et al. 2023; Dubois et al. 2024). In the context of LVLMs, existing works mainly focused on the adaptation of the human-preference alignment techniques (Sun et al. 2023; Li et al. 2023a; Yu et al. 2024a). A significant challenge here was the scarcity of visual preference data. To address this challenge, many efforts have been made to create visual preference data, including collecting human preferences (Sun et al. 2023), and acquiring preferences from a strong LVLM (Li et al. 2023a; Yu et al. 2024b). Different from these works, we investigate how to leverage rich, high-quality textual preference data to offset the scarcity of visual preference data.

*https://github.com/NiuTrans/Vision-LLM-Alignment

†https://arxiv.org/abs/2408.12109

## Our Method

We first review the preliminaries of the human-preference alignment training for language models. Then, we present the three-phase progressive training for use with RoVRM. Last, we introduce the proposed preference data selection.

### Preliminaries

**Reinforcement Learning with Human Feedback** RLHF is a key technique for aligning language models with human preferences. It typically consists of two main steps: 1) training a reward model (also known as preference model) from preference data, and 2) using an RL algorithm, such as PPO (Schulman et al. 2017), to maximize the reward. In step 1, we usually employ the Bradley-Terry model (Bradley and Terry 1952). When the preference data existed in a comparison pair, the loss function can be written as:

$$\mathcal{L}_{reward} = -\log(\sigma(r_\theta(x, y_w) - r_\theta(x, y_l))) \tag{1}$$

where $\sigma$ is the Sigmoid activation function, $r(\cdot)$ is a reward model and $\theta$ is its parameters. $y_w$ and $y_l$ are two different responses for the human prompt $x$, where $y_w$ is more preferred than $y_l$. When dealing with multiple responses more than two, we can induce $\mathcal{L}_{reward}$ based on the more general Plackett-Luce model (Luce 2005):

$$\mathcal{L}_{reward} = -\sum_{i=1}^{k} \log \frac{\exp\left(r_\theta(x, y_i)\right)}{\sum_{j=i}^{k} \exp\left(r_\theta(x, y_j)\right)} \tag{2}$$

where $k$ denotes the number of responses. These responses are ranked by the defined preferences: $(y_1 \succ \cdots \succ y_k|x)$, where $y_1$ is the best while $y_k$ is the worst. In step 2, the reward signals produced by the trained reward model are instrumental in adjusting the parameters of the language models. Thus, the alignment of the language model is significantly influenced by how well the reward model is trained.

**Direct Preference Optimization** To bypass the complex RL procedure, Rafailov et al. (2024) proposed the direct preference optimization (DPO) which employs a reward model training objective to maximize rewards:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} = &-\log\sigma[\beta\log(\frac{p_{\theta'}(y_w|x)}{p_{\theta'_{old}}(y_w|x)}) \\ &-\beta\log(\frac{p_{\theta'}(y_l|x)}{p_{\theta'_{old}}(y_l|x)})]\end{aligned} \tag{3}$$

where $\theta'$ denotes the parameters of the language model, $\theta'_{old}$ denotes the parameters of the language model trained via supervised fine-tuning, $\beta$ denotes a scaling factor, and $\sigma$ denotes a Sigmoid function.

