TARS: MinMax Token-Adaptive Preference Strategy for Hallucination Reduction in MLLMs

Kejia Zhang[1], Keda Tao[2], Zhiming Luo[1†], Chang Liu[4⋆], Jiasheng Tang[3,5], and Huan Wang[2†]

[1]Xiamen University, China [2]Westlake University, China
[3]DAMO Academy, Alibaba Group, China
[4]AWS AI Lab, Amazon, United States [5]Hupan Laboratory, China

Project Page: https://kejiazhang-robust.github.io/tars_web

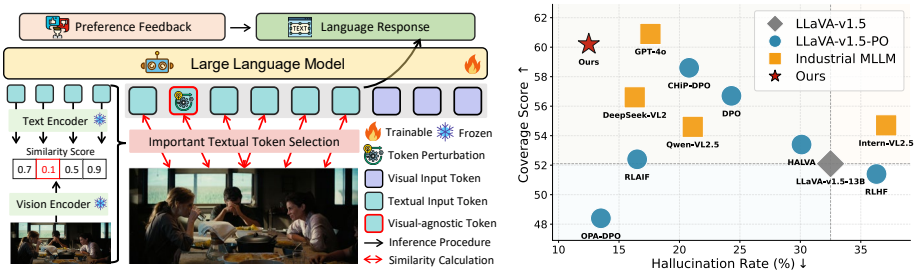


Fig. 1: Left: *TARS*, a token-adaptive preference strategy for mitigating hallucinations in MLLMs, recasts DPO as a min-max objective: (1) minimizing behavioral misalignment via structured preference feedback and (2) maximizing distributional adaptability by perturbing visual-agnostic tokens. **Right:** On LLaVA-v1.5-13B under AMBER [69], TARS improves over preference-optimization baselines and approaches GPT-4o [31] on key hallucination metrics.

Abstract. Multimodal large language models (MLLMs) often hallucinate, producing fluent but visually ungrounded outputs, partly because direct preference optimization (DPO) overfits to superficial linguistic cues under static preference supervision. We propose TARS, a token-adaptive preference strategy that reformulates DPO as a min-max optimization problem: the inner maximization perturbs visual-agnostic tokens to induce worst-case distributional shifts, while the outer minimization enforces alignment with causal visual signals rather than surface-level patterns. A spectral alignment loss further regularizes hidden representations in the frequency domain via the Fast Fourier Transform (FFT). With only 4.8k preference samples and no expert feedback, TARS halves the hallucination rate (26.4% → 13.2%) and reduces the cognition score from 2.5 to 0.4, outperforming standard DPO. It also surpasses 5× larger LLM-based data augmentation (28.8k samples; Hal-Rate 16.0% vs. 13.2%) and narrows the gap to GPT-4o, suggesting that token-adaptive

⋆ Work done prior to joining Amazon.

† Corresponding to: Huan Wang <wanghuan@westlake.edu.cn>, Zhiming Luo <zhiming.luo@xmu.edu.cn>

optimization improves data efficiency beyond simply scaling augmented preference data.

Keywords: Multimodal Large Language Models · Hallucination · Preference Optimization

1 Introduction

Multimodal large language models (MLLMs) extend the reasoning capabilities of LLMs [13, 26, 35, 70] to visual inputs, enabling grounded vision-language understanding [23, 30, 66]. Despite strong performance across diverse tasks [36, 60], MLLMs remain prone to hallucinations, producing outputs that appear plausible but are factually incorrect or lack visual grounding [27, 29, 34, 37, 57]. Mitigating such failures is crucial for deploying reliable MLLMs.

Modern MLLMs follow a two-stage pipeline of knowledge pretraining [5, 19, 79] and instruction tuning [40, 41]. Hallucinations often stem not from knowledge deficits but from behavioral biases that produce plausible yet ungrounded outputs [11, 47]. Preference optimization (PO) addresses this by fine-tuning with ranked response pairs from human [49, 64] or AI feedback [61, 78], aligning outputs with factual expectations [2, 58]. Direct preference optimization (DPO) [53] is widely used for hallucination reduction [24, 75], yet current methods can overfit to shallow textual cues such as high-frequency phrases [29, 43], generating plausible but visually ungrounded responses (Fig. 2(c)). Our analysis (Fig. 2(d)) further reveals that DPO-trained models assign high preference to outputs with spurious correlation tokens (*e.g.*, prepositions or frequently mentioned objects) that lack visual grounding [67, 72]. This reliance on static preference signals hinders generalization under shifting visual-textual contexts, leading to brittle alignment and increased hallucination [24, 59].

We formulate this challenge as a **min-max token-adaptive alignment problem**: maximizing distributional variation under semantic constraints, then minimizing preference loss under these perturbations. A natural alternative is data augmentation, which diversifies the preference distribution by generating additional training pairs. However, our experiments (Sec. 4.7) show that even 5× LLM-based augmentation (28.8k samples) underperforms TARS using only the original 4.8k samples (Hal-Rate: 16.0% vs. 13.2%). This result suggests that adversarial token perturbation improves this alignment setting more effectively than simply scaling augmented preference data. The key distinction is that data augmentation enriches the **data distribution**, whereas TARS actively reshapes the **optimization landscape**: the inner maximization dynamically generates worst-case token configurations per sample, forcing the outer minimization to learn alignment features that are invariant to distributional shifts rather than memorizing surface-level patterns.

Specifically, we perturb visual-agnostic tokens, *i.e.*, textual elements with minimal cross-modal grounding, to shift the input distribution without altering semantic content. This forces the model to rely on causally grounded visual

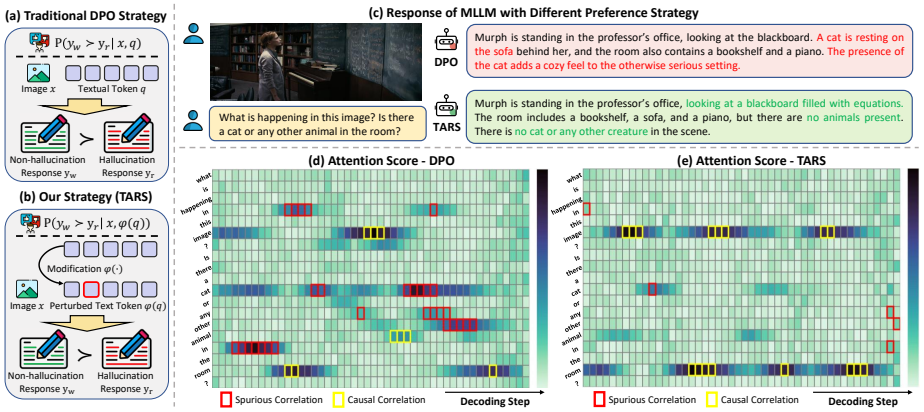


Fig. 2: Motivation illustration for TARS. (a) and (b) illustrate standard DPO and our token-adaptive perturbation strategy. (c) shows a VQA example where DPO hallucinates, while TARS effectively avoids ungrounded output. (d) and (e) visualize token-to-query attention maps during autoregressive decoding. DPO over-attends to spurious tokens, while TARS attends to causally grounded visual-semantic cues.

