REDUCING HALLUCINATION IN MULTIMODAL LARGE LANGUAGE MODELS THROUGH HARD GROUNDING PREFERENCE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains a major challenge in vision-language models (VLMs), particularly when linguistically plausible responses are unsupported by visual evidence. We study whether multimodal hallucination can be reduced by concentrating post-training supervision on hard grounding boundaries, where preferred and rejected responses are semantically close but differ in their support from observable visual evidence. Under a frozen visual encoder and cross-modal alignment pathway, we first use supervised fine-tuning (SFT) to establish broad decoder-side multimodal behavior and then construct hard grounding preference pairs for Direct Preference Optimization (DPO). These pairs target evidence utilization, calibration, and grounding consistency across fine-grained recognition, spatial reasoning, OCR, ambiguous or insufficient evidence, and false-premise queries. The resulting DPO model consistently improves over its SFT initialization on DocVQA, TextVQA, MMBench, and VQAv2. Across approximately 839 additional multimodal examples, it further achieves 6.2–8.2 percentage-point higher pairwise win rates under three independent LLM judges. We additionally release HardVQA-DPO, a curated and growing resource with more than 3K hard-grounding SFT examples and an initial set of DPO preference pairs.[1] These results show that decoder-side adaptation alone can improve a meaningful subset of grounding failures under a fixed multimodal representation, while preserving broader multimodal capabilities.

1 INTRODUCTION

1.1 MULTIMODAL HALLUCINATION AND VISUAL GROUNDING

Multimodal large language models (MLLMs) have achieved strong performance across visual question answering, document understanding, multimodal reasoning, and open-ended image understanding. Despite these advances, hallucination remains a persistent challenge: models can generate responses that are linguistically coherent and semantically plausible but unsupported by observable visual evidence. Prior work has studied hallucinations involving object existence, attributes, and relations (Li et al., 2023; Wang et al., 2023), while more diagnostic studies show that language priors can conflict with or dominate visual context (Guan et al., 2024; Favero et al., 2024). These failures become particularly difficult when evidence is small, occluded, crowded, ambiguous, or otherwise difficult to recognize, where a plausible interpretation is not necessarily a visually grounded one.

MLLMs typically model $p_\theta(y \mid x_v, x_t) = \prod_t p_\theta(y_t \mid x_v, x_t, y_{<t})$, where x_v is the visual input, x_t the textual query, and y the generated response. Standard autoregressive supervision increases the likelihood of a target response but does not explicitly contrast a visually supported continuation with a plausible but unsupported alternative. Generation therefore reflects both visual evidence and statistical regularities encoded by the language model. POPE, for example, finds that frequently occurring or co-occurring objects are particularly susceptible to hallucination (Li et al., 2023); HallusionBench identifies cases where language bias overshadows visual context (Guan et al., 2024); and Favero et al. (2024) show that dependence on the visual prompt can decrease as autoregressive

[1] https://huggingface.co/datasets/vlmgrounding/hardvqa-sft-dpo

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generation proceeds. Once an unsupported claim enters the generated context $y_{<t}$, subsequent tokens are conditioned on that claim, potentially extending it into a coherent but visually unsupported response.

Figure 1 illustrates that fine-grained grounding failures are not fully characterized by conventional task categories. Visual evidence may be incorrectly resolved across overlapping instances or missed when partially occluded; difficult yet sufficient evidence may trigger unnecessary rejection; an unsupported premise may induce a specific, plausible continuation; and even a correct final decision may be accompanied by reasoning inconsistent with the image. Notably, multiple frontier MLLMs exhibit the first failure on the same subtle counting example. Thus, categories such as object existence, counting, attributes, OCR, and spatial relations describe *what* visual content is evaluated, but not necessarily *how* that evidence is used. We therefore consider three complementary grounding behaviors: **evidence utilization**, **evidence calibration**, and **grounding consistency**. The goal is not generic conservatism, but to match response specificity and reasoning to the available visual evidence.

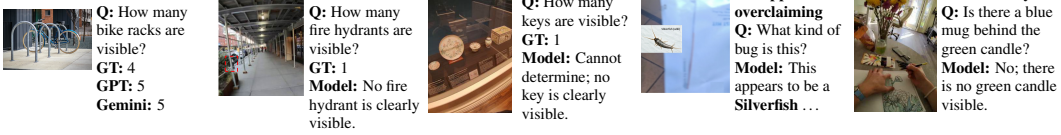
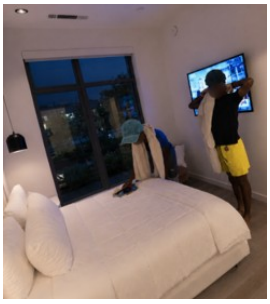


Figure 1: Representative fine-grained grounding failures in MLLMs. **(a)** Multiple frontier MLLMs miscount overlapping bicycle racks. **(b)** Available but partially occluded evidence is under-utilized. **(c)** Difficult yet sufficient evidence produces an unnecessary rejection. **(d)** A false premise induces a plausible but unsupported continuation. **(e)** Responses reach the same final decision, but one explanation is inconsistent with visible evidence. Together, these examples expose complementary failures in evidence utilization, evidence calibration, and grounding consistency.