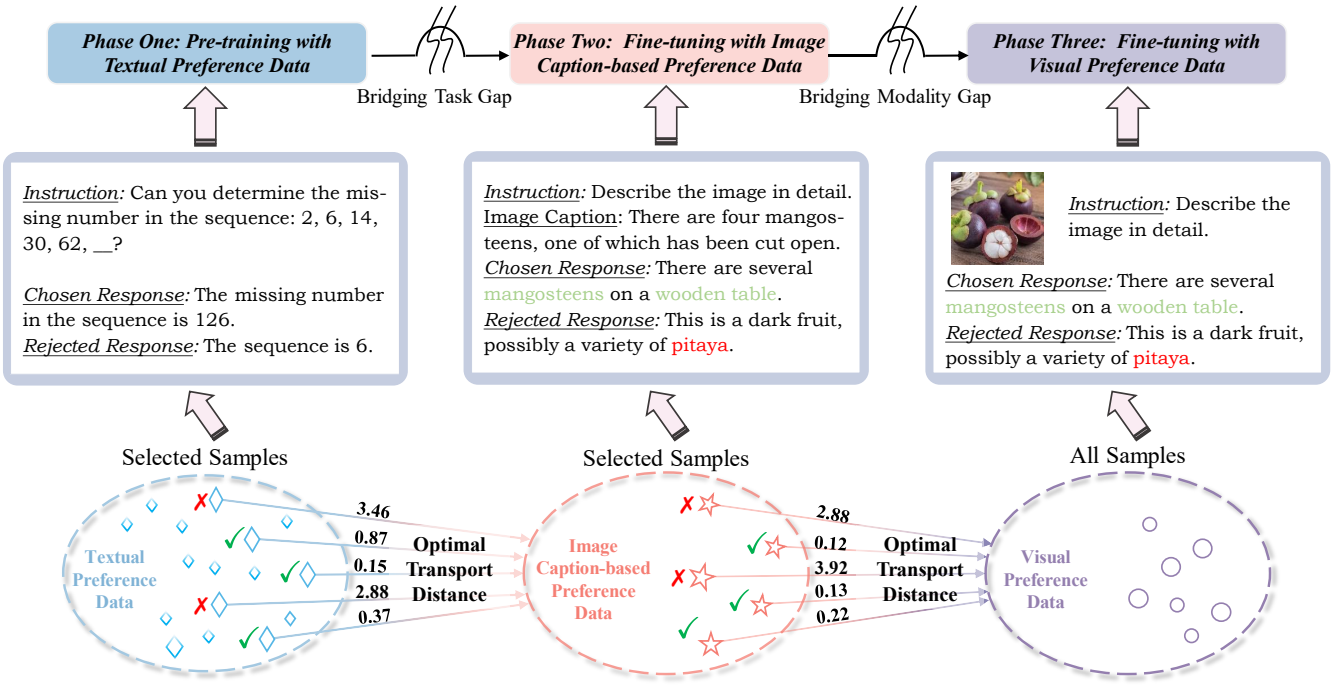



Figure 1: We propose three-phase progressive training and optimal transport-based preference data selection approaches to train RoVRM. For three-phase progressive training, we take full advantage of textual preference data to compensate for the limited availability of visual preference data. Using this preference selection, samples for phases one and two are selected based on those selected for the subsequent phase. Green ✓denotes a selected sample, while red ✗ denotes one that is not selected.

**Best-of-$n$ Sampling** Best-of-$n$ sampling (also known as re-ranking) refers to reordering or reevaluating a set of candidate responses sampled from a trained model (Lee, Auli, and Ranzato 2021; Fernandes et al. 2022). Given a set $\mathcal{Y}$ of $n$ candidate responses for $x$, we can also use the best-of-$n$ sampling approach to maximize the reward, thereby aligning the response with human preferences. Typically, we employ the reward model to score the candidate responses and select a final response that has a maximum reward score.

We can notice that when applying these alignment training methods to LVLMs, sufficient visual preference data is required either to train a VRM or to perform DPO training. However, in practice, visual preference data is often insufficient and expensive to acquire.

## A Robust Visual Reward Model

We aim to provide a RoVRM for human-preference alignment in LVLMs. The overview of training RoVRM is depicted in Figure 1. As shown in the figure, we present a three-phase progressive training and preference data selection to improve the robustness of RoVRM.

**Three-Phase Progressive Training** In response to the scarcity of visual preference data, we propose a three-phase progressive training approach that effectively solves this issue. Phase one is to conduct preference pre-training using a large amount of textual preference data. This phase can help our RoVRM to pre-learn general preferences. Ideally, the RoVRM would inherit these general preferences when processing vision-language tasks. However, this faces two serious obstacles: *task gap* and *modality gap*, which prevent these preferences from being directly applicable to vision-language tasks (see experiments in Figure 4). Here, we design phases two and three to bridge these gaps progressively. Phase two is to bridge the task gap by constructing vision-language preference data based on image captions and fine-tuning the RoVRM. Specifically, we use image captions to replace the images for visual preference data, *i.e.*, changing the human prompt $x$=[`Instruction`; `Image`] to $x$=[`Instruction`; `Image Caption`] in Eqs. 1 and 2. Building on phase two, phase three is to bridge the modality gap by using the visual preference data to continue fine-tuning the RoVRM with a visual projector. Compared to training a VRM directly with visual preference data, this three-phase training process incurs additional time costs due to an extra preference training session. However, it can leverage auxiliary textual preference data to improve robustness and respond to the scarcity of visual preference data. Furthermore, although pre-training followed by fine-tuning is widely used in machine learning (Devlin et al. 2019; Liu et al. 2019), our approach is the first to demonstrate the feasibility of optimizing a VRM through this paradigm.

**Preference Data Selection** Not all preference data aligns with the preferences used in subsequent phases, and conflicts may arise. Thus, during each training phase, we expect to employ samples that more closely align with the preferences contained in the data for the next phase. To

achieve this, we propose an optimal transport-based preference data selection approach. We apply this approach to perform preference data selection for phases one and two, based on the preference data used in the next phase. For instance, in phase one, following Xia et al. (2024)'s work, we first extract gradient features for all samples in the textual preference dataset $\mathcal{D}_{\mathrm{T}} = \{s_1^t, s_2^t, \cdots, s_m^t\}$. Based on these features, we compute the distance score between each sample in $\mathcal{D}_{\mathrm{T}}$ and the image caption-based preference dataset $\mathcal{D}_{\mathrm{C}} = \{s_1^c, s_2^c, \cdots, s_n^c\}$ using optimal transport. The details are described as follows.

*Gradient Feature.* Xia et al. (2024) construct gradient features for each sample of general supervised fine-tuning data to select the data that more effectively improves the specific downstream task. Here, using these gradient features, we conduct the preference data selection. Specifically, we firstly use LoRA (Hu et al. 2022) to efficiently perform a warmup reward model training with a small subset of preference data $\mathcal{D}_{\mathrm{Warmup}}$, where $\mathcal{D}_{\mathrm{Warmup}}$ is a subset extracted randomly from $\mathcal{D}_{\mathrm{T}} \cup \mathcal{D}_{\mathrm{C}}$. Then, we extract the gradient features for each preference sample in $\mathcal{D}_{\mathrm{T}}$ and $\mathcal{D}_{\mathrm{C}}$ through the forward and backpropagating on the warmed-up reward model:

$$g = \mathrm{RP}(\nabla \mathcal{L}_{reward}(s; \theta_{warmup})) \tag{4}$$

where $g$ is the gradient feature of the preference sample $s$ and $\theta_{warmup}$ is the parameters of the warmed-up reward model. $\mathrm{RP}(\cdot)$ is a random projection (Xie, Li, and Xue 2017) that reduces the dimensionality of gradient features.

*Optimal Transport-based Distance.* Unlike the Xia et al. (2024) who use the cosine similarity to compute sample distance scores, we use optimal transport (Villani et al. 2009), endowed with the capability to compute the distance transferring an arbitrary data feature to a specific data feature (Gurumoorthy, Jawanpuria, and Mishra 2021; Kang et al. 2024). Our motivation is to gather preference data for easy integration into the next training phase. To reduce computational overhead, we select a representative subset $\mathcal{D}_{\mathrm{SubC}}$ from $\mathcal{D}_{\mathrm{C}}$. This subset approximates the distance computation for the entire dataset $\mathcal{D}_{\mathrm{C}}$ when selecting samples from $\mathcal{D}_{\mathrm{T}}$. We define the distance score of $i$-th sample in $\mathcal{D}_{\mathrm{T}}$ by:

$$c_i = \frac{1}{|\mathcal{D}_{\mathrm{SubC}}|} \sum_{j=1}^{|\mathcal{D}_{\mathrm{SubC}}|} \mathrm{OT}(g_i^t, g_j^c) \tag{5}$$

where $g_i^t$ and $g_j^c$ denote the gradient features for the preference samples $s_i^t$ and $s_j^c$, respectively. $\mathrm{OT}(\cdot)$ denotes the function of computing the transfer distance. Given gradient features $g_i^t$, $g_j^c$ over a gradient space $\mathcal{Z}$, the optimal transport-based transfer distance can be defined as:

$$\mathrm{OT}(g_i^t, g_j^c) := \min_{\gamma \in \Gamma(g_i^t, g_j^c)} \int_{\mathcal{Z}^2} C(z, z')\, d\gamma(z, z') \tag{6}$$

where $C(\cdot)$ denotes a symmetric positive-definite cost function, and $\Gamma(g_i^t, g_j^c)$ denotes a collection of couplings between two gradients $g_i^t$ and $g_j^c$. Here, we utilize $L_2$-norm as the cost function and define the sum of the solved $\gamma$ as the distance score. A lower distance score indicates that the textual preference sample has preferences more easily transferable to the vision-language task. Our implementation of optimal transport solvers is done using Python Optimal Transport (`POT`)[‡]. While optimal transport distance has been used in data selection before (Kang et al. 2024), this is the first application to preference data selection.

To ensure that the ultimate goal of selecting preference data is to transfer preferences from textual preference data to vision-language tasks, we start by selecting image caption-based preference data for phase two. Next, we choose the textual preference data for phase one based on the preference data selected in phase two.

# Experiments

## Experimental Setups

**Datasets** The datasets used in this work are as follows:

- *Textual Preference Dataset*: We used UltraFeedback (Cui et al. 2023), a large-scale, high-quality, and diversified preference dataset, as our textual preference dataset. It comprises 64k instructions, each with 4 responses, leading to over 340k comparison preference pairs.
- *Image Caption-based Preference Dataset*: We constructed an image caption-based preference dataset to bridge the task gap. Specifically, we employed `GPT-4o-mini` to generate detailed image captions that replace the visual content in our preference data. Note that when the image is present in the COCO caption dataset[§], we used the human-annotated captions directly.
- *Visual Preference Dataset*: We employed the visual preference dataset from RLAIF-V (Yu et al. 2024b), which consists of about 83k comparison preference pairs. To our knowledge, it is the largest scale open source preference dataset in computer vision.
- *RL Training*: We sampled 50k instructions from LLaVA-Instruct-150K (Liu et al. 2024b) for training.

**Settings** For training RoVRM, we used the LLaVA-1.5-7B model to initialize the visual reward model. The learning rates for the three-phase progressive training were set to 2e-5 for phase one, and 1e-6 for phases two and three. For optimal transport-based preference data selection, we used 5k samples to warm up the VRM, consisting of 2k samples from the dataset to be selected and 3k samples from the target preference dataset. The representative subset size was set to 5k samples. For best-of-$n$ sampling and RL training, we employed the LLaVA-1.5-7B as the initial model. In the process of best-of-$n$ sampling, we set the sampling size to 8. We also tested other sampling sizes in Figure 5.

**Evaluation** We evaluated the RoVRM in two aspects: trustworthiness, which denotes the level of hallucination, and helpfulness, which reflects overall interaction capability. Trustworthiness was evaluated using two benchmarks: MMHal-Bench (Sun et al. 2023) and AMBER (Wang et al. 2023). GPT-4 was employed to evaluate the response-level hallucination rate (**HalRate**) and informativeness score

[‡]https://pythonot.github.io/index.html

[§]https://huggingface.co/datasets/lmms-lab/COCO-Caption2017

| Method | #Param | MMHalBench | | AMBER | | $\text{LLaVA}^{\text{W}}$ | MMIns |
|---|---|---|---|---|---|---|---|
| | | Score ↑ | HalRate ↓ | Cover. ↑ | HalRate ↓ | Score ↑ | WinRate ↑ |
| Qwen-VL-Chat | 10B | 2.76 | 38.5 | 53.2 | 31.0 | 71.9 | 73.58 |
| OmniLMM | 12B | 3.14 | 36.5 | - | - | 72.7 | - |
| MiniGemini | 34B | 3.08 | 38.5 | - | - | 79.2 | - |
| LLaVA-NeXT | 34B | 3.31 | 34.4 | 63.2 | 43.6 | 77.7 | 93.83 |
| LURE | 7B | 1.64 | 60.4 | - | - | 36.9 | - |
| HA-DPO | 7B | 1.98 | 60.4 | 49.5 | 29.1 | 60.3 | - |
| VCD | 7B | 2.12 | 54.2 | 51.5 | 39.0 | 65.8 | 42.56 |
| Silkie | 10B | 3.19 | 32.3 | 56.0 | 28.4 | 73.2 | 63.64 |
| LLaVA-RLHF | 13B | 2.02 | 62.5 | 52.0 | 39.2 | 61.5 | 74.24 |
| ***Best-of-$n$ Sampling*** | | | | | | | |
| LLaVA-1.5-7B | 7B | 2.12 | 55.0 | 50.3 | 37.1 | 66.7 | 46.16 |
| +VRM-Vanilla | 7B | 2.39 | 47.9 | 50.8 | 29.0 | 73.6 | 57.69 |
| +RoVRM-Random | 7B | 2.52 | 43.8 | 51.7 | 26.9 | 77.2 | 58.49 |
| +RoVRM | 7B | **2.68** | **40.6** | **53.2** | **23.9** | **82.0** | **61.91** |
| LLaVA-1.5-13B | 13B | 2.30 | 53.8 | 50.6 | 37.2 | 75.6 | 50.00 |
| +VRM-Vanilla | 13B | 2.41 | 51.0 | 51.4 | 26.6 | 84.0 | 73.08 |
| +RoVRM-Random | 13B | 2.43 | 48.3 | 51.9 | 25.7 | 86.4 | 74.42 |
| +RoVRM | 13B | **2.57** | **47.3** | **53.6** | **22.8** | **89.8** | **78.75** |
| ***Reinforcement Learning*** | | | | | | | |
| LLaVA-1.5-7B | 7B | 2.12 | 55.0 | **50.3** | 37.1 | 66.7 | 46.16 |
| +VRM-Vanilla | 7B | 2.17 | 53.2 | 49.1 | 29.1 | 72.8 | 51.11 |
| +RoVRM-Random | 7B | 2.21 | 50.8 | 48.7 | 24.3 | 74.2 | 54.35 |
| +RoVRM | 7B | **2.36** | **48.9** | 48.2 | **23.4** | **78.3** | **58.69** |
| LLaVA-1.5-13B | 13B | 2.30 | 53.8 | **50.6** | 37.2 | 75.6 | 50.00 |
| +VRM-Vanilla | 13B | 2.49 | 50.0 | 41.1 | 23.2 | 78.2 | 52.63 |
| +RoVRM-Random | 13B | 2.34 | 47.9 | 48.6 | 21.0 | 79.5 | 60.53 |
| +RoVRM | 13B | **2.57** | **43.8** | 47.7 | **19.5** | **81.7** | **65.79** |

Table 1: Experimental results on different vision-language tasks. The best results for each group are in bold.

(**Score**) on the MMHalBench. We also provided the object coverage (**Cover.**) and hallucination rate metrics for AMBER. To assess helpfulness, we used two benchmarks: MM-Instruct (Liu et al. 2024c) and LLaVA-Bench (In-the-Wild) (Liu et al. 2024b). GPT-4, following the settings in `lmms-eval`[¶], was used to score responses in LLaVA-Bench. For MM-Instruct, responses from LLaVA-1.5-13B were used as a baseline, and we computed the win rate (**WinRate**) as per Liu et al. (2024c).

**Baselines** Our baselines were the **LLaVA-1.5-7B** and **-13B** models without human-preference alignment. We also compared with other general LVLMs, including **Qwen-VL-Chat** (Bai et al. 2023), **OmniLMM** (Hu et al. 2023), and **MiniGemini** (Li et al. 2024). Furthermore, we compared RoVRM with commonly used methods to solve the hallucination, including **LURE** (Zhou et al. 2023), **HA-DPO** (Zhao et al. 2023), **VCD** (Leng et al. 2024), **Silkie** (Li et al. 2023a), and **LLaVA-RLHF** (Sun et al. 2023). The traditional VRM training was also our baseline, where we optimized a VRM only using our visual preference dataset (**VRM-Vanilla**). To evaluate the effectiveness of optimal transport, we chose **RoVRM-Random** as a baseline, where we randomly selected samples during the preference data selection.

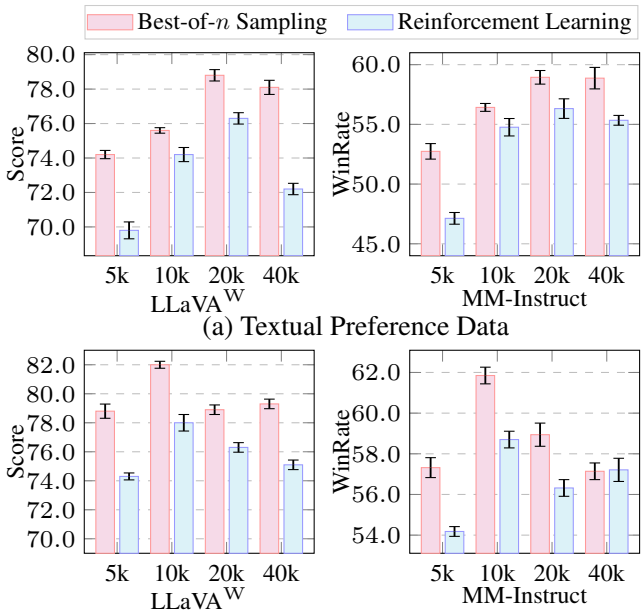
## Experimental Results

**Results of Best-of-$n$ Sampling** Table 1 summarizes the performance of our RoVRM on the best-of-$n$ sampling. On all vision-language tasks, RoVRM consistently outperforms the VRM-Vanilla which does not use textual preference data. For instance, when using the LLaVA-1.5-7B model, RoVRM can outperform VRM-Vanilla by 8.4 points on the LLaVA-Bench. We also observe this consistent phenomenon on the LLaVA-1.5-13B model. Moreover, from the results, we find that RoVRM significantly reduces visual hallucinations, *e.g.*, lowering the hallucination rate by 13.2 points in the LLaVA-1.5-7B model. We attribute this improvement to the extensive use of textual preference data, which improves VRM's capacity to evaluate facticity. Interestingly, we also find that RoVRM enables the LLaVA-1.5 models to outperform stronger LVLMs, with the LLaVA-

[¶]https://github.com/EvolvingLMMs-Lab/lmms-eval

| Method | AMBER | | LLaVA$^{W}$ |
|---|---|---|---|
| | Cover. ↑ | HalRate ↓ | Score ↑ |
| LLaVA-1.5-7B | 50.3 | 37.1 | 66.7 |
| ***Best-of-$n$ Sampling*** | | | |
| RoVRM | **53.2** | **23.9** | **82.0** |
| w/o PDS | 52.4 | 25.1 | 80.6 |
| w/o TPT-One | 51.0 | 26.7 | 71.3 |
| w/o TPT-Two | 51.8 | 24.9 | 78.0 |
| ***Reinforcement Learning*** | | | |
| RoVRM | **48.2** | **23.4** | **78.3** |
| w/o PDS | 46.2 | 32.2 | 75.2 |
| w/o TPT-One | 44.3 | 35.0 | 73.0 |
| w/o TPT-Two | 47.5 | 28.2 | 76.1 |

Table 2: The suffixes "-One" and "-Two" denote the removal of phases one and two, respectively, in the three-phase progressive training approach. "w/o PDS" denotes that all data is used for each training phase without employing preference data selection. PDS: preference data selection; TPT: three-phase progressive training.




(a) Textual Preference Data

(b) Image Caption-based Preference Data

Figure 2: We train RoVRM with varying amounts of textual and image caption preference data. Experiments are conducted on the LLaVA-1.5-7B model using three different seeds, and we report the average results along with their standard deviation.

1.5-7B model even surpassing the LLaVA-1.5-13B model on most of the benchmarks, such as MMHalBench and LLaVA-Bench. This finding shows a promising direction for achieving *weak-to-strong generalization* (Burns et al. 2023).

**Results of Reinforcement Learning** Compared to best-of-$n$ sampling, RL typically requires a more robust reward model: The reward model not only evaluates responses as "good" or "bad" but also provides an accuracy score margin between the responses (Zhou et al. 2024a). From the results, we find that RoVRM fulfills this requirement more effectively than VRM-Vanilla, resulting in improved RL training performance in LVLMs. For instance, in RL training on the LLaVA-1.5-7B model, RoVRM surpasses VRM-Vanilla by 7.58 points on MM-Instruct. This finding demonstrates that RoVRM is robust and can deliver high-quality reward signals across various alignment techniques. Additionally, we observe that RL training reduces hallucinations but slightly decreases the "Cover." metric, which is consistent with the findings of Meng, Xia, and Chen (2024)'s work and DPO training in Table 3. We conjecture that preference alignment training may slightly hurt the instruction-following capability of LVLMs (Wang et al. 2024a).

Furthermore, compared to RoVRM-Random, RoVRM shows better performance across all benchmarks. This indicates that optimal transport-based preference data selection outperforms random selection. However, RoVRM-Random also significantly improves performance over VRM-Vanilla.

## Ablation Study

We present detailed ablation studies to investigate the effects of three-phase progressive training and our preference data selection approach. The experiments are conducted on the LLaVA-1.5-7B model and the impacts of removing each approach were thoroughly examined. Furthermore, we study the impact of eliminating the distinct designs of phases one and two. The results are summarized in Table 2. Through the results, we can see that three-phase progressive training significantly improves the performance of RoVRM in both best-of-$n$ sampling and RL. Notably, removing phase one leads to a substantial performance decline (*e.g.*, a loss of 10.7 points on the LLaVA-Bench for best-of-$n$ sampling), highlighting the importance of textual preference data in training RoVRM. Likewise, removing image caption-based preference data also results in performance loss, indicating the need to address the task gap. Additionally, we see that using the preference data selection can train a better RoVRM. It shows the effectiveness of using optimal transport to conduct preference data selection.

## Analysis

**Performance on Different Numbers of Selected Preference Samples** We investigate the impact of different numbers of selected preference samples using a three-phase progressive training with LLaVA-Bench and MM-Instruct. We test sample sizes of 5k, 10k, 20k, and 40k, alongside 20k image caption-based preference samples (Figure 2(a)). Our results show that using 20k textual preference samples yields strong performance, even outperforming the 40k sample scenario. Consequently, we choose 20k textual preference samples for phase one to train our RoVRM. Similarly, we evalu-

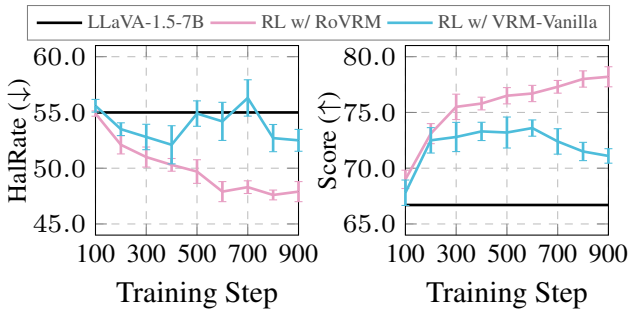



Figure 3: Performance during RL training is evaluated on the MMHalBench (left) and LLaVA-Bench (right) benchmarks using three different seeds.

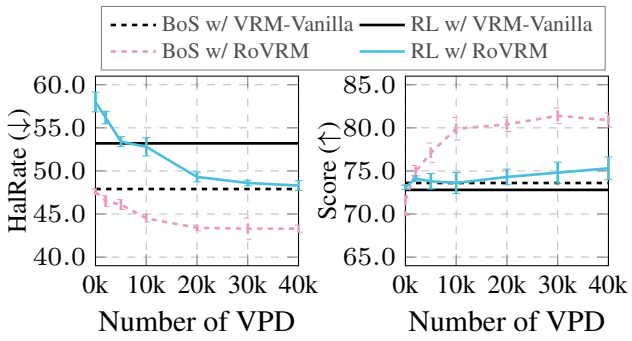



Figure 4: Performance of best-of-$n$ sampling (BoS) and RL on MMHalBench (left) and LLaVA-Bench (right) across three different seeds.

ate sample sizes of 5k, 10k, 20k, and 40k for phase two, *i.e.*, image caption-based preference data selection (Figure 2(b)), identifying 10k as the optimal sample size.