signals rather than superficial correlations (Fig. 2(b)). Supervising perturbed representations requires maintaining semantic consistency without rigid token-level correspondence. Standard spatial constraints (*e.g.*, ℓ_2 or contrastive losses) indiscriminately penalize any positional deviation equally, reintroducing the spurious correlations our perturbation aims to eliminate [65,83]. We instead propose a novel **spectral alignment loss** via the Fast Fourier Transform (FFT). Under the shift property of the Discrete Fourier Transform, a local token perturbation at position l contributes a bounded additive term proportional to $e^{-2\pi ikl/L}$ to each frequency bin k, regardless of l [18]. This means spectral magnitudes vary smoothly under local perturbation, whereas spatial metrics (ℓ_2, cosine) exhibit sharp discontinuities at the perturbed positions. Meanwhile, the dominant low-frequency components capture global semantic structure (*e.g.*, topic, intent, visual grounding), while high-frequency components encode position-dependent lexical details [73, 80]. This makes frequency-domain alignment a theoretically principled choice for preserving global semantic invariance while tolerating the local distributional shifts inherent in our min-max formulation (Sec. 3.3).

We refer to this approach as **TARS** (token-adaptive preference strategy), a lightweight and generalizable approach that enhances preference learning by combining adaptive token perturbation with frequency-domain spectral regularization. We evaluate TARS on LLaVA-v1.5 [42] at 7B and 13B scales across generative and discriminative hallucination benchmarks. TARS achieves strong performance across all benchmarks and matches or exceeds GPT-4o [31] on selected hallucination metrics, supporting the effectiveness of token-adaptive preference optimization. Our contributions are as follows:

– We reformulate preference learning as a min-max optimization that dynamically generates worst-case token perturbations per sample, reshaping the

optimization landscape rather than augmenting data. With only 4.8k samples, TARS outperforms 5× LLM-based augmentation (28.8k samples).
- We propose a **spectral alignment loss** via FFT that preserves global semantic structure in dominant low-frequency components while tolerating local distributional shifts from token perturbation, avoiding spurious positional correlations reintroduced by spatial constraints (*e.g.*, ℓ_2, cosine).
- We present **TARS**, combining adaptive token perturbation with spectral regularization. Using only 4.8k samples without expert feedback, TARS strongly reduces hallucinations and matches or exceeds GPT-4o on selected hallucination metrics while preserving general reasoning capabilities.

2 Preliminaries

Multimodal Large Language Models. MLLMs extend LLMs by incorporating visual inputs alongside textual prompts [79]. Formally, given an image x and a prompt q, the model generates a textual response $y = (y_1, \ldots, y_l)$ in an autoregressive manner [42]:

$$y_t \sim \pi_\theta(y_t \mid y_{<t}, x, q), \tag{1}$$

where π_θ denotes the conditional generation policy parameterized by θ. Given a textual input q and a visual input x, the model tokenizes them into discrete sequences: textual tokens $q = \{q_1, \ldots, q_m\}$ and visual tokens $x = \{x_1, \ldots, x_n\}$. In practice, a pretrained vision encoder extracts patch-level features from the image, which are projected into the language model's embedding space through a learnable alignment module. These tokens are mapped to embeddings and fused via cross-attention to integrate semantic signals from both modalities. The resulting multimodal context is then used by the decoder to autoregressively generate the output sequence [20, 74].

Direct Preference Optimization. Direct preference optimization (DPO) [53] is an effective and widely adopted approach for aligning model behavior with human preferences. It bypasses explicit reward models by directly optimizing preferences from pairwise comparisons.

Traditional methods such as reinforcement learning with human feedback (RLHF) [49] and AI feedback (RLAIF) [78] rely on training a scalar reward model $r_\psi(x, q, y)$ from preference pairs. This reward model is typically trained using the Bradley-Terry formulation [7]:

$$\begin{aligned} P\left(y_w \succ y_r \mid x, q\right) &= \frac{\exp(r_\psi(x, q, y_w))}{\exp(r_\psi(x, q, y_w)) + \exp(r_\psi(x, q, y_r))} \\ &= \sigma\left(r_\psi(x, q, y_w) - r_\psi(x, q, y_r)\right), \end{aligned} \tag{2}$$

where (x, q, y_w, y_r) is sampled from the preference data distribution $\mathcal{D}$, and $\sigma(z) = \frac{1}{1+\exp(-z)}$ denotes the sigmoid function. y_w and y_r denote the preferred and dispreferred responses, respectively. The reward model $r_\psi(x, q, y)$ is then

trained to maximize the log-likelihood of correctly ranking the preferred response over the dispreferred one during optimization:

$$\min_{r_\psi} \mathbb{E}_{(x,q,y_w,y_r)\sim\mathcal{D}} \left[-\log \sigma \left(r_\psi(x,q,y_w) - r_\psi(x,q,y_r) \right) \right]. \tag{3}$$

After training, the learned reward model $r_\psi(x,q,y)$ is used to guide the fine-tuning of the policy π_θ. Specifically, the policy is optimized to generate high-reward responses while minimizing divergence from a fixed reference policy π_{ref}, typically using KL-regularized objectives:

$$\min_{\pi_\theta} \mathbb{E}_{(x,q)\sim\mathcal{D},\, y^*\sim\pi_\theta} \big[-\big(r_\psi(x,q,y^*) - \alpha \cdot \mathbb{D}_{\text{KL}}(\pi_\theta(y^* \mid x,q) \| \pi_{\text{ref}}(y^* \mid x,q)) \big) \big]. \tag{4}$$

where α controls the strength of KL regularization, which ensures alignment with the learned preferences. Rather than relying on the explicitly trained reward model, DPO [53] simplifies the learning process by leveraging the insight that the optimal policy can be expressed in closed form using relative log-likelihoods under π_θ and π_{ref}:

$$\min_{\pi_\theta} \mathbb{E}_{(x,q,y_w,y_r)\sim\mathcal{D}} \big[-\log \sigma \big(\alpha \log \tfrac{\pi_\theta(y_w|x,q)}{\pi_{\text{ref}}(y_w|x,q)} - \alpha \log \tfrac{\pi_\theta(y_r|x,q)}{\pi_{\text{ref}}(y_r|x,q)} \big) \big]. \tag{5}$$

This formulation enables direct policy optimization from preference pairs, aligning the output probabilities with human preferences and improving alignment stability and sample efficiency compared to RL-based approaches.

3 Method

We propose a token-adaptive min-max strategy with perturbations on visual-agnostic tokens and a frequency-based regularizer for improved alignment. An overview is shown in Fig. 3, and the detailed algorithm is provided in Supplementary Section 8. We first define notation used throughout: x denotes the visual input, $q = \{q_1, \ldots, q_m\}$ the tokenized textual prompt, y_w and y_r the preferred and dispreferred responses, π_θ the trainable policy, π_{ref} the frozen reference policy, and $\mathcal{D}$ the preference dataset.