Grounding inconsistency can also propagate across conversational turns. As illustrated in Figure 2, an early misinterpretation of a visual or spatial relationship can become part of the linguistic context for subsequent questions. Later responses may then remain internally coherent while inheriting the earlier unsupported assumption. Thus, grounding consistency concerns not only individual predictions or rationales, but also whether visually unsupported commitments are propagated through subsequent generation.



Q: Is the right person facing toward the TV?
A: No, the person on the right is facing away from the TV. Their back is toward the screen as they appear to be putting on or adjusting a shirt. *(Initially plausible interpretation.)*
Q: Look at the whole body and analyze again.
A: The person is still facing away from the TV. Their torso and shoulders suggest their back is toward the screen, and their arms are raised as they pull on a shirt. Their back is clearly visible to the camera; their heels are closest to the TV stand; their legs are oriented toward the room. *(Unsupported spatial inference.)*
Q: You said "their back is clearly visible to the camera"—do you see the back?
A: Yes, the posture and arm positioning indicate that the back is facing the camera, consistent with the person facing away from the TV. *(Unsupported visual evidence.)*
Q: You said "their heels are closest to the TV stand"—where is the TV stand?
A: The TV stand refers to the surface beneath the wall-mounted television. The person's heels are positioned near that area, suggesting they are facing into the room. *(Introduces a non-existent object and spatial relation.)*
Q: Wait, is the "low surface" the baseboard beneath the TV?
A: No, it refers to a horizontal furniture piece (e.g., a media console) under the TV. The baseboard runs along the wall, while the person is positioned near the furniture piece. *(Reinforces the hallucinated object and 3D relation.)*

Figure 2: **Multi-turn propagation of a grounding error.** An initially plausible interpretation of the person's orientation is retained across follow-up questions and progressively supported by unsupported spatial claims. Blue text highlights unsupported claims that accumulate from body orientation to a non-existent TV stand or media console and associated spatial relationships. The example illustrates how an initial grounding error can become part of the conversational context and support increasingly elaborate but internally coherent reasoning.

1.2 Related Work and Remaining Gap

Existing work approaches multimodal hallucination from complementary directions. POPE focuses on object hallucination and identifies effects of object frequency and co-occurrence (Li et al., 2023), while AMBER evaluates existence, attribute, and relation hallucinations (Wang et al., 2023). HallusionBench further distinguishes language hallucination from visual illusion and studies conflicts between language priors and visual context (Guan et al., 2024). Closely related to our motivation, Favero et al. (2024) study the competition between visual conditioning and the language prior during autoregressive generation and propose an inference-time mechanism that amplifies dependence on the image, with an extension using DPO. These works motivate studying not only the visual category in which an error occurs, but also how visual evidence influences generation.

Training-time alignment provides a complementary intervention. Multimodal instruction tuning and curated SFT can improve cross-modal capabilities and grounding (Moon et al., 2024; Xu et al., 2026), while RLHF and DPO provide comparative supervision between preferred and rejected responses (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Recent multimodal work further shows that preference-data construction is important. OPA-DPO (Yang et al., 2025) addresses mismatch introduced by off-policy preference data, while LPOI (Pesaran Zadeh et al., 2025) introduces object-aware listwise optimization using different levels of object visibility. These approaches demonstrate the value of multimodal preference optimization, but leave an important data-design question: how should preference supervision be constructed when the rejected response is not obviously incorrect, but differs from the preferred response only at a fine-grained grounding boundary? In such cases, both responses may be fluent, plausible, and largely semantically aligned, while differing in a subtle count, attribute, spatial relation, false premise, or unsupported reasoning step.

1.3 Hard Grounding Preference Supervision

We address this problem through *hard grounding preference supervision*. We construct preference pairs in which preferred and rejected responses are deliberately close in fluency and semantic content but differ in their support from visual evidence. The pairs cover conventional capabilities such as counting, OCR, attributes, spatial reasoning, and false-premise questions, while targeting the underlying grounding behaviors described above. Importantly, the supervision operates in both directions: when evidence is insufficient, the preferred response avoids unsupported specificity; when subtle but sufficient evidence is available, it preserves the supported answer rather than defaulting to rejection.

Our experimental design further isolates this intervention to the language decoder. As shown in Figure 3, the ViT visual encoder and Perceiver-based alignment module remain frozen throughout both SFT and DPO, while only LoRA parameