Reducing Hallucination in Vision-Language Models via Stage-wise Preference Optimization under Distribution Shift

Anonymous ACL submission

Abstract

Hallucination remains a fundamental challenge in vision-language models (VLMs), where autoregressive generation may produce linguistically plausible yet visually ungrounded or physically inconsistent responses due to likelihood maximization under joint probabilistic modeling.

We propose a stage-wise preference optimization framework for hallucination reduction through targeted multimodal preference construction. Our approach first establishes grounded perceptual alignment through supervised fine-tuning (SFT), and then progressively constructs hallucination-focused preference pairs near known failure boundaries under controlled distribution shift. The framework emphasizes ambiguous spatial reasoning, object relationships, OCR uncertainty, and adversarial false-premise reasoning. Hallucinated negatives are generated through minimally perturbed yet visually inconsistent alternatives, enabling Direct Preference Optimization (DPO) to better separate grounded reasoning from plausible hallucination.

Experiments on open-source benchmarks and real-world multimodal scenarios demonstrate improved grounding consistency, reduced hallucination, and more informative grounded responses. Cross-model qualitative analysis further shows that the proposed framework produces more visually grounded responses than several frontier proprietary VLMs in ambiguous reasoning settings.

Overall, the results suggest that hallucination arises not only from limited model capacity, but also from autoregressive generation favoring linguistically plausible continuations under weak visual grounding. Future work may explore physical consistency modeling, uncertainty-aware multimodal reasoning, and alternatives beyond standard autoregressive decoding.

1 Introduction

1.1 Theoretical Perspective on Hallucination in VLMs

VLMs aim to approximate the conditional distribution

$$p_\theta(y \mid x_v, x_t),$$

where x_v denotes the visual input, x_t the textual input, and $y = (y_1, \ldots, y_T)$ the output sequence. Training is typically performed using the autoregressive cross-entropy objective:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid x_v, x_t, y_{<t}).$$

Using the chain rule of probability, the autoregressive formulation and its log format can be factorized as:

$$p_\theta(y \mid x_v, x_t) = \prod_{t=1}^{T} p_\theta(y_t \mid x_v, x_t, y_{<t}),$$

$$\log p_\theta(y \mid x_v, x_t) = \sum_{t=1}^{T} \log p_\theta(y_t \mid x_v, x_t, y_{<t}).$$

Therefore, minimizing cross-entropy is equivalent to maximizing the likelihood of the target sequence. Although highly effective for multimodal generation, this objective does not explicitly enforce grounding in visual evidence. Instead, it encourages outputs that are statistically consistent with the joint image–text distribution (Rohrbach et al., 2018; Bender et al., 2021). Consequently, VLMs may assign high probability to responses that are linguistically plausible but visually unsupported:

$$p_\theta(y \mid x_v, x_t) \gg 0 \quad \text{while} \quad y \notin \mathcal{Y}(x_v),$$

where $\mathcal{Y}(x_v)$ denotes the set of visually faithful responses.

This issue becomes more severe under weak, ambiguous, or partially observed visual evidence,

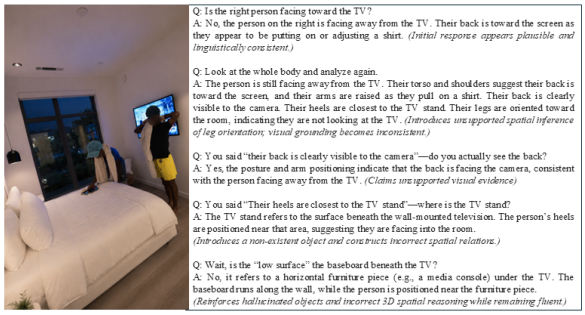


Figure 1: A representative multimodal reasoning failure under iterative questioning (photo generated by AI

where generation increasingly relies on learned language priors rather than grounded perception. Due to autoregressive conditioning, early deviations from grounded content can propagate through subsequent token predictions, producing globally coherent yet visually inconsistent responses.

Figure 1 illustrates several representative failure modes of autoregressive multimodal LLMs. Although the model initially produces a plausible response, subsequent reasoning steps progressively drift away from grounded visual evidence, introducing unsupported spatial assumptions, nonexistent objects, and physically inconsistent interpretations while maintaining fluent natural language generation. The example also reveals weaknesses in physical and spatial reasoning, and instability across multi-turn reasoning despite operating on the same visual input. These behaviors suggest that hallucination in multimodal systems reflects not only factual errors, but also broader limitations in grounding consistency and autoregressive probabilistic generation under ambiguous visual evidence.

Decoding strategies further influence hallucination behavior. Methods such as top-k and nucleus (top-p) sampling introduce stochasticity into generation (Holtzman et al., 2020), while temperature scaling reshapes the output distribution. Higher diversity settings can improve fluency and descriptiveness, but also increase the likelihood of selecting tokens that deviate from grounded visual evidence.

Notably, stronger reasoning capability does not necessarily eliminate hallucination. Techniques such as chain-of-thought prompting improve logical coherence and internal consistency (Wei et al., 2022), but may still extrapolate beyond observable evidence. As a result, hallucinations can become more convincing and harder to detect because the generated reasoning remains linguistically coherent despite lacking factual grounding (Dziri et al., 2023; Mündler et al., 2023).

These observations suggest that standard likelihood-based objectives primarily optimize distributional alignment rather than faithful cross-modal grounding. Addressing hallucination therefore requires supervision signals that explicitly distinguish grounded responses from plausible but unsupported alternatives. And thus, we explore this direction through stage-wise preference optimization and hallucination-targeted data construction.

1.2 Related Work

Hallucination in LLMs and VLMs has been widely studied, with mitigation strategies spanning prompting, supervised fine-tuning, preference optimization, retrieval, and inference-time reasoning. Prompt-based methods encourage models to avoid unsupported claims and rely on observable evidence, while reasoning methods such as chain-of-thought, self-consistency, self-refinement, and multi-agent verification improve logical coherence at inference time (Wei et al., 2022; Wang et al., 2022; Madaan et al., 2023; Du et al., 2024). However, these approaches do not fundamentally modify the underlying model distribution and may still produce visually unsupported responses under ambiguity or distribution shift.

Training-time methods offer a more direct way to reshape model behavior. SFT improves factuality by aligning models to curated grounded responses (Xu et al., 2026a), but the likelihood objective does not explicitly penalize plausible hallucinated alternatives (Ouyang et al., 2022). Preference-based methods such as RLHF and DPO introduce comparative supervision, encouraging preferred responses over rejected ones (Christiano et al., 2017; Rafailov et al., 2023). In particular, DPO provides a simple and stable pairwise optimization signal that can separate grounded responses from hallucinated alternatives when the preference data captures visual faithfulness. GRPO extends preference learning to group-wise relative comparisons, but requires diverse candidate responses and reliable preference estimation (Shao et al., 2024).

Retrieval-augmented generation improves factual grounding by conditioning generation on external evidence (Lewis et al., 2020; Shuster et al., 2021), but retrieval quality and system complexity can limit its effectiveness, and retrieval does not directly address hallucinations caused by perceptual

or visual reasoning failures. Table 1 summarizes the strengths and limitations of representative hallucination mitigation approaches.

Table 1: Summary of hallucination mitigation methods

Method	Core Mechanism	Strengths	Limitations
Prompting	Instruction guidance	Simple, low cost	Limited behavioral control
SFT	Likelihood training	Stable, simple	No explicit negative signal
RLHF/DPO	Preference optimization	Direct alignment	Data-quality dependent
GRPO	Group-wise preference	Richer supervision	Requires diverse candidates
RAG	External evidence	Strong factuality	Retrieval-dependent
CoT/Multi-agent	Inference-time reasoning	Improves coherence	Costly; not sufficient for grounding

In contrast to these approaches, our work focuses on data-centric hallucination modeling through progressive multimodal preference construction near hallucination-prone reasoning boundaries.

1.3 Positioning of Our Approach

1.4 Positioning of Our Approach

Existing hallucination mitigation approaches primarily focus on modifying optimization objectives, retrieval augmentation, or inference-time reasoning. While preference optimization methods such as DPO have been widely adopted for general alignment, our work focuses on how hallucination-oriented preference data is constructed from grounded SFT data and progressively introduced during training.

Starting from grounded SFT data emphasizing simpler factual and perceptual understanding, we construct more challenging preference-oriented samples through data-centric augmentation under controlled distribution shift. The resulting DPO data includes detailed grounded responses paired with minimally perturbed hallucinated alternatives that remain linguistically plausible while introducing subtle visually unsupported attributes, relations, or reasoning. These preference pairs are intentionally concentrated near hallucination-prone decision boundaries, producing more informative supervision than trivial positive–negative contrasts.

Unlike inference-time approaches, the proposed framework operates during training and progressively refines grounding robustness through stage-wise SFT-to-DPO preference optimization. Overall, our contribution lies not in DPO itself, but in structured multimodal preference construction and progressive alignment for hallucination reduction under challenging multimodal reasoning conditions.

2 Methodology

2.1 Model Architecture

We instantiate our framework on a large-scale VLM consisting of a visual encoder, a Perceiver-based cross-modal alignment module, and a LLaMA-3 70B language decoder. Rather than modifying the underlying architecture, our approach focuses on improving hallucination robustness through stage-wise preference optimization and hallucination-targeted data construction. Specifically, the model is first trained on large-scale supervised grounding data and subsequently refined on more challenging hallucination-prone scenarios under distribution shift, as illustrated in Figure 2.

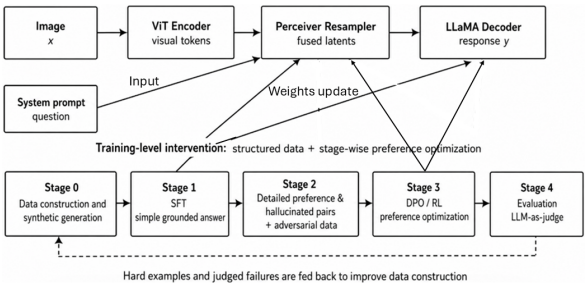


Figure 2: Overview of model and training pipeline. A ViT extracts visual features, which are compressed via a Perceiver resampler and fused with text inputs for LLaMA decoding. The proposed framework improves grounding through stage-wise preference optimization and structured hallucination-targeted data construction.

2.2 Problem Setup

Let $x = (x_v, x_t)$ denote a multimodal input consisting of an image x_v and a textual query x_t, and let y denote the generated response. We consider two related training distributions:

- $\mathcal{D}_1$: a large-scale multimodal dataset emphasizing basic grounding tasks, including object recognition, OCR, and standard VQA;
- $\mathcal{D}_2$: a substantially smaller but more challenging distribution focused on spatial reasoning, compositional understanding, and hallucination-prone scenarios.

The second-stage distribution $\mathcal{D}_2$ is derived from $\mathcal{D}_1$ through a data-centric augmentation pipeline that transforms grounded examples into more challenging reasoning-oriented training samples. These include long-form responses, adversarial false-premise examples, and grounded-versus-hallucinated preference pairs. Compared with $\mathcal{D}_1$,

$\mathcal{D}_2$ is intentionally smaller and more targeted, reflecting its role in hallucination-focused refinement rather than broad capability acquisition. Each sample in $\mathcal{D}_2$ contains a preference pair (y^+, y^-), where $y^+ \succ y^-$. Our objective is to learn a policy $p_\theta(y \mid x)$ that improves hallucination robustness while preserving strong multimodal grounding performance.

2.3 Preferred-Conditioned Data Augmentation

A key component of our framework is the construction of $\mathcal{D}_2$. Rather than sampling it independently, we generate it through structured augmentation of grounded multimodal examples from $\mathcal{D}_1$. In addition to long-form response enrichment and adversarial false-premise generation, we also construct hard preference pairs in which negative responses remain linguistically plausible and semantically close to the preferred responses. Many samples involve ambiguous spatial relationships, partial visual evidence, or subtle grounding errors that are difficult to distinguish. This design encourages the preference optimization stage to focus on fine-grained hallucination boundaries rather than trivial negative cases.

Given $x \sim \mathcal{D}_1$ with grounded response y^+, we generate a perturbed sample: $x' \sim \mathcal{A}(x; y^+)$,where $\mathcal{A}$ denotes a structured augmentation operator that increases reasoning difficulty while preserving semantic consistency. The augmentation process includes, but is not limited to: i) introducing spatial reasoning challenges (e.g., relative positioning and occlusion-aware counting), ii) adding compositional and contextual constraints, and iii) constructing ambiguity-aware and hallucination-prone query conditions. We then construct preference pairs (y'^+, y'^-) for x', where negative responses are minimally perturbed variants of grounded answers that introduce subtle hallucinations while remaining linguistically plausible. This induces a controlled distribution shift of $\mathcal{D}_1 \rightarrow \mathcal{D}_2$, with $\mathcal{D}_2$ concentrated near decision boundaries between grounded and hallucinated outputs.

Figure 3 illustrates representative examples from the proposed stage-wise data curation pipeline. Starting from grounded SFT data $\mathcal{D}_1$, selected samples are progressively transformed through data-centric augmentation into a more targeted preference-oriented distribution $\mathcal{D}_2$, introducing increasingly challenging hallucination-prone scenarios such as spatial reasoning, contextual understanding, and fine-grained attribute recognition.

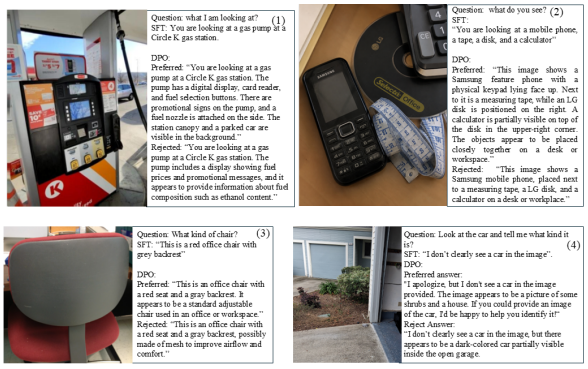


Figure 3: Stage-wise data curation pipeline. In the first stage (SFT), training emphasizes concise and visually grounded responses. In the second stage (DPO), detailed preferred responses are paired with hallucinated alternatives to construct challenging reasoning scenarios, including spatial understanding, contextual inference, and fine-grained attribute recognition (2nd photo was AI generated; all others were captured by the author).

2.3.1 Theoretical Interpretation

More generally, the optimization objective can be conceptually interpreted as :

$$\max_{\theta} \; E_{\mathcal{D}_2}[\mathrm{pref}_\theta] \quad \text{s.t.} \quad \mathrm{KL}(p_\theta \,\|\, p_{\theta_1}) \leq \epsilon,$$

where pref_θ denotes the preference score assigned to grounded responses over hallucinated alternatives, and $\mathrm{KL}(p_\theta \,\|\, p_{\theta_1})$ constrains the updated policy p_θ to remain close to the Stage-1 SFT policy p_{θ_1}. Intuitively, the objective encourages improved hallucination discrimination while avoiding large deviations from the original grounded behavior learned during supervised fine-tuning.

Thus, Stage 1 acts as an implicit regularizer, constraining updates to remain within a parameter region that preserves fundamental grounding behavior.

From a complementary perspective, multimodal reasoning can be viewed as involving progressively more difficult latent factors $z = (z_{\text{easy}}, z_{\text{hard}})$, where z_{easy} corresponds to basic perception and grounding (e.g., object recognition and OCR), and z_{hard} represents higher-order reasoning factors such as spatial understanding, ambiguity resolution, and hallucination-sensitive reasoning. Under this view, the conditional generation process can be conceptually expressed as

$$p(y \mid x, z_{\text{easy}}, z_{\text{hard}}).$$

Our stage-wise optimization approximates a progressive alignment process:

$$p_{\theta_1}(y \mid x, z_{\text{easy}}) \;\rightarrow\; p_{\theta_2}(y \mid x, z_{\text{easy}}, z_{\text{hard}}),$$

where Stage 1 first establishes stable multimodal grounding on simpler distributions, while Stage 2 further adapts the model to harder hallucination-prone reasoning scenarios. This staged decomposition simplifies optimization and improves robustness under distribution shift by avoiding direct optimization over highly complex reasoning behaviors from the beginning of training.

2.4 Auxiliary Mechanisms to reduce Hallucination

In addition to stage-wise preference optimization, we incorporate two complementary mechanisms that further improve grounding: visual prompting and adversarial false-premise training. These mechanisms operate at different levels—conditioning and data construction—but share the goal of aligning generation with observable visual evidence.

2.4.1 Visual Prompt with Few-Shot Guidance

We employ a visual prompting strategy as an initial guard layer that encourages structured examination of the image prior to answer generation. The system prompt explicitly instructs the model to attend to visual evidence and avoid unsupported guessing, while few-shot examples demonstrate grounded reasoning patterns for challenging multimodal tasks such as counting, spatial understanding, and object relationships.

Formally, the prompting strategy can be viewed as introducing an intermediate latent reasoning trajectory $r = (r_1, \ldots, r_K)$, representing a sequence of implicit grounding and reasoning steps before answer generation:

$$p_\theta(y \mid x) = \sum_r p_\theta(y \mid x, r)\, p_\theta(r \mid x).$$

Here, $p_\theta(r \mid x)$ models the probability of selecting a particular reasoning trajectory r conditioned on the multimodal input x, while $p_\theta(y \mid x, r)$ denotes answer generation conditioned on that reasoning process. For example, in an object-counting task, a reasoning trajectory may involve sequential grounding steps such as identifying visible objects in different image regions, checking partially occluded areas, and progressively aggregating counts before producing the final answer. The prompting strategy reshapes the distribution $p_\theta(r \mid x)$ toward more visually grounded reasoning trajectories, thereby reducing hallucination-prone generation.

From a probabilistic perspective, the prompting mechanism reshapes the distribution $p_\theta(r \mid x)$, suppressing reasoning trajectories that omit important visual evidence and thereby reducing hallucination. However, as an inference-time guidance method, its effectiveness is inherently limited by the coverage and diversity of the few-shot examples, making it difficult to generalize across broad multimodal reasoning scenarios.

2.4.2 Adversarial False-Premise Training

We introduce an adversarial training scheme targeting hallucinations induced by incorrect or misleading queries. Specifically, we construct inputs x' where the question contains a false premise (e.g., referencing an object absent from the image). For each such input, we define: 1) a preferred response y^+ that correctly rejects or questions the premise, and 2) a negative response y^- that complies with the premise and hallucinates content.

Augmented examples are incorporated into the preference dataset $\mathcal{D}_2$ for hallucination-targeted DPO training. This directly penalizes the model's tendency to generate plausible but visually unsupported completions, a known failure mode of autoregressive models. By training on such adversarial cases, the model learns to better distinguish factual validity from linguistic plausibility, improving robustness under ambiguous or misleading conditions.

Figure 3, case (4), presents an example of adversarial false-premise training data, where the question assumes the presence of an object (a car) that does not actually exist in the image, despite the scene context (a garage environment with dark-colored structures near the entrance) making such an assumption appear visually plausible.

2.4.3 Integration with Stage-wise Optimization

These mechanisms integrate naturally with the stage-wise framework.

- **Visual prompting** operates at inference time, improving conditioning by structuring intermediate reasoning.

- **Adversarial training** operates at training time, enriching $\mathcal{D}_2$ with targeted hallucination

cases.

Together with preferred-conditioned augmentation, they expose the model to diverse failure modes, including: 1) weak or distant visual evidence, 2) compositional reasoning challenges, and 3) false or misleading premises.

This combination improves alignment between generated outputs and observable inputs, leading to consistent reductions in hallucination without requiring additional inference-time computation beyond prompting.

To improve response completeness while preserving grounding, we additionally explore response-length augmentation and iterative expansion strategies that encourage richer yet visually supported generations. Additional formulation details and analysis of the detail–hallucination trade-off are provided in **Appendix A**.

2.5 Practical Implications

The proposed framework yields several practical benefits: 1) *Improved optimization stability*: initialization from θ_1 stabilizes preference optimization on more challenging hallucination-targeted data distributions; 2) *Better distributional coverage*: large-scale Stage-1 training provides broad multimodal grounding and linguistic coverage prior to targeted refinement; 3) *Targeted robustness*: preference-oriented augmentation concentrates learning on difficult reasoning boundaries involving ambiguity, spatial understanding, and hallucination-prone conditions; and 4) *Improved detail–grounding balance*: the preference construction encourages richer responses while discouraging unsupported elaborations.

Together, these components form a progressive data-centric refinement pipeline that adapts multimodal models to increasingly complex reasoning scenarios while improving hallucination robustness without modifying the underlying model architecture.

3 Experimental Results and Analysis

3.1 Experimental Design

We evaluate our approach using a combination of open-source benchmarks and curated evaluation protocols, comparing model behavior before and after hallucination-targeted training. The primary evaluation tasks include visual question answering (VQA) (Antol et al., 2015), document visual question answering (DocVQA) (Mathew et al., 2021), and multimodal reasoning benchmarks such as MMBench (Liu et al., 2023).

To capture different aspects of model quality, we employ an LLM-as-a-judge framework for pairwise evaluation. In this setting, responses from two models are compared on the same input, and a judge model selects the preferred answer. Prior work has shown that pairwise evaluation can reduce common scalar-scoring biases of pointwise method, particularly verbosity bias, where longer responses are systematically favored regardless of factual correctness or grounding (Xu et al., 2026b; Zheng et al., 2023).

In addition to quantitative evaluation, we conduct qualitative analysis to examine improvements in visual grounding, multimodal reasoning, hallucination robustness, and response informativeness. We further evaluate performance on long-tail and hallucination-prone scenarios, comparing our approach against strong baselines including LLaMA3 multimodal (MM) DPO (ours), Gemini Flash, and GPT-series models.

It is worth noting that we also include a small subset of AI generated stylized images, such as watercolor-like renderings generated by Nano Banana, as a stylized-image transformation for visual question answering (VQA). These transformations obscure Personally Identifiable Information (PII) while preserving key spatial and semantic content, consistent with prior work onidentity-obscured VQA and vision-language modeling (Bara et al., 2022). Additionally, such stylization introduces mild distribution shifts, allowing us to probe whether models rely on superficial visual cues or more robust semantic understanding, as explored in prior stylized-image visual tasks (Geirhos et al., 2019).

3.2 Quantitative Results

Performance on Open-Source Benchmarks We first evaluate the Stage-1 SFT checkpoint on several open-source benchmarks to verify that the model maintains strong general multimodal capability prior to hallucination-targeted refinement. As shown in Table 2, the model achieves competitive performance across standard visual understanding tasks, including DocVQA, VQA, and MMBench.

For the Stage-2 DPO-trained model with the new stage-wise and hallucination-targeted refinement, we further conduct pairwise evaluation on approximately 839 diverse multimodal evaluation scenarios, including indoor scenes, outdoor envi-

Table 2: Performance comparison on standard multimodal benchmarks.

Model	DocVQA	VQA	MMBench
LLaMA-3 MM SFT	82.30	81.45	78.99
LLaMA-3 MM DPO	83.00	82.18	80.14

Table 3: Comparison of model performance across evaluation metrics.

Metric	GPT-4o	Gemini-2.5-Flash	LLaMA-4-Marvrick
LlaMA-3 MM SFT Wins	207 (24.7%)	256 (30.5%)	265 (31.6%)
LlaMA-3 MM DPO Wins	276 (32.9%)	316 (37.7%)	317 (37.8%)
Equivalent (TIE/C)	294 (35.0%)	221 (26.3%)	220 (26.2%)
Both Wrong (D)	49 (5.8%)	20 (2.4%)	19 (2.3%)
Error Rate	13 (1.5%)	26 (3.1%)	18 (2.1%)
Δ Win Rate (%)	+8.2%	+7.2%	+6.2%

ronments, food, landscapes, and plants, etc. In this evaluation, responses from two models are compared on the same input and judged by an LLM-based evaluator. Results reveal substantial improvements in grounding quality and response informativeness. As shown in Table 3, the DPO-trained model achieves consistent gains across multiple judge models, with win rates increasing by $+6.2\%$ to $+8.2\%$ over the baseline.

Interestingly, relatively small hallucination-targeted preference datasets (approximately 50k–120k samples) already produce substantial grounding improvements, suggesting that preference structure and data quality may play an important role during alignment refinement. Additional discussion is provided in Appendix C.1.