3.1 Min-Max Reformulation of DPO

To address the limitations of traditional DPO, we reformulate preference optimization as a *token-adaptive min-max game*. The inner maximization introduces a controlled token-level perturbation function $\varphi(\cdot)$, which modifies selected tokens in the prompt q to induce input distribution shifts. The outer minimization then aligns the policy π_θ with preference signals under these perturbations. Formally, we define the min–max preference objective as:

$$\min_{\pi_\theta} \max_{\varphi\in\Phi(\mathcal{A})} \mathbb{E}_{(x,q,y_w,y_r)\sim\mathcal{D}} \left[\mathcal{L}_{\text{TARS}}\big(x, \varphi(q), y_w, y_r\big) \right], \tag{6}$$

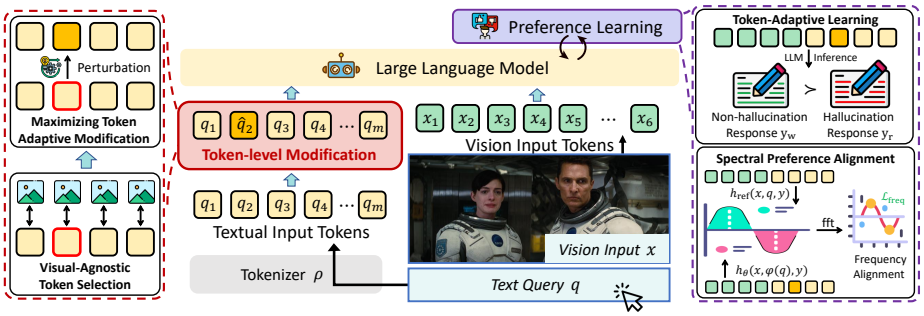


Fig. 3: Overview of **TARS**. TARS reformulates preference optimization as a Min–Max problem: (1) The maximization branch perturbs visual-agnostic tokens to simulate semantically shifted contexts (red dashed box); (2) The minimization branch fine-tunes the model to align with human preferences via the DPO objective (purple dashed box). TARS encourages the model to attend to causally grounded visual signals rather than spurious correlations, thereby reducing hallucinations.

where φ is constrained to modify only visually agnostic tokens. Specifically, $\Phi(\mathcal{A})$ denotes the set of admissible perturbation functions: $\Phi(\mathcal{A}) = \{\varphi \mid \{i \mid \varphi(q_i) \neq q_i\} \subseteq \mathcal{A}(x, q)\}$, where $\mathcal{A}(x, q)$ is the set of token indices identified as visually agnostic (defined in Eq. (9)). This min–max objective promotes preference alignment under distributional shifts, helping to mitigate spurious correlations and reduce hallucinated outputs.

3.2 Maximizing with Token Perturbations

As shown in Eq. (5), DPO aligns models with preferred responses via log-likelihood ratios against a reference model. However, we observe that this formulation can encourage overfitting to superficial patterns such as frequent phrases and stylistic tokens, which in turn reduce effective alignment with the visual context and hinder robust multimodal understanding [24, 59].

To counter this, we apply token-wise maximization to introduce distribution shifts and reduce overfitting to preference signals. Formally, we define:

$$\varphi(q) = \arg\max_{\varphi \in \Phi(\mathcal{A})} \mathrm{Sim}(\varphi(q), q), \tag{7}$$

where $\Phi(\mathcal{A})$ denotes allowable perturbations constrained to $\mathcal{A}(x, q)$, and $\mathrm{Sim}(\varphi(q), q)$ measures token-level deviation. In practice, we approximate $\varphi^*(q)$ by applying token-level transformations:

$$\varphi(q) = \{\mathbb{I}[i \in \mathcal{A}(x, q)] \cdot \varphi(q_i) + \mathbb{I}[i \notin \mathcal{A}(x, q)] \cdot q_i\}_{i=1}^{|q|}, \tag{8}$$

where $\varphi(q_i)$ denotes the perturbed token, constructed using either masking (replacing with `[MASK]`) or synonym substitution; implementation details are provided in Supplementary Section 7.2. $\mathbb{I}[\cdot]$ is the indicator function. This approximation simulates worst-case alignment uncertainty while preserving semantic integrity.

To preserve semantics, we restrict changes to visual-agnostic tokens with minimal impact on cross-modal alignment. We compute token-level visual relevance using a visual encoder $\mathcal{G}_v(\cdot)$ and a text encoder $\mathcal{G}_t(\cdot)$ (both instantiated by the CLIP encoder [52]) as the cosine similarity between visual features $\mathcal{G}_v(x) \in \mathbb{R}^d$ and each token embedding $\mathcal{G}_t(q_i) \in \mathbb{R}^d$. We then identify a set $\mathcal{A}$ of N_t visually agnostic tokens with the lowest cross-modal alignment scores:

$$\mathcal{A} = \mathrm{Top}_{N_t}\big(-\mathcal{G}_v(x)\mathcal{G}_t(q_i)^\top\big), \quad N_t = \lfloor \omega \cdot \Delta P^{-1} \rfloor + 1. \tag{9}$$

where $\lfloor\cdot\rfloor$ denotes the floor operation and ω is a scaling coefficient controlling perturbation intensity. The matrix $P \in \mathbb{R}^m$ is the negated similarity score vector, with $P_i = -\mathcal{G}_v(x) \cdot \mathcal{G}_t(q_i)^T$. The confidence margin $\Delta P = \max_j P_j - \max_{k \neq j} P_k$ quantifies the predictive uncertainty of the cross-modal alignment: confident predictions (large ΔP) lead to fewer perturbations, while greater uncertainty (small ΔP) induces broader variation. Supplementary Section 9.5 compares this adaptive selection strategy against random and uniform perturbation baselines.

3.3 Spectral Regularization for Token Alignment

Token-level perturbation introduces distribution shifts, yet the supervision from preference pairs (y_w, y_r) is static. This discrepancy may cause the model to learn distribution-specific artifacts under strong alignment constraints [17, 24].