**Comparison of RL Training Process on Different VRMs** Figure 3 illustrates the performance of the LLaVA-1.5-7B model comparing RL training with VRM-Vanilla and RoVRM. The results show that RL training with RoVRM improves performance more effectively than VRM-Vanilla. Additionally, we observe that RoVRM can lead to a more stable RL training process by mitigating reward over-optimization (Gao, Schulman, and Hilton 2023).

**Enabling Few-Shot Learning in VRM** Figure 4 shows RoVRM's performance with different numbers of visual preference data. Note that when the visual preference dataset is small (*i.e.*, 1k, 5k, and 10k), we use the entire dataset without image caption-based preference data selection. From the results, we find that pre-training with textual preference data enables effective few-shot learning in VRM (Wang et al. 2020). Based on these textual preferences, the reward model quickly generalizes to vision-language tasks using only a few visual preference samples. Notably, using only 5k visual preference samples can achieve a performance comparable to that of VRM-Vanilla trained with 83k samples. However, while it is feasible to directly use a textual reward model (*i.e.*, using 0k visual preference data) to optimize LVLM, the results are worse, particularly during RL training.

| Method | AMBER | | LLaVA$^W$ |
|---|---|---|---|
| | Cover. ↑ | HalRate ↓ | Score ↑ |
| LLaVA-1.5-7B | 50.3 | 37.1 | 66.7 |
| +DPO | 49.6 | 22.2 | 80.9 |
| +RoDPO | **50.7** | **17.6** | **83.7** |
| LLaVA-1.5-13B | **50.6** | 37.2 | 75.6 |
| +DPO | 49.2 | 15.7 | 84.2 |
| +RoDPO | 49.8 | **12.8** | **86.4** |

Table 3: Performance on the direct preference optimization.

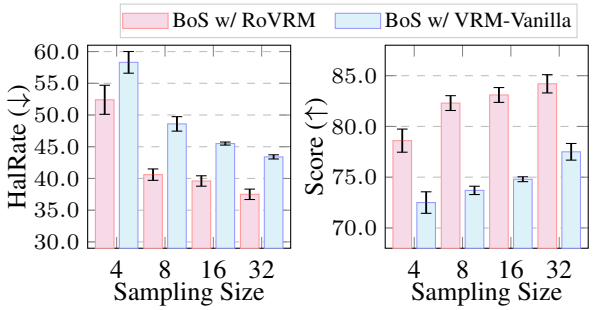



Figure 5: Performance of best-of-$n$ sampling (BoS) with different sampling sizes: 4, 8, 16, and 32.

**Integration with Direct Preference Optimisation** Despite bypassing reward model training, direct preference optimization (DPO) still requires preference data to train the language model with a ranking-loss function. Consequently, DPO also faces the challenge of limited visual preference data in LVLMs. To address this, we propose a **Ro**bust **DPO** (namely RoDPO) by integrating our three-phase progressive training and preference data selection. Our experiments on the LLaVA-1.5-7B and -13B models show that RoDPO performs better than DPO, as summarized in Table 3.

**Performance on Different Sampling Sizes** We evaluate the performance of best-of-$n$ sampling with varying sample sizes using the LLaVA-1.5-7B model. Figure 5 presents a comparison of RoVRM and VRM-Vanilla on the MMHal-Bench (left) and LLaVA-Bench (right) benchmarks. The experimental results indicate that RoVRM consistently enhances performance across different sampling sizes, highlighting its improved robustness.

## Conclusion

In this paper, we focus on improving the human-preference alignment for LVLMs. We present a **Ro**bust **V**isual **R**eward **M**odel (namely RoVRM) via three-phase progressive training and preference data selection approaches. Our extensive experiments demonstrate that our RoVRM significantly outperforms the traditional visual reward model.

## Acknowledgments

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Natural Science Foundation of Liaoning Province of China (2022-KF-26-01), the Fundamental Research Funds for the Central Universities (Nos. N2216016 and N2316002), the Yunnan Fundamental Research Projects (No. 202401BC070021), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt; et al. 2023. Gpt-4 technical report. *ArXiv preprint*.

AI, A. 2023. Fuyu-8b: A multimodal architecture for ai agents.

Aiello, E.; Yu, L.; Nie, Y.; Aghajanyan, A.; and Oguz, B. 2023. Jointly training large autoregressive multimodal models. *ArXiv preprint*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Proc. of NeurIPS*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Proc. of NeurIPS*.