3.3 Qualitative Analysis

3.3.1 Context Richness and Hallucination Reduction

Figure 4 presents representative comparisons between SFT and DPO model outputs across diverse multimodal scenarios. Compared with the SFT baseline, the DPO-trained model generally produces richer and more contextually useful responses while maintaining stronger visual grounding. In the shopping example (Example 2), for instance, the model not only identifies visible storefront elements but also provides more useful contextual interpretation aligned with the query intent. Similar trends are observed across multiple examples, where preference optimization improves response informativeness and contextual completeness without substantially increasing unsupported details.

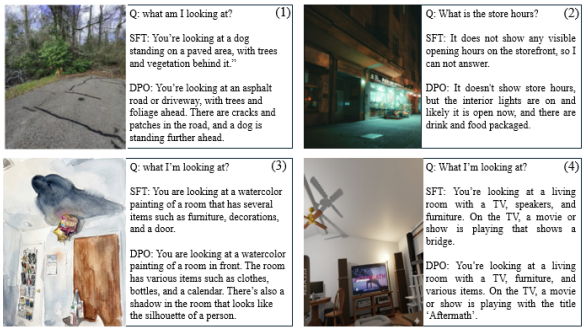


Figure 4: Representative comparisons between SFT and DPO model outputs across diverse multimodal scenarios (1st photo captured by the author, 2nd by Krüger 2024 - Unsplash, 3rd and 4th are AI generated).

Figure 5 further illustrates hallucination mitigation across spatial reasoning, OCR understanding, and contextual interpretation tasks. In several examples, the queried information is not directly observable from the image. For instance, when asked to estimate the number of pages in a book (Example 1) or determine food prices (Example 3), the SFT baseline generates plausible but unsupported guesses, whereas the DPO-trained model more reliably abstains or explicitly acknowledges insufficient visual evidence. Similarly, under challenging OCR conditions where text is only partially readable, the SFT model tends to infer unsupported details from scene context, while the DPO model avoids such ungrounded extrapolation (Example 4).

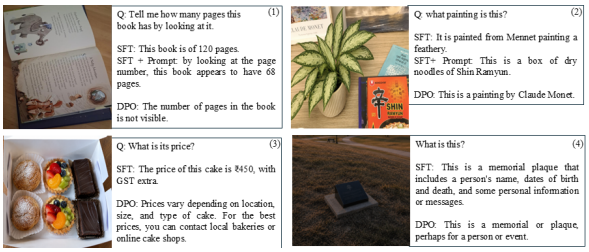


Figure 5: Representative examples of hallucination mitigation across spatial reasoning, OCR, and contextual interpretation scenarios (1st/2nd photos captured by the author; 3rd/4th images generated by AI).

We also compare the proposed approach with visual prompt-based hallucination mitigation. Although prompting can occasionally reduce fabricated details, its effectiveness remains inconsistent and may lead to overly conservative or partially irrelevant responses under ambiguous conditions. In contrast, the DPO-based approach achieves more reliable grounding through preference-based training, reducing hallucination while preserving re-

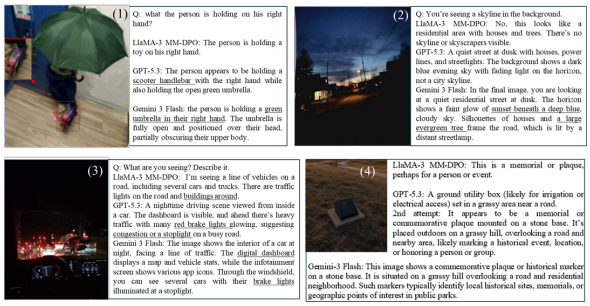


Figure 6: Cross-model comparison across challenging multimodal reasoning scenarios involving ambiguity, spatial understanding, and contextual interpretation (1st/4th images generated by AI; 2nd/3rd photos captured by the author).

sponse informativeness and task relevance. These examples highlight the limitations of prompt-level control and the advantages of data-centric preference optimization for robust multimodal grounding.

Finally, we compare our model with several frontier multimodal models. Figure 6 compares multiple VLMs across challenging multimodal reasoning scenarios involving ambiguity, spatial understanding, and contextual interpretation. The examples suggest that the proposed LLaMA-3 MM-DPO model tends to produce more conservative responses that remain closely tied to visible evidence, whereas GPT-series and Gemini models often generate richer and more contextually detailed interpretations. Under ambiguous or weak visual evidence, however, the latter models may occasionally introduce unsupported contextual inferences or over-interpretations. These qualitative comparisons highlight differing trade-offs between descriptiveness, contextual reasoning, and grounding consistency across multimodal systems. Many more examples and analysis are provided in **Appendix B**.

These observations further suggest a trade-off between descriptiveness and hallucination under ambiguous visual conditions. Models producing richer and more contextual responses may become more susceptible to unsupported inference, whereas more conservative grounding can reduce hallucination at the cost of descriptive richness. Additional discussion is provided in Appendix C.2.

4 Conclusion

We present a data-centric framework for reducing hallucination in vision-language models through stage-wise preference optimization under distribution shift. Rather than modifying the underlying architecture, our approach focuses on structured supervision and hallucination-targeted preference construction. By contrasting grounded responses with minimally perturbed hallucinated alternatives, the proposed framework improves the model's ability to distinguish linguistic plausibility from evidence-supported generation.

Our experiments suggest several key findings. First, carefully constructed preference data can produce substantial alignment gains despite relatively modest training scale, indicating that supervision quality and preference structure may play an important role during post-training refinement. Second, pairwise evaluation provides a useful signal for hallucination assessment, particularly in long-form generation settings where pointwise scoring may be affected by verbosity bias. Third, hallucination-targeted preference optimization can improve response informativeness while preserving stronger visual grounding, partially shifting the trade-off between descriptiveness and hallucination.

Qualitative analysis further suggests that hallucination often emerges near ambiguous reasoning boundaries, where autoregressive generation becomes increasingly influenced by linguistic priors rather than observable evidence. Although the proposed approach improves grounding robustness, current autoregressive multimodal models still exhibit limitations in physical reasoning, uncertainty handling, and stability under ambiguous visual conditions. Additional discussion of these limitations is provided in Appendix C.3.

Overall, this work highlights the importance of data-centric alignment for multimodal reasoning systems and demonstrates that targeted preference optimization can improve grounding behavior without requiring architectural modification. Future work will explore stronger uncertainty modeling, physical reasoning mechanisms, and alternative multimodal generation paradigms beyond standard next-token prediction objectives.

5 Limitations

Although the proposed framework improves grounding robustness and reduces hallucination in challenging multimodal scenarios, several limitations remain.

First, while our evaluation covers established open-source benchmarks (such as DocVQA, VQA,

and MMBench) alongside pairwise hallucination and benchmark assessments, it primarily focuses on static, short-context visual question answering. Exploring how stage-wise preference optimization scales to massive, multi-turn conversational agents, long-document understanding, or highly specialized domain-specific benchmarks (e.g., medical or advanced scientific imaging), etc., remains an important direction for comprehensive future validation.

Second, the preference construction pipeline relies on targeted augmentation strategies designed around known failure modes, such as spatial ambiguity, OCR uncertainty, and false-premise reasoning. Although these domains address critical multimodal challenges, they do not exhaustively cover the entire spectrum of real-world edge cases.

Third, the framework remains fundamentally dependent on the underlying autoregressive generation paradigm. While stage-wise preference optimization significantly mitigates language-prior dominance and stabilizes reasoning trajectories under ambiguous visual evidence, it operates within the constraints of next-token prediction and cannot completely eliminate intrinsic modeling uncertainties.

Finally, the cross-model comparisons involving proprietary systems are qualitative and based on publicly available API versions at evaluation time. Since these systems evolve rapidly and may employ different architectures, training procedures, or external tools, the comparisons should be interpreted as illustrative case studies rather than definitive benchmarking results.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, and 1 others. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

C.-P. Bara and 1 others. 2022. Privacy preserving visual question answering. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Paul F. Christiano, Jan Leike, Tom Brown, and 1 others. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*.