We align in the frequency domain rather than in the spatial domain (*e.g.*, ℓ_2). The key motivation is that a spatial loss $\|z - z'\|_2$ penalizes every positional deviation equally; when specific tokens are perturbed, this rigid constraint reintroduces spurious positional correlations [65, 83]. In contrast, by the shift property of the Discrete Fourier Transform, perturbing a single token z_l affects every frequency bin k by an additive term proportional to $e^{-2\pi ikl/L}$, whose magnitude is bounded regardless of l. The spectral representation thus absorbs local perturbations into a smooth global envelope, preserving dominant low-frequency semantics while tolerating position-specific noise [18].[1]

Concretely, we extract hidden states for $(x, \varphi(q), y_w)$ and contrast them with (x, q, y_w) and (x, q, y_r). Let $z \in \mathbb{R}^{L \times D}$ denote a hidden-state sequence. The spectral representation is:

$$\mathcal{F}(z)_k = \left\| \mathrm{Re}\left[\sum_{l=0}^{L-1} z_l e^{-2\pi ikl/L} \right] \right\|_2, \quad k = 0, \ldots, L-1. \tag{10}$$

where the FFT is applied along the token axis, $\mathrm{Re}[\cdot]$ extracts the real part, and $\|\cdot\|_2$ yields a scalar spectral summary. The spectral preference loss is:

$$\mathcal{L}_{\text{freq}} = -\log \sigma\Big(\beta\big[\log \tfrac{\mathcal{F}(h_\theta(x,\varphi(q),y_w))}{\mathcal{F}(h_{\text{ref}}(x,q,y_w))} - \log \tfrac{\mathcal{F}(h_\theta(x,\varphi(q),y_r))}{\mathcal{F}(h_{\text{ref}}(x,q,y_r))}\big]\Big). \tag{11}$$

[1] Supplementary Section 9.3 provides a formal analysis of spectral alignment under token-level adversarial perturbation.

Here $h_\theta(\cdot)$ and $h_{\text{ref}}(\cdot)$ are hidden states from the policy and reference models, and β is a scaling temperature. This objective extends DPO alignment to the spectral domain, improving frequency-aware consistency and reducing hallucinations from overfitting to fixed preferences.

3.4 Minimization Objective in TARS

We integrate the standard DPO loss with spectral regularization to yield the final TARS training objective. Given a perturbed input $\varphi(q)$ obtained from the inner maximization, and its original counterpart q, the overall loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{TARS}}(x, q, \varphi(q), y_w, y_r) = \mathcal{L}_{\text{DPO}}(x, \varphi(q), y_w, y_r) \\ + \lambda \cdot \mathcal{L}_{\text{freq}}(x, q, \varphi(q), y_w, y_r).\end{aligned} \tag{12}$$

where λ is a weighting coefficient that balances preference alignment and spectral consistency. This joint formulation explicitly encourages the model to preserve causal alignment with preference signals under adversarial perturbation, thereby mitigating spurious correlation. Extended ablation studies on the sensitivity of ω and λ are presented in Supplementary Sections 9.1 and 9.2.

4 Experiments

4.1 Experiment Details

Experiment Setups. We evaluate our approach on the multimodal LLM LLaVA-v1.5 [42] at both 7B and 13B scales, and on Muffin-13B [24, 76] (Supplementary Section 11). All methods are performed with greedy decoding and a temperature of 0.

For fair comparison, we carefully align our training configuration with the most data-efficient preference optimization baselines. Specifically, we randomly sample 4.8k instances from the RLHF-V-Dataset [77], consistent with OPA-DPO [75], and adopt the same training strategy as CHiP-DPO [24]. All models are trained on eight NVIDIA A100 (80GB) GPUs with identical hyperparameters. We set $\alpha = 1$ in Eq. (5) and $\beta = 1$ in Eq. (11) for preference optimization. We implement $\varphi(\cdot)$ using both token masking (Mask) and replacement (Replace) strategies in Eq. (8), and set the perturbation constraint strength to $\omega = 0.1$ in the adversarial min-max formulation Eq. (9). We use a frequency-domain loss weight of $\lambda = 0.1$ in Eq. (12). Full implementation details and ablation studies are reported in the supplementary material.[2]

Evaluation Benchmarks. We evaluate TARS across four established hallucination benchmarks spanning generative and discriminative settings: AMBER [69] for fine-grained generative hallucination (CHAIR [54], Cover, Hal-Rate, Cog), MMHal [63] for VQA hallucination scored by GPT-4V, OBJHal [77] for

[2] See Supplementary Sections 7, 9.1, and 9.2 for configurations and hyperparameter ablations.

Table 1: Comparison across benchmarks. We evaluate SOTA MLLMs as references, denoted by §. For algorithms with available checkpoints, re-tested results are marked with †; for those without, we reproduce results using settings from [24, 38], denoted by ‡. **Bold** denotes the best performance, and underlined denotes the second-best.

Algorithm	AMBER				MMHal		POPE		OBJHal	
	CHAIR↓	Cover↑	Hal-Rate↓	Cog↓	Score↑	Hal-Rate↓	Acc↑	Pre↑	CR_s↓	CR_i↓
Intern-VL2.5-7B [16]§	7.9	54.7	37.1	3.2	3.54	0.26	-	-	36.0	9.1
Qwen-VL2.5-8B [3]§	4.6	54.6	21.1	1.3	3.29	0.27	-	-	40.7	8.6
DeepSeek-VL2-27B [71]§	2.4	56.6	16.3	0.9	2.84	0.27	-	-	10.0	7.0
GPT-4o [31]§	2.5	60.9	17.6	0.8	3.87	0.24	-	-	29.3	6.7
LLaVA-v1.5-7B [42]§	7.6	51.7	35.4	4.2	2.02	0.61	80.0	61.8	54.0	15.8
+ *RLHF* [64]†	8.3	52.2	41.8	4.5	1.93	0.67	82.0	69.3	56.0	15.2
+ *RLAIF* [78]†	3.0	50.3	16.5	1.0	**2.89**	**0.42**	88.1	88.0	13.7	4.2
+ *HALVA* [56]†	6.9	52.8	33.2	3.5	2.12	0.59	87.5	79.6	47.3	14.6
+ *DPO* [38]‡	4.9	56.6	26.4	2.5	2.19	0.61	87.8	82.0	14.0	5.0
+ *CHiP-DPO* [24]‡	2.9	57.3	19.9	1.0	2.32	0.57	81.1	91.8	**7.3**	4.3
+ *OPA-DPO* [75]†	2.7	47.4	**12.5**	0.9	2.78	0.46	87.4	86.2	13.3	4.5
+ TARS (Mask)	2.4	**59.6**	13.2	**0.4**	2.48	0.45	**88.7**	**97.5**	12.0	**3.2**
+ TARS (Replace)	**2.1**	59.3	14.9	0.7	2.54	0.46	87.9	97.0	13.4	3.3
LLaVA-v1.5-13B [42]§	6.7	52.1	32.5	3.5	2.39	0.53	74.6	55.2	50.0	14.5
+ *RLHF* [64]†	7.1	51.4	36.3	3.6	2.10	0.67	83.6	71.2	46.7	11.6
+ *HALVA* [56]†	6.5	53.4	30.1	3.3	2.28	0.56	86.8	75.6	42.7	12.1
+ *DPO* [38]‡	4.1	56.7	24.3	2.2	2.48	0.50	85.2	84.3	19.0	7.2
+ *CHiP-DPO* [24]‡	3.8	58.6	20.8	1.7	2.70	0.46	86.6	74.9	30.0	6.2
+ *OPA-DPO* [75]†	2.8	48.4	13.5	1.0	**3.02**	**0.40**	87.2	80.7	18.3	5.1
+ TARS (Mask)	**2.1**	**59.8**	**12.5**	**0.6**	2.89	0.45	**87.6**	**93.0**	**14.6**	**2.8**
+ TARS (Replace)	**2.1**	59.4	13.6	0.7	2.63	0.47	86.9	92.5	14.9	3.4

captioning hallucination (response-level CR_s and object-level CR_i), and POPE [39] for binary discriminative object hallucination. Detailed benchmark descriptions and evaluation protocols are provided in Supplementary Section 7.3.

Baseline Methods. We compare against two categories:

(1) Advanced multimodal foundation models: Intern-VL2.5-7B [16], Qwen-VL2.5-8B [3], DeepSeek-VL2-27B [71], and GPT-4o [31].

(2) LLaVA-v1.5 with RL techniques: We evaluate multiple RL-based approaches applied to both the 7B and 13B variants of LLaVA-v1.5, including RLHF [64], RLAIF [78], HALVA [56], as well as three state-of-the-art methods based on direct preference optimization (DPO): DPO [51], CHiP-DPO [24], and OPA-DPO [75]. A comparison of algorithmic properties is provided in Tab. 4.

4.2 Evaluation on Hallucination Benchmarks

Tab. 1 presents results across four hallucination benchmarks. We adopt token masking and synonym replacement as perturbation strategies; extended results on Muffin-13B are reported in Supplementary Section 11.

(1) Consistent hallucination mitigation across benchmarks. On the 7B scale, TARS lowers the AMBER Hal-Rate from 35.4% to 13.2% (–22.2 pp) while raising Cover from 51.7% to 59.6% (+7.9 pp) and reducing Cog from 4.2 to 0.4. On OBJHal, CR_s drops sharply from 54.0% to 12.0%.

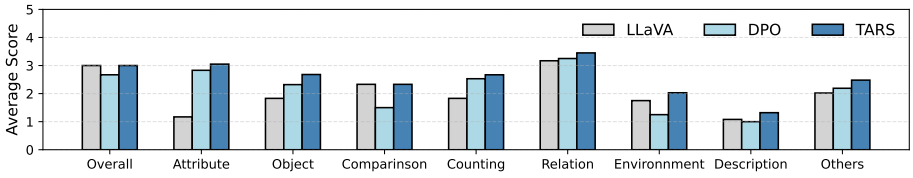


Fig. 4: Comparison of average scores across question categories on MMHal. TARS achieves consistently higher scores, demonstrating stronger visual grounding.

Table 2: Ablation of token-level perturbation (TP), cross-modal alignment score (CAS), and spectral preference alignment (SPA).

Algorithm	AMBER			OBJHal	
	Cover↑	Hal-Rate↓	Cog↓	CR_s↓	CR_i↓
LLaVA-v1.5-7B [42]	51.7	35.4	4.2	54.0	15.8
TARS	**59.6**	**13.2**	**0.4**	**12.0**	**3.2**
w/o **TP**	56.6	26.4	2.5	14.0	5.0
w/o **CAS**	55.9	17.7	1.3	12.7	3.5
w/o **SPA**	58.3	15.1	0.7	12.5	3.7
w/o **CAS&SPA**	55.1	18.5	1.5	12.6	3.8

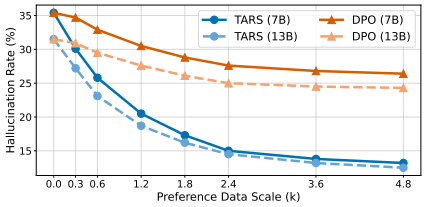


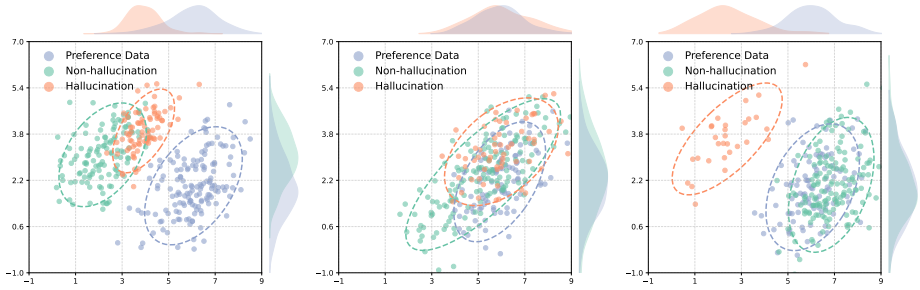
Fig. 5: Comparison of AMBER hallucination rate versus preference data scale.

(2) Strong data efficiency under limited supervision. Using only 4.8k public preference samples without expert feedback (Tab. 4), TARS outperforms OPA-DPO on AMBER-13B in both Cover (+11.4 pp) and Hal-Rate (–1.0 pp).
(3) Scalability across model sizes. From 7B to 13B, CHAIR improves from 2.4 to 2.1, Hal-Rate drops from 13.2% to 12.5%, and Cog stays below 0.7. TARS-13B consistently surpasses all 13B baselines by 1.0–1.5 pp in Hal-Rate.
(4) Competitiveness with larger proprietary models. At 13B, TARS approaches GPT-4o in Cover (59.8% vs. 60.9%) while achieving a notably lower Hal-Rate (12.5% vs. 17.6%), and remains competitive with DeepSeek-VL2-27B in both metrics despite its smaller scale. POPE accuracy reaches 88.7% (+8.7 pp over LLaVA-7B), confirming preserved factual grounding.[3]

4.3 Ablation Analyses on Component

We analyze the key TARS components through ablations in Tab. 2, focusing on three elements:
(1) Token-level perturbation (TP) in Eq. (6), which introduces distributional shifts and proves essential for revealing token-level vulnerabilities and improving robustness. Removing it substantially increases Cog from 0.4 to 2.5, highlighting its critical role in capturing fine-grained alignment uncertainty.
(2) Cross-modal alignment score (CAS) in Eq. (9), which targets visually agnostic tokens to preserve semantic fidelity. Its absence leads to a 4.5-point increase in hallucination and a 0.9 rise in Cog, indicating weaker suppression of spurious correlations and reduced visual grounding. Supplementary Section 9.5

[3] Fine-grained AMBER and multimodal-understanding breakdowns are provided in Supplementary Sections 10.1 and 10.2.



(a) LLaVA (b) DPO (c) TARS

Fig. 6: Distribution of hidden representations across preference-aligned, non-hallucinated, and hallucinated responses of different MLLMs. Top and right margins show marginal distributions along key feature dimensions. We extract representations from 100 preference training instances and 200 AMBER inputs across text and vision modalities. Responses to AMBER inputs are categorized as non-hallucinated or hallucinated based on factual coherence. TARS aligns with preference data while avoiding overfitting to spurious correlations, demonstrating superior factual fidelity.

directly compares adaptive and random/uniform token selection, confirming the necessity of the proposed mechanism.