Burns, C.; Izmailov, P.; Kirchner, J. H.; Baker, B.; Gao, L.; Aschenbrenner, L.; Chen, Y.; Ecoffet, A.; Joglekar, M.; Leike, J.; et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *ArXiv preprint*.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv preprint*.

Coste, T.; Anwar, U.; Kirk, R.; and Krueger, D. 2023. Reward model ensembles help mitigate overoptimization. *ArXiv preprint*.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *ArXiv preprint*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

Dubois, Y.; Li, C. X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P. S.; and Hashimoto, T. B. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Proc. of NeurIPS*, 36.

Eisenstein, J.; Nagpal, C.; Agarwal, A.; Beirami, A.; D'Amour, A.; Dvijotham, D.; Fisch, A.; Heller, K.; Pfohl, S.; Ramachandran, D.; et al. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *ArXiv preprint*.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-modal hallucination control by visual information grounding. In *Proc. of CVPR*.

Fernandes, P.; Farinhas, A.; Rei, R.; C. de Souza, J. G.; Ogayo, P.; Neubig, G.; and Martins, A. 2022. Quality-Aware Decoding for Neural Machine Translation. In *Proc. of NAACL*.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *Proc. of ICML*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proc. of AAAI*.

Gurumoorthy, K. S.; Jawanpuria, P.; and Mishra, B. 2021. SPOT: A framework for selection of prototypes using optimal transport. In *Proc. of KDD*.

Hong, J.; Lee, N.; and Thorne, J. 2024. Orpo: Monolithic preference optimization without reference model. *ArXiv preprint*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of ICLR*.

Hu, J.; Yao, Y.; Wang, C.; Wang, S.; and Pan, e. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *ArXiv preprint*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proc. of CVPR*.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; et al. 2024b. Language is not all you need: Aligning perception with language models. *Proc. of NeurIPS*.

Kang, F.; Just, H. A.; Sun, Y.; Jahagirdar, H.; Zhang, Y.; Du, R.; Sahu, A. K.; and Jia, R. 2024. Get more for less: Principled Data Selection for Warming Up Fine-Tuning in LLMs. *ArXiv preprint*.

Lee, A.; Auli, M.; and Ranzato, M. 2021. Discriminative Reranking for Neural Machine Translation. In *Proc. of ACL*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proc. of CVPR*.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023a. Silkie: Preference distillation for large visual language models. *ArXiv preprint*.

Li, Y.; Zhang, C.; Yu, G.; Wang, Z.; Fu, B.; Lin, G.; Shen, C.; Chen, L.; and Wei, Y. 2023b. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *ArXiv preprint*.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *ArXiv preprint*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proc. of CVPR*.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023. Aligning large multi-modal model with robust instruction tuning. *ArXiv preprint*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proc. of CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Proc. of NeurIPS*.

Liu, J.; Huang, X.; Zheng, J.; Liu, B.; Wang, J.; Yoshie, O.; Liu, Y.; and Li, H. 2024c. MM-Instruct: Generated Visual Instructions for Large Multimodal Model Alignment. *ArXiv preprint*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*.

Luce, R. D. 2005. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *ArXiv preprint*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Proc. of NeurIPS*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Proc. of NeurIPS*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *ArXiv preprint*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. In *Proc. of NeurIPS*.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*.

Villani, C.; et al. 2009. *Optimal transport: old and new*. Springer.

Wang, C.; Zhou, H.; Chang, K.; Li, B.; Mu, Y.; Xiao, T.; Liu, T.; and Zhu, J. 2024a. Hybrid Alignment Training for Large Language Models. *ArXiv preprint*.

Wang, F.; Zhou, W.; Huang, J. Y.; Xu, N.; Zhang, S.; Poon, H.; and Chen, M. 2024b. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. *ArXiv preprint*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *ArXiv preprint*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2024c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Proc. of NeurIPS*.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2024. Fine-grained human feedback gives better rewards for language model training. *Proc. of NeurIPS*.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. Less: Selecting influential data for targeted instruction tuning. *ArXiv preprint*.

Xie, H.; Li, J.; and Xue, H. 2017. A survey of dimensionality reduction techniques based on random projection. *ArXiv preprint*.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proc. of CVPR*.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *ArXiv preprint*.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *ArXiv preprint*.

Zhou, H.; Wang, C.; Hu, Y.; Xiao, T.; Zhang, C.; and Zhu, J. 2024a. Prior Constraints-based Reward Model Training for Aligning Large Language Models. *ArXiv preprint*.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024b. Aligning modalities in vision large language models via preference fine-tuning. *ArXiv preprint*.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and mitigating object hallucination in large vision-language models. *ArXiv preprint*.