Nouha Dziri and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *Transactions of the Association for Computational Linguistics*.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Yuan Liu, Haotian Li, Yuhang Wu, and 1 others. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Aman Madaan, Niket Tandon, Peter Clark, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Long Ouyang, Jeffrey Wu, Xu Jiang, and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Zhihong Shao, Peiyi Wang, Qihao Zhu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of EMNLP*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and 1 others. 2022. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qinwu Xu, Yifan Jiang, and Haoyu Ren. 2026a. Multilingual ocr-aware fine-tuning and prompt-guided chain-of-thought reasoning for multimodal large language models. In *arXiv:2605.16409 [cs.CV]*.

Qinwu Xu, Zhuoheng Li, and Jessie Salas. 2026b. Robust checkpoint selection for multimodal llms via agentic evaluation and stability-aware ranking. In *arXiv:2605.18852 [cs.LG]*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Appendix

Appendix A:Controlling Response Detail under Grounding Constraints

To improve response completeness while maintaining factual grounding, we introduce data-centric strategies that explicitly reshape the distribution of response detail during training. We employ two complementary augmentation strategies to increase response richness:

- **Long-form duplication:**

  We apply a length-aware resampling strategy based on exponential tilting. Specifically, each sample with length x is assigned a weight:

  $$w(x) = e^{\beta(x-\mu_a)}$$

  where μ_a is the mean length of the original distribution and $\beta > 0$ controls the degree of up-weighting for longer responses. This formulation increases the relative contribution of longer answers while preserving the overall distributional structure. The parameter β is chosen to match the empirical shift in mean length between the source and target distributions. In practice, we additionally cap $w(x)$ to avoid over-amplifying extreme long-tail samples.

- **Iterative expansion:**

  We generate second-turn responses that progressively refine and elaborate the initial answers used in SFT training, adding attributes, spatial relations, and contextual details.

These strategies shift the distribution of preferred responses toward greater length and specificity and complexity such including spatial relationship and reasoning. As shown in Figure 7 , the augmented dataset exhibits a consistent increase in response length, along with improved coverage of fine-grained visual attributes.

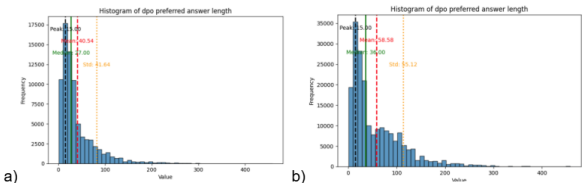


Figure 7: DPO response length of words distribution: a) original one; b) new one with data duplication and augmentation

Qualitatively, the resulting responses contain richer object descriptions, more precise spatial relations, and more explicit grounding in visual evidence.

Trade-off Between Detail and Hallucination

Increasing response detail introduces an inherent risk of hallucination. The model learns a conditional distribution $p_\theta(y \mid x)$, which combines evidence-supported generation with prior-driven completion. When visual evidence is incomplete or ambiguous, generating longer and more specific responses increases reliance on learned priors, thereby raising the likelihood of unsupported details.

This phenomenon can be interpreted through conditional entropy: increasing response richness effectively enlarges the space of plausible continuations, increasing the probability of ungrounded generations.

To mitigate this effect, our preference construction emphasizes detailed yet visually grounded responses while contrasting them against similarly detailed but subtly hallucinated alternatives. Rather

than rewarding verbosity itself, the optimization encourages response richness only when supported by observable evidence.

This preference construction and optimization strategy encourage the model to generate richer and more informative responses while suppressing unsupported elaborations. Rather than enforcing a simple trade-off between detail and hallucination, the optimization reallocates probability mass toward responses that are both detailed and visually grounded. Although residual ambiguities may persist in highly challenging scenarios, the overall effect is improved informativeness together with reduced hallucination. We empirically analyze this trade-off and its mitigation in Section 4.3.2.

Appendix B: Cross-Model Comparison

To further analyze the trade-off between response richness and hallucination, we conduct a comparative study on multimodal examples across multiple models, including our LLaMA3-MM DPO model, Gemini Flash, and GPT-series models. The representative examples are demonstrated in the appendix - Across Model Comparison and Analysis.

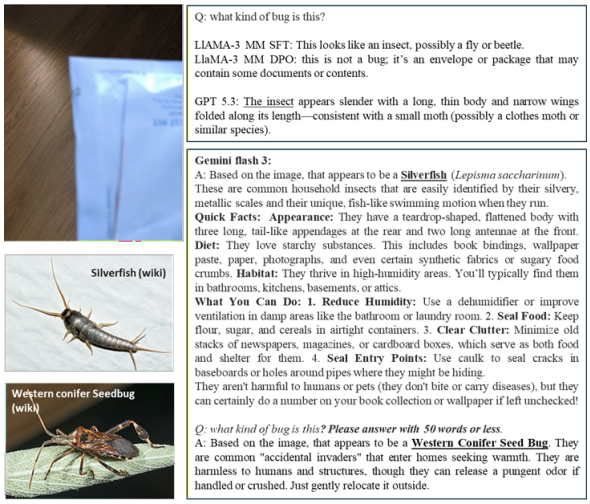


Figure 8: Adversarial hallucination example for fine-grained object identification. The DPO-trained model avoids unsupported classification and remains grounded in visible evidence, whereas GPT-series and Gemini Flash models generate specific but unverified bug categories (photo captured by author).

Figure 8 presents a representative adversarial example involving fine-grained bug classification. The queried object is visually ambiguous and lacks sufficient evidence for reliable species-level identification. In this case, the DPO-trained model responds conservatively and avoids unsupported categorization. In contrast, both GPT-series and Gemini Flash models generate specific bug labels despite limited visual evidence.

Notably, the Gemini model produces inconsistent predictions (e.g., *silverfish* versus *seed bug*) under different prompting constraints, even though these categories correspond to visually distinct insect types. This behavior suggests that the generated responses may be influenced more strongly by linguistic priors or token-level completion preferences than by stable visual grounding. These observations further highlight the importance of hallucination-targeted preference optimization for improving consistency and grounding in ambiguous multimodal scenarios.

Figure 6 presents a comparative evaluation of three multimodal models across four challenging visual question answering scenarios, highlighting differences in hallucination behavior, grounding reliability, and multimodal reasoning consistency. Overall, LLaMA-3 MM-DPO produces more conservative and visually grounded responses, avoiding unsupported inferences under ambiguous conditions. In contrast, GPT-5.3 and Gemini Flash more frequently introduce additional details that are not strictly supported by the image, reflecting a stronger tendency toward language-driven completion and contextual extrapolation.

Example 1 reveals limitations in physical reasoning and spatial grounding. In this case, Gemini Flash hallucinates the presence of a scooter despite the implausible physical configuration required for the child to carry such an object horizontally. GPT-5.3, while avoiding the scooter hallucination, incorrectly interprets the spatial relationship between the person's left and right hands under partial occlusion. Example 2 illustrates ambiguity sensitivity and reasoning instability, where Gemini Flash produces inconsistent interpretations of the same scene across different prompting conditions (e.g., sunset with visible evergreen trees versus a low-light nighttime scene). In Examples 3 and 4, all models capture the overall scene semantics; however, weaker grounding leads to varying degrees of over-interpretation, such as inferring brake-light states or contextual details beyond observable evidence.

These results suggest that stage-wise preference optimization improves grounding consistency and reduces hallucination under challenging multimodal conditions, although physical reasoning and uncertainty handling remain difficult failure modes for current vision-language models.

Overall, the results highlight differences in how multimodal models balance visual evidence against language-driven inference. While all models perform reasonably well on simpler and well-grounded queries, larger differences emerge in scenarios involving ambiguity, long-form reasoning, spatial relationships, or incomplete visual evidence. Additional challenging VQA examples are provided in the Appendix (Mini-Benchmark of different multimodal LLM models - examples).

LLaMA-3 MM (DPO) tends to produce more conservative responses, particularly under ambiguous or adversarial conditions. Its outputs remain closely tied to observable evidence, resulting in a lower incidence of unsupported inference and improved grounding consistency. However, this conservatism can occasionally reduce descriptive richness in highly uncertain scenarios.

GPT-series model (API version accessed at evaluation time) demonstrates comparatively balanced behavior between descriptiveness and grounding. The model generally produces structured and cautious interpretations while maintaining reasonable alignment with visual content. In some cases, however, it appears less consistent in leveraging fine-grained visual cues, such as small or partially occluded text, which may be omitted or only partially captured without explicit prompting.

Gemini Flash model (API version accessed at evaluation time) generates the most detailed and context-rich responses among the evaluated systems, capturing a broad range of objects, scene relationships, and contextual cues. However, this stronger descriptiveness is accompanied by a higher tendency toward over-interpretation, where additional attributes or contextual details are inferred without sufficient visual evidence or beyond the scope required by the query.

Regarding OCR-related performance, the GPT-series model shows comparatively weaker robustness under challenging text conditions, including small, blurred, or partially occluded text regions. We note that comparisons involving proprietary systems (GPT and Gemini) are based on publicly available API versions at the time of evaluation and may evolve with future model updates. Furthermore, no external OCR or computer-vision toolkits were integrated into any of the evaluated systems during testing.

Importantly, our method shifts this trade-off frontier. Relative to the LLaMA-3 MM (SFT) baseline, the stage-wise DPO-trained model improves response richness while maintaining or reducing hallucination, indicating more effective allocation of probability mass toward evidence-supported details.

A detailed comparison across multiple evaluation judges is summarized in Table 4. Across the manually analyzed challenging examples, Gemini Flash exhibited the highest frequency of hallucination cases, followed by GPT-series models, while LLaMA-3 MM (DPO) showed the lowest frequency of unsupported inference. Note that the qualitative examples are intended as illustrative case studies rather than exhaustive benchmarking. Future work may explore larger-scale human evaluation, statistical analysis, and standardized annotation protocols.

Table 4: Strengths and limitations observed across evaluated multimodal models.

Model	Strengths	Limitations
LLaMA-3 MM (DPO)	Strong grounding to visible evidence; lower tendency toward unsupported inference; more consistent handling of hallucination-prone scenarios	Can be comparatively conservative in ambiguous cases, occasionally reducing descriptive richness
GPT-5.3 Instant	Balanced reasoning and interpretation; generally structured and cautious responses; good overall grounding quality	Less consistent utilization of fine-grained visual cues (e.g., small text or partially occluded objects); occasional incomplete extraction of visual details
Gemini 3 Flash	Rich and detailed scene descriptions; strong semantic and contextual interpretation capability	Higher tendency toward over-interpretation and unsupported contextual inference under ambiguous conditions

Appendix - Mini-Benchmark of different multimodal LLM models - examples

Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
 Q: What do you see? (photo captured by author)	SFT: You're looking at a dog standing on a paved area, with trees and vegetation behind it. DPO: You're looking at an asphalt road or driveway, with trees and foliage ahead. There are cracks and patches in the road, and a dog is standing further ahead.	You are looking at an asphalt surface with visible black sealant cracks in the foreground. In the background, a small, tan-colored dog is standing near a lush green bush at the edge of a wooded area with tall, slender trees.	A cracked asphalt road or path in a wooded area, with trees and dense green bushes along the edge.	Gemini and LLaMA-3 MM (DPO) both correctly identify the key object (dog) and provide reasonable spatial grounding. GPT misses the dog entirely, indicating a failure in salient object detection. While Gemini is more detailed, both Gemini and LLaMA are grounded and accurate.
 Q: What kind of bug is this?(photo captured by author)	SFT: This looks like an insect, possibly a fly or beetle. DPO: This is not a bug; it's an envelope or package that may contain some documents or contents.	Based on the image, that appears to be a Western Conifer Seed Bug. They are common "accidental invaders" that enter homes seeking warmth. They are harmless to humans and structures, though they can release a pungent odor if handled or crushed. Just gently relocate it outside.	The insect appears slender with a long, thin body and narrow wings folded along its length—consistent with a small moth (possibly a clothes moth or similar species).	This is an adversarial case. LLaMA-3 MM correctly rejects the premise (no bug present). Both GPT and Gemini hallucinate specific insect types, demonstrating over-reliance on prior knowledge instead of visual evidence.
 Q: What painting is this? (photo captured by author)	SFT: It is painted from Mennet painting a feathery. SFT+ Prompt: This is a box of SHIN Ramyun instant noodles, a popular Korean spicy noodle brand produced by Nongshim. DPO: This is a painting by Claude Monet.	The painting visible in the top-left corner of the second image is by Claude Monet, as indicated by the text on the print. Based on the visible colors and style, it appears to be one of his famous Water Lilies (Nymphéas) series.	A framed print referencing Claude Monet is visible in the background, likely an Impressionist-style artwork rather than a clearly identifiable specific painting.	GPT gives a cautious and grounded answer, avoiding overcommitment. LLaMA-3 MM (DPO) identifies the artist correctly but lacks specificity. Gemini over-specifies by hallucinating a specific series ("Water Lilies") without sufficient evidence.
 Q: How is this game played? (photo - AI generated)	SFT: To run on a treadmill, first warm up by walking for 2 minutes at a speed of 4–5 km/h. Then jog for 10 minutes at 6–7 km/h and finally cool down by walking again for 2 minutes. You can adjust this routine according to your pace and needs. DPO: To play this game, players should sit together, and one player should spin the clock's hand. Wherever the hand stops, that time indicates how many minutes the player must perform a task.	This clock rug is an educational tool for learning to tell time. Children can physically jump or place objects on the numbers. They can also use their arms or long toys to mimic clock hands, practicing terms like "Quarter To" and "Half Past" printed on the rug.	Players spin the center arrow. When it stops, they move to or interact with the number it points to—such as stepping on it, counting, or doing a simple task. Turns rotate among players. The game focuses on learning numbers, counting, and basic turn-taking through play.	The image does not provide a clearly defined rule set. All models infer plausible gameplay mechanics, but none can be verified as correct. This is an inherently ambiguous case where hallucination cannot be strictly judged.

Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
Q: Tell me how many pages this book has by looking at it (photo captured by author).	SFT: This book is 120 pages. SFT + Prompt: by looking at the page number, this book appears to have 68 pages. DPO: The number of pages in the book is not visible.	The visible page numbers are 102 and 103. Given the visible thickness of the remaining pages underneath, the book likely contains at least 200 to 250 pages in total. It appears to be a large treasury or collection of children's stories.	I can't determine the exact number of pages from this image alone. It only shows one open spread of an illustrated children's book, without visible page numbers, spine thickness, or cover details.	LLaMA-3 MM (DPO) and GPT correctly refuse to infer beyond visible evidence. Gemini hallucinates a page number based on a visible page index, which does not indicate total page count.
Q: What is this? (photo - AI generated)	SFT: This is a memorial plaque that includes a person's name, dates of birth and death, and some personal information or messages. DPO: This is a memorial or plaque, perhaps for a person or event.	This image shows a commemorative plaque or historical marker on a stone base. It is situated on a grassy hill overlooking a road and residential neighborhood. Such markers typically identify local historical sites, memorials, or geographic points of interest in public parks.	A ground utility box (likely for irrigation or electrical access) set in a grassy area near a road. 2nd attempt: It appears to be a memorial or commemorative plaque mounted on a stone base. It's placed outdoors on a grassy hill, overlooking a road and nearby area, likely marking a historical event, location, or honoring a person or group.	LLaMA-3 MM and Gemini both correctly identify the object as a memorial/plaque. GPT initially misclassifies it as a utility box but corrects itself in a second attempt. Gemini provides richer context but slightly overextends beyond the question.
Q: What am I looking at?	SFT: You are looking at a dark screen or surface, possibly a view of a room or area that is dimly lit. DPO: You are looking at a dark-colored screen with no visible scene or text.	You are looking at a completely black, solid image. There are no visible objects, colors, or patterns present in the file. It appears to be either a blank frame or an image taken in total darkness.	The image is too dark to make out any clear objects or details.	All three models appropriately express uncertainty without hallucinating content. Differences are stylistic rather than substantive.
Q: What is in front of me? (photo captured by author)	SFT: In front of you is a large aquarium with many types of fish and other aquatic animals. DPO: In front of you is a large aquarium with various types of fish.	You are looking into a large aquarium or fish tank. A school of small, silvery fish with dark spots **near** their tails is swimming in front of you. The tank features a sandy bottom, large rocks or driftwood, and some green aquatic plants. (note: dark spots are at the center)	An aquarium tank with fish and a sea turtle swimming underwater.	All models correctly identify an aquarium. Gemini provides the most detailed description but includes minor spatial inaccuracies. GPT hallucinates a turtle, which is not clearly supported. LLaMA-3 MM is concise and fully grounded.

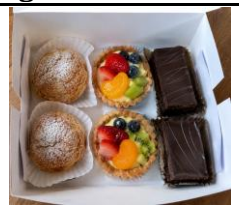
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
 Q: What is its price? (AI generated)	SFT: The price of this cake is ₹450, with GST extra. DPO: Prices vary depending on location, size, and type of cake. For the best prices, you can contact local bakeries or online cake shops.	You are looking at an assortment of six cake slices in a white bakery box. The box contains two pink-frosted slices, two dark chocolate or black forest slices, and two triangle-shaped honey or mille-feuille slices. While prices vary by bakery, a set like this typically costs between $20.00 and $35.00.	The image shows a box with desserts (cake slices and a grilled sandwich), but no price is visible, so the cost cannot be determined.	LLaMA-3 MM and GPT correctly refuse to infer price without visual evidence. Gemini hallucinates a plausible price range, which is not grounded in the image.
 Q: Is this teacher a good instructor? (AI generated)	SFT: I don't know exactly how the instructor is, but they are working hard to engage their students and provide valuable information. DPO: Evaluating the quality of an instructor is difficult because it depends on many factors, such as teaching style, subject matter expertise, and interaction with students. In this image, the instructor is shown giving a lecture, but there's no specific information about their teaching skills or interaction with students. Therefore, I cannot conclude whether they are good instructors or not.	The instructor uses visual aids and maintains an orderly environment. While students appear focused, assessing teaching quality requires hearing their explanations and seeing student-teacher interaction. Currently, the setup looks professional and well-organized for a lecture-style class.	There's no clear way to judge the teacher's quality from this image alone. It just shows a classroom setting with a teacher presenting and students listening; effectiveness depends on things like teaching style, clarity, engagement, and student outcomes, which aren't visible here.	All models appropriately avoid making unsupported judgments. Differences lie only in verbosity and explanation depth.
 Q: Look over here and read the sides (photo captured by author).	SFT: French Fries, Tater Tots, Hush Puppies, Coleslaw, and Chef's Veg. DPO: The menu lists the following sides: French Fries / Tater Tots – $1.75 Hush Puppies (s) – $1.75 Coleslaw – $1.50 Chef's Veg – $1.50 Side Caesar Salad – $1.50 Side House Salad – $1.50 Millibilly Cheddar Biscuit – $1.50 Beer-Battered Onion Rings – $2.50	The Sides listed on the menu are: French Fries | Tater Tots (3.75) Hush Puppies (5) (3.95) Coleslaw (2.75) Beer-Battered Onion Rings Side Caesar Salad (5.75) Side House Salad (5.75) Chef's Veg (3.50)	Sides listed: French fries ("tater tots"), hush puppies, coleslaw, beer-battered onion rings, side Caesar salad, side house salad, and Chef's veg.can. *Follow-up Q: give price details.* *A:* French fries ("tater tots"): about $3, Hush puppies (5): about $6 Coleslaw: about $3, Beer-battered onion rings: about $6, Side Caesar salad: about $6, Side house salad: about $6, Chef's veg: about $6	LLaMA-3 MM provides the most accurate OCR extraction including prices. Gemini captures items but partially misses or distorts price details. GPT omits prices unless explicitly prompted and later produces inaccurate estimates.

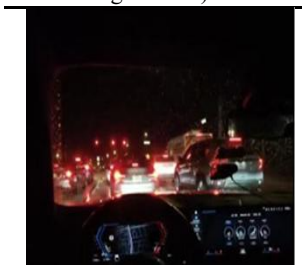
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
Q: You're seeing a skyline (photo of PxHere, licensed under CC0).	SFT: I see a peaceful street scene with buildings and trees in the background. The sky is cloudy, and the streetlights are glowing. DPO: No, this looks like a residential area with houses and trees. There's no skyline or skyscrapers visible.	No, this is a train station platform at night. A skyline shows a city's profile against the sky. This image features a brick building with overhanging lights, a dark platform, and railroad tracks, capturing a localized, industrial mood rather than a distant urban horizon.	I don't see a skyline. The image appears to show a dark nighttime alley or walkway beside a building, with colorful artificial lights and railroad tracks extending into the distance under a deep blue sky.	LLaMA-3 MM and GPT correctly reject the premise (no skyline). Gemini provides a detailed but partially hallucinated description (e.g., inferred sunset, specific tree types).
Q: What am I looking at? (AI augmented)	SFT: You're looking inside an elevator, where a small child is standing. DPO: You're looking inside an elevator, where a small child is standing with a toy.	You are looking at a child standing inside an elevator with blue chevron-patterned walls and a wood-look floor. The child is wearing a dark jacket with a yellow lining and is carrying a red and blue backpack. The elevator's control panel and a informational sign are also visible.	A small child standing in an elevator corner, holding a scooter. The elevator has padded blue walls, wood-pattern flooring, and a control panel with buttons and notices.	LLaMA-3 MM provides a correct and minimal description. Gemini and GPT hallucinate objects (backpack, scooter) with weak physical plausibility.
Q: What am I looking at? (AI generated)	SFT: A person is looking out from inside a vehicle, possibly a car or SUV." DPO: You're looking at an impressionist painting. A person is sitting **inside** a vehicle, possibly a car or SUV. They're in the **passenger** seat and the **driver's seat is empty**. **Outside**, you can see a residential area with houses and lawns.	You're viewing an impressionist painting of a car's interior from the backseat. The artist uses thick, visible brushstrokes to capture light and form, showing the front passenger seat, a seatbelt, and a suburban street through the window, transforming a routine drive into a textured, artistic scene.	You're looking at a stylized, impressionistic view from inside a car. The interior frame, seat, and a passenger are visible. Through the window, a suburban street with houses and a parked car appears, all painted with loose, colorful brushstrokes.	All models correctly identify both interior and exterior context. Differences are mainly stylistic; no major hallucinations.
 Q: What are you seeing? (photo captured by author)	SFT: I'm seeing several cars on a road, along with traffic lights and other vehicles. DPO: I'm seeing a line of vehicles on a road, including several cars and trucks. There are traffic lights on the road and buildings around.	The image shows the interior of a car at night, facing a line of traffic. The digital dashboard displays a map and vehicle stats, while the infotainment screen shows various app icons. Through the windshield, you can see several cars with their brake lights illuminated at a stoplight.	A nighttime driving scene viewed from inside a car. The dashboard is visible, and ahead there's heavy traffic with many red brake lights glowing, suggesting congestion or a stoplight on a busy road.	Gemini hallucinates dashboard map details. GPT and Gemini both assume "brake lights," which may simply be tail lights at night. LLaMA-3 MM remains more conservative but slightly vague.