(3) Spectral preference alignment (SPA) in Eq. (11), which regularizes frequency-aware consistency across token representations. Removing it increases the hallucination rate by 1.9 points and CR_i from 3.2 to 3.7, suggesting degraded fine-grained factual grounding and less stable preference alignment. Supplementary Section 9.4 compares spectral alignment against spatial alternatives (ℓ_2, cosine).

4.4 Ablation Analyses on Preference Scale Impact

We examine how preference data scale influences alignment efficiency (Fig. 5). TARS consistently outperforms DPO across all scales and exhibits sharper gains in the low-data regime: from 0 to 1.8k examples, the 7B and 13B variants reduce hallucination rates by over 15 pp, indicating that TARS captures the core alignment signal early. Beyond 3.6k examples, marginal gains saturate while performance remains stable, highlighting strong data efficiency for practical scenarios where preference annotations are costly or scarce. This behavior is consistent with the dual-space regularization discussed in Sec. 3.4: by enforcing alignment in both the output probability space and the hidden representation space, TARS extracts richer supervision from each preference pair, reducing the sample complexity needed to reach competitive performance. In contrast, standard DPO requires substantially more data to compensate for its tendency to overfit to surface-level patterns, as reflected by the persistent hallucination rate gap across all data scales in Fig. 5.

4.5 Stability of Semantic Representations

We analyze how preference optimization reshapes hidden-state distributions in Fig. 6. TARS yields a more structured latent space in which hallucinated and preference-aligned representations are clearly separated, whereas DPO produces entangled clusters that interleave hallucinated and preference features, indicating overfitting to superficial signals. Crucially, TARS selectively aligns non-hallucinated responses with preference features while isolating hallucinated content, creating a semantically faithful representation space that reinforces only factually grounded outputs.

The marginal distributions (top and right panels in Fig. 6) further reveal the mechanism behind this separation. Under DPO, hallucinated and preference-aligned marginals overlap heavily, reflecting a representation space where surface-level patterns dominate over grounded semantics. In contrast, TARS produces near-disjoint marginals for hallucinated content while maintaining tight overlap between non-hallucinated and preference-aligned distributions. This selective clustering indicates that spectral regularization combined with adversarial perturbation shapes a representation geometry that naturally distinguishes factual from hallucinated outputs, without requiring explicit labels during training.

4.6 Training Dynamics Analysis

To validate the min–max formulation (Eq. (6)), we compare training dynamics of standard DPO, random perturbation (RandPert), and TARS (Min-Max) on LLaVA-v1.5-7B over 20 epochs (Fig. 7; all losses normalized to $[0, 1]$). DPO converges to the lowest training loss ($\sim$0.2) but exhibits the worst robust-sample test loss (0.65–0.75), confirming severe overfitting to static preference patterns [51]. RandPert shows intermediate test performance but with heavy oscillation, as indiscriminate perturbation yields inconsistent gradients. TARS maintains a moderately higher training loss (0.35–0.4) yet achieves the *lowest* test loss on both clean ($<$0.3) and robust ($<$0.5) splits with stable convergence. This confirms that the min–max formulation acts as an implicit regularizer: adversarial token-level shifts during training prevent premature convergence and yield alignment features that generalize to unseen and perturbed data.

4.7 Comparison with Data Augmentation

We compare TARS against data augmentation baselines across all four benchmarks (Tab. 3).[4] We consider paraphrasing via back-translation and LLM-based instruction augmentation at 1$\times$ and 5$\times$ expansion rates, all applied on top of standard DPO training. As shown in Tab. 3, augmentation alone improves coverage and moderately reduces hallucination, yet even 5$\times$ LLM augmentation still falls short of TARS in Hal-Rate (16.0% vs. 13.2%) and Cog (1.2 vs. 0.4). The trend is consistent across MMHal and OBJHal: augmentation narrows the gap

[4] See Supplementary Section 10.3 for detailed augmentation protocols and analysis.

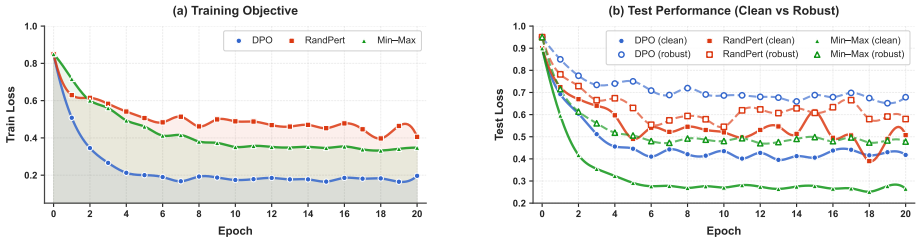


Fig. 7: Training dynamics of DPO, RandPert, and TARS (Min-Max); all losses are normalized to $[0, 1]$. **(a)** Training objective: DPO converges fastest but overfits; Min-Max maintains a higher loss reflecting adversarial difficulty. **(b)** Test performance on clean and robust (perturbed) samples: Min-Max achieves the lowest test loss on both splits, while DPO exhibits a large clean–robust gap indicating poor robustness.

Table 3: Comparison of TARS against data augmentation baselines on four hallucination benchmarks. *Paraphrasing* denotes back-translation-based response diversification; *LLM Aug.* denotes LLM-generated preferred responses, where 1× and 5× indicate the expansion factor relative to the original 4.8k preference set. All methods build on standard DPO. TARS is complementary to augmentation strategies.

Method	**AMBER**				**MMHal**		**POPE**		**OBJHal**	
	CHAIR↓	Cover↑	Hal-Rate↓	Cog↓	Score↑	Hal-Rate↓	Acc↑	Pre↑	CR_s↓	CR_i↓
LLaVA-7B-DPO	4.9	56.6	26.4	2.5	2.19	0.61	87.8	82.0	14.0	5.0
+ Paraphrasing	4.5	57.9	24.8	2.3	2.24	0.58	87.9	83.5	13.6	4.8
+ LLM Aug. (1×)	4.1	58.8	22.9	2.0	2.30	0.55	88.1	85.8	13.2	4.5
+ LLM Aug. (5×)	3.1	59.3	16.0	1.2	2.38	0.50	88.3	90.4	12.6	4.0
+ TARS	2.4	59.6	13.2	0.4	2.48	0.45	88.7	97.5	12.0	3.2
+ TARS & Para.	2.4	60.1	12.8	0.4	2.50	0.45	88.9	97.5	11.6	3.0
+ TARS & Aug. (1×)	2.2	60.5	12.5	0.3	2.53	0.43	88.9	97.7	11.2	2.9
+ TARS & Aug. (5×)	**2.1**	**60.8**	**12.3**	**0.2**	**2.55**	**0.42**	**89.1**	**97.8**	**10.6**	**2.7**

but cannot close it without TARS's token-level robustness mechanism. Combining TARS with augmentation yields the best results on every metric (CHAIR 2.1, Hal-Rate 12.3%, CR_s 10.6%), confirming that the two strategies target complementary error sources.