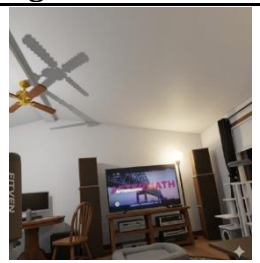
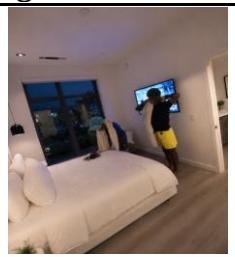
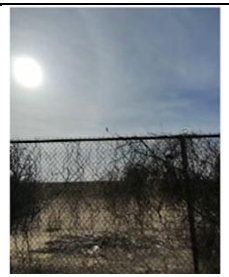
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
 Q: What am I looking at? (AI generated)	SFT: You're looking at a living room with a TV, speakers, and furniture. On the TV, a movie or show is playing that shows a bridge." DPO: You're looking at a living room with a TV, furniture, and various items. On the TV, a movie or show is playing with the title 'Aftermath'.	you're looking at a 3D-rendered room featuring a unique visual glitch: a realistic ceiling fan casting a pixelated, Minecraft-style shadow. The scene includes a TV displaying "AFTERMATH," large speakers, a cat tree, and a desk, blending realistic textures with digital anomalies.	You're looking at a living room. A TV is centered on a stand, showing a program. Surrounding it are speakers, shelves, and furniture like a chair and table. A ceiling fan with lights is above, casting shadows on the ceiling.	Gemini provides the most complete scene understanding including OCR text. LLaMA-3 MM is accurate but less detailed. GPT misses OCR signals.
 Q: What am I looking at? (AI generated)	SFT: You're looking at a room that has several items such as furniture, decorations, and a door. DPO: You are looking at a watercolor painting of a room in front. The room has various items such as clothes, bottles, and a calendar. There's also a shadow in the room that looks like the silhouette of a person.	You are looking at a watercolor painting of a room interior. It features a prominent dark, smoky wash on the ceiling above a cluster of gold and purple balloons. The scene includes a wooden door, an over-the-door organizer filled with tools, and a calendar hanging on the wall.	You're looking at a stylized, watercolor-like interior scene. A door with a curtain is on the right, a refrigerator with magnets or photos on the left, and balloons hang near the ceiling. A dark cloud-like shape appears overhead, giving the scene a slightly surreal or moody atmosphere.	Gemini provides the most accurate and detailed description. GPT introduces hallucinated objects (refrigerator). LLaMA-3 MM is correct but less descriptive.
 Q: Tell me what I'm looking at right now? (AI generated)	SFT: You're looking at a room with wooden flooring, curtains, and various furniture. DPO: You're looking at a painting of a room. The room has wooden flooring and curtains on the walls. The room also has some furniture, such as a table and a few chairs.	You are looking at an impressionist oil painting of a home interior. Thick, textured brushstrokes depict a living area with a small wooden table, a larger dining table with chairs, and patterned curtains. The dark, reflective flooring contrasts with the soft, muted light filtering through the room.	You're looking at a painterly interior of a living room or bedroom. Light filters through patterned curtains on the left, illuminating a small table and objects. On the right, a darker adjoining space with more furniture recedes, creating contrast between bright and shadowed areas.	GPT provides the most balanced and accurate description. Gemini slightly over-interprets objects. LLaMA-3 MM is the least detailed.

Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3 Instant	Analysis
 Q: What am I seeing? (AI generated)	SFT: You are looking at a bedroom with two people, perhaps changing clothes or organizing. DPO: You are in a bedroom where two people are clothed. One person is watching TV, and the other is perhaps looking for something on the bed.	You are looking at a bedroom illuminated with purple light. Two people are present: one is sitting on a bed with white linens, and another stands nearby wearing yellow shorts and a dark shirt. A television is mounted on the wall, and a ceiling fan is visible above.	A bedroom scene with two people: one standing near a wall-mounted TV adjusting or looking at it, and another bending over the bed. The bed is unmade, and a ceiling fan is above.	All exhibit some errors in spatial reasoning and physical implausity.
 Q: What am I looking at? (photo captured by author)	SFT: You're looking at a fence on which vines or plants are growing. In the background, there's an open area and a clear blue sky above. DPO: You're looking at a chain-link fence on which vines or plants are growing. Behind the fence, there's an open area, and a bird is sitting on top of the fence. In the background, there's a clear blue sky with the sun shining.	You are looking at an outdoor scene on a bright, sunny day. In the foreground, there is a chain-link fence overgrown with dry, leafless vines. Beyond the fence is a flat, open field of dry grass or dirt under a clear sky with a prominent sun.	An outdoor scene with a chain-link fence in the foreground, dry grass and sparse vegetation beyond it, and a bright sun in a lightly clouded sky overhead.	LLaMA-3 MM hallucinates a bird. GPT and Gemini remain grounded and accurate.
  Q: read the text (brochure of Operation Homefront)	SFT: HOLIDAY TOY DRIVE Experience the joy of giving. DPO: HOLIDAY TOY DRIVE Experience the joy of giving. Buy toys to support military families through Operation Homefront.	HOLIDAY TOY DRIVE Experience the joy of giving. Please purchase toys to support military families through Operation Homefront.	HOLIDAY TOY DRIVE Experience the joy of giving. There's also a circular logo at the top right that appears to include: U.S. Marine Corps Reserve – Toys for Tots (based on the recognizable train logo)	LLaMA-3 MM and Gemini provide accurate and faithful OCR. GPT adds extra contextual content not present in the image.

Note: (1) the underlined words are hallucinations, (2) Some models tend to generate excessively long responses; therefore, prompts explicitly constrain outputs to approximately 50-60 words or fewer, (3) due to stochastic generation, evaluation noise, image resolution variations, and other nondeterministic factors, exact responses may not be fully reproducible across runs.

666

Appendix C: Discussion

Appendix C.1:Data Scaling and Preference Efficiency

A notable finding is the strong data efficiency of preference-based optimization. With only approximately 50k–120k hallucination-targeted DPO samples, the model exhibits consistent improvements in grounding consistency and hallucination reduction.

This contrasts with pretraining and supervised fine-tuning (SFT), which typically require substantially larger datasets to establish broad multimodal capabilities. While large-scale training remains essential for general perception and language understanding, our results suggest that carefully constructed preference data can produce disproportionately large alignment gains during later refinement stages.

We attribute this efficiency primarily to data structure rather than raw scale. The proposed hallucination-targeted preference pairs are constructed near difficult reasoning and grounding boundaries, yielding high-information training signals that directly target failure modes of multimodal generation. These findings suggest that, for alignment-oriented objectives, improving data quality and preference structure may be more impactful than simply increasing training scale.

Interestingly, these findings are closely related to the observed detail–grounding trade-off. The effectiveness of relatively small preference datasets suggests that hallucination behavior is concentrated near ambiguous reasoning boundaries, where carefully structured supervision can strongly influence the balance between descriptive richness and visual grounding.

Appendix C.2: Trade-offs Between Detail and Hallucination

Our experiments reveal a consistent relationship between response richness and hallucination risk. As model outputs become longer and more descriptive, the probability of introducing unsupported attributes or contextual inferences increases, particularly under incomplete, ambiguous, or weak visual evidence.

This behavior is illustrated in Figure 6. Models favoring highly detailed responses (e.g., Gemini Flash) exhibit stronger tendencies toward over-interpretation, whereas more conservative models reduce hallucination at the cost of descriptiveness and contextual richness.

This phenomenon can be interpreted as a shift from evidence-conditioned generation toward prior-driven completion. Increasing response length expands the space of plausible continuations, increasing reliance on learned priors and language-level completion behavior.

Our approach mitigates this effect through the preference construction strategy introduced in Section 3.4.2, where detailed grounded responses are contrasted against similarly detailed but hallucinated variants. This isolates grounding—not verbosity—as the primary optimization signal.

As a result, the proposed stage-wise preference optimization reshapes the generation distribution toward evidence-supported detail rather than merely suppressing response length. Empirically, this shifts the detail–grounding trade-off frontier, enabling richer responses while maintaining or improving factual consistency.

Nevertheless, residual trade-offs persist in highly ambiguous scenarios, indicating that fully disentangling descriptiveness from hallucination remains an open challenge for multimodal reasoning systems.

Appendix C.3: Limitations of Autoregressive Multimodal Generation

Despite substantial improvements in grounding and hallucination reduction, several fundamental limitations remain difficult to address within autoregressive likelihood-based generation frameworks.

First, current multimodal models still exhibit weaknesses in physical and spatial reasoning, particularly under occlusion, ambiguity, or complex human-object interactions. In such cases, token prediction may favor statistically plausible completions that violate physical consistency.

Second, hallucination is closely tied to the language modeling objective itself. Because autoregressive models optimize joint token likelihood, generation can become dominated by linguistic plausibility and contextual priors rather than strict visual evidence, especially when the image provides incomplete or ambiguous information.

Third, model outputs can vary noticeably across prompting conditions, decoding strategies, or response-length constraints. As illustrated in several adversarial examples, relatively small prompt variations may lead to substantially different interpretations of the same visual input, indicating instability in the underlying reasoning trajectories.

These observations suggest that hallucination is not solely a data or alignment problem, but

also reflects intrinsic limitations of autoregressive probabilistic generation. Future work may require stronger uncertainty modeling, explicit physical reasoning mechanisms, or alternative multimodal generation paradigms beyond standard next-token prediction objectives.