Discussion. Across all experiments, TARS's improvements manifest systematically across generative (AMBER, OBJHal), discriminative (POPE), and open-ended (MMHal) evaluations. The ablation studies (Tab. 2) confirm that each component contributes independently, the scale analysis (Fig. 5) demonstrates strong data efficiency, and the training dynamics (Fig. 7) validate the regularization effect of the min-max formulation. This convergence of evidence across complementary evaluation axes supports the claim that token-adaptive perturbation with spectral regularization addresses a fundamental limitation of static preference optimization rather than exploiting benchmark-specific biases.

Table 4: Comparison of preference optimization strategies. TARS achieves hallucination mitigation and causal alignment with minimal data and no expert feedback.

Algorithm	Data Size	Feedback	Reward-Free	Hal. Mitigation	Causal Align.
LLaVA-v1.5 [42]	-	-	✗	✗	✗
+ *RLHF* [64]	122k	self-reward	✓	✗	✗
+ *RLAIF* [78]	16k	LLaVA-Next	✗	✓	✗
+ *HALVA* [56]	22k	GPT-4V	✗	✗	✗
+ *DPO* [38]	5k	self-reward	✓	✓	✗
+ *CHiP-DPO* [24]	5k	self-reward	✓	✓	✗
+ *OPA-DPO* [75]	4.8k	GPT-4V	✗	✓	✗
+ ***TARS (Ours)***	4.8k	self-reward	✓	✓	✓

5 Related Work

Multimodal large language models (MLLMs) extend LLMs by integrating visual inputs to support multimodal reasoning [12, 21, 32]. Typically, visual features are extracted by a vision encoder, aligned through a connector, and processed by the LLM [42, 50], with complementary efforts compressing or pruning such models for efficiency [68,84]. Despite strong performance, MLLMs often produce factually incorrect or visually ungrounded outputs, undermining reliability [4, 11]. This issue is more severe than in unimodal LLMs [15, 34], mainly due to modality imbalance [28,45] and ineffective fusion [6,33]. Recent studies attribute these failures to persistent misalignment between multimodal representations and human expectations, rather than model capacity [14, 44, 55].

A key bottleneck in addressing MLLM hallucinations lies in aligning model outputs with human preferences for factual consistency. Unlike knowledge-intensive pretraining [9,46] and instruction tuning [10,42], recent methods typically leverage small-scale human preference data refined via reinforcement learning [8, 22, 77]. Direct preference optimization (DPO) [51, 53] has become a leading approach due to its simplicity and effectiveness, demonstrated in CHiP-DPO [24] and OPA-DPO [75]. However, DPO's reliance on limited data can cause overfitting to superficial linguistic cues [24, 59], leading to distributional rigidity and reduced adaptability to modality-specific semantics [48, 62]. These limitations call for more adaptive alignment strategies that capture token-level variability and cross-modal dependencies for stable multimodal reasoning. Adversarial training has proven effective for improving robustness in vision-language pretraining [25, 82], and injecting visual noise has recently been shown to mitigate object hallucinations [81], yet integrating such perturbation into preference optimization remains largely unexplored.

To address these challenges, we propose a token-adaptive min-max alignment strategy with spectral regularization that enhances preference learning without relying on high-resource expert feedback (*e.g.*, GPT-4V [1]). Our spectral alignment loss operates in the frequency domain via FFT, encouraging global semantic consistency while tolerating the local distributional shifts induced by token perturbation. Using only a small public preference dataset, our method effectively mitigates hallucinations and consistently outperforms RL-based baselines

across benchmarks. Tab. 4 compares preference optimization methods in terms of data scale, supervision, and alignment.

6 Conclusion

This work studies a limitation of static preference optimization: its tendency to latch onto superficial linguistic correlations rather than causally grounded visual evidence. We address this issue by reshaping the optimization landscape rather than only scaling data. TARS recasts preference learning as a min-max game in which the inner adversarial maximization over visual-agnostic tokens encourages the model to separate visual signals from spurious textual cues, while a spectral alignment loss in the frequency domain regularizes global semantic invariance under the resulting distributional shifts. Empirically, using only 4.8k preference samples without expert feedback, TARS halves the hallucination rate ($26.4\% \rightarrow 13.2\%$), reduces cognitive inconsistency from 2.5 to 0.4, and surpasses $5\times$ LLM-based augmentation on 28.8k samples while narrowing the gap with GPT-4o on AMBER metrics at a smaller model scale. These results suggest that adversarial token perturbation can improve data efficiency for robust multimodal preference alignment.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62276221). This paper is supported by Young Scientists Fund of the National Natural Science Foundation of China (NSFC) (No. 62506305), Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (No. 2024R01007), Key Research and Development Program of Zhejiang Province (No. 2025C01026), Scientific Research Project of Westlake University (No. WU2025WF003), Chinese Association for Artificial Intelligence (CAAI) & Ant Group Research Fund - AGI Track (No. 2025CAAI-ANT-13). It is also supported by the research funds of the National Talent Program and Hangzhou Municipal Talent Program.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Achiam, J., Held, D., Tamar, A., Abbeel, P.: Constrained policy optimization. In: ICML (2017)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)

5. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In: NeurIPS (2022)
6. Bellagente, M., Brack, M., Teufel, H., Friedrich, F., Deiseroth, B., Eichenberg, C., Dai, A.M., Baldock, R., Nanda, S., Oostermeijer, K., et al.: Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation. In: NeurIPS (2023)
7. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
8. Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al.: Open problems and fundamental limitations of reinforcement learning from human feedback. TMLR (2023)
9. Chang, H., Park, J., Ye, S., Yang, S., Seo, Y., Chang, D.S., Seo, M.: How do large language models acquire factual knowledge during pretraining? In: NeurIPS (2024)
10. Chen, C., Zhu, J., Luo, X., Shen, H., Song, J., Gao, L.: Coin: A benchmark of continual instruction tuning for multimodel large language models. In: NeurIPS (2024)
11. Chen, C., Liu, M., Jing, C., Zhou, Y., Rao, F., Chen, H., Zhang, B., Shen, C.: Perturbollava: Reducing multimodal hallucinations with perturbative visual training. In: ICLR (2025)
12. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: ICML (2024)
13. Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., Liu, Z.: Large language models are visual reasoning coordinators. In: NeurIPS (2023)
14. Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H.: Unified hallucination detection for multimodal large language models. In: ACL (2024)
15. Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D., Chai, J.: Multi-object hallucination in vision language models. In: NeurIPS (2024)
16. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
17. Chowdhury, S.R., Kini, A., Natarajan, N.: Provably robust dpo: Aligning language models with noisy feedback. In: ICML (2024)
18. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of computation **19**(90), 297–301 (1965)
19. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
20. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., et al.: An empirical study of training end-to-end vision-and-language transformers. In: CVPR (2022)
21. Feng, S., Fang, G., Ma, X., Wang, X.: Efficient reasoning models: A survey. TMLR (2025)
22. Feng, S., Tuo, K., Wang, S., Kong, L., Zhu, J., Wang, H.: Rewardmap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. In: ICLR (2026)

23. Feng, S., Wang, S., Ouyang, S., Kong, L., Song, Z., Zhu, J., Wang, H., Wang, X.: Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. In: CVPR (2026)
24. Fu, J., Huangfu, S., Fei, H., Shen, X., Hooi, B., Qiu, X., Ng, S.K.: Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In: ICLR (2025)
25. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. In: NeurIPS (2020)
26. Gandhi, K., Fränken, J.P., Gerstenberg, T., Goodman, N.: Understanding social reasoning in language models with language models. In: NeurIPS (2023)
27. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: AAAI (2024)
28. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: CVPR (2024)
29. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR (2024)
30. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al.: Language is not all you need: Aligning perception with language models. In: NeurIPS (2023)
31. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
32. Jain, J., Yang, J., Shi, H.: Vcoder: Versatile vision encoders for multimodal large language models. In: CVPR (2024)
33. Ji, Y., Wang, J., Gong, Y., Zhang, L., Zhu, Y., Wang, H., Zhang, J., Sakai, T., Yang, Y.: Map: Multimodal uncertainty-aware vision-language pre-training model. In: CVPR (2023)
34. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: CVPR (2024)
35. Jiang, Y., Yan, X., Ji, G.P., Fu, K., Sun, M., Xiong, H., Fan, D.P., Khan, F.S.: Effectiveness assessment of recent large vision-language models. Visual Intelligence **2**(1), 17 (2024)
36. Jiang, Y., Sun, K., Sourati, Z., Ahrabian, K., Ma, K., Ilievski, F., Pujara, J., et al.: Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. In: NeurIPS (2024)
37. Kim, J., Kim, H., Yeonju, K., Ro, Y.M.: Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In: NeurIPS (2024)
38. Li, S., Lin, R., Pei, S.: Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In: ACL (2024)
39. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023)
40. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
41. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
42. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
43. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In: ECCV (2024)

44. Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., JianHao, Z., Zhang, C., Zheng, X., Huang, X.J.: Aligning large language models with human preferences through representation engineering. In: ACL (2024)
45. Ma, F., Xue, H., Zhou, Y., Wang, G., Rao, F., Yan, S., Zhang, Y., Wu, S., Shou, M.Z., Sun, X.: Visual perception by large language model's weights. In: NeurIPS (2024)
46. McKinzie, B., Gan, Z., Fauconnier, J.P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Belyi, A., et al.: Mm1: methods, analysis and insights from multimodal llm pre-training. In: ECCV (2024)
47. Oh, J., Kim, S., Seo, J., Wang, J., Xu, R., Xie, X., Whang, S.: Erbench: An entity-relationship based automatically verifiable hallucination benchmark for large language models. In: NeurIPS (2024)
48. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: ECCV (2024)
49. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
50. Parekh, J., Khayatan, P., Shukor, M., Newson, A., Cord, M.: A concept-based explainability framework for large multimodal models. In: NeurIPS (2024)
51. Pi, R., Han, T., Xiong, W., Zhang, J., Liu, R., Pan, R., Zhang, T.: Strengthening multimodal large language model with bootstrapped preference optimization. In: ECCV (2024)
52. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
53. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS (2023)
54. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018)
55. Ruan, H., Lin, J., Lai, Y., Luo, Z., Li, S.: Hccm: Hierarchical cross-granularity contrastive and matching learning for natural language-guided drones. In: ACM MM (2025)
56. Sarkar, P., Ebrahimi, S., Etemad, A., Beirami, A., Arık, S.Ö., Pfister, T.: Data-augmented phrase-level alignment for mitigating object hallucination. In: ICLR (2025)
57. Sarkar, P., Ebrahimi, S., Etemad, A., Beirami, A., Arik, S.O., Pfister, T.: Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In: ICLR (2025)
58. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
59. Setlur, A., Garg, S., Geng, X., Garg, N., Smith, V., Kumar, A.: Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. In: NeurIPS (2024)
60. Shao, K., Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Holitom: Holistic token merging for fast video large language models. arXiv preprint arXiv:2505.21334 (2025)
61. Sharma, A., Keh, S.S., Mitchell, E., Finn, C., Arora, K., Kollar, T.: A critical evaluation of ai feedback for aligning large language models. In: NeurIPS (2024)
62. Song, Y., Swamy, G., Singh, A., Bagnell, J., Sun, W.: The importance of online data: Understanding preference fine-tuning via coverage. In: NeurIPS (2024)

63. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525 (2023)
64. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings: ACL (2024)
65. Tian, X., Zou, S., Yang, Z., He, M., Zhang, J.: Black sheep in the herd: Playing with spuriously correlated attributes for vision-language recognition. In: ICLR (2025)
66. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: NeurIPS (2024)
67. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mdpo: Conditional preference optimization for multimodal large language models. In: EMNLP (2024)
68. Wang, H., Jin, X., Lu, L., Li, C., Chen, J., Liu, Q., Wang, H.: Earlytom: Early token compression completes fast video understanding. In: CVPR (2026)
69. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023)
70. Wang, S., Lin, J., Guo, X., Shun, J., Li, J., Zhu, Y.: Reasoning of large language models over knowledge graphs with super-relations. In: ICLR (2025)
71. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
72. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: Findings: EMNLP (2024)
73. Yang, W., Zhong, R., Chen, Y., Lu, C., Jiang, P.: Structured spectral reasoning for frequency-adaptive multimodal recommendation. In: NeurIPS (2025)
74. Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: CVPR (2021)
75. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: CVPR (2025)
76. Yu, T., Hu, J., Yao, Y., Zhang, H., Zhao, Y., Wang, C., Wang, S., Pan, Y., Xue, J., Li, D., et al.: Reformulating vision-language foundation models and datasets towards universal multimodal assistants. arXiv preprint arXiv:2310.00653 (2023)
77. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: CVPR (2024)
78. Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.S., et al.: Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. In: CVPR (2025)
79. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE TPAMI **46**(8), 5625–5644 (2024)
80. Zhang, K.J., Ding, X., Xiang, B.B., Zhang, H.F., Bao, Z.K.: Extracting higher order topological semantic via motif-based deep graph neural networks. IEEE Transactions on Computational Social Systems **11**(4), 5444–5453 (2024)
81. Zhang, K., Tao, K., Tang, J., Wang, H.: Poison as cure: Visual noise for mitigating object hallucinations in lvms. In: NeurIPS (2025)
82. Zhang, K., Weng, J., Li, S., Luo, Z.: Towards adversarial robustness via debiased high-confidence logit alignment. In: ICCV (2025)

83. Zhou, Y., Xu, P., Liu, X., An, B., Ai, W., Huang, F.: Explore spurious correlations at the concept level in language models for text classification. In: ACL (2024)
84. Zhu, J., Wang, H., Su, M., Wang, Z., Wang, H.: Obs-diff: Accurate pruning for diffusion models in one-shot. In: ICLR (2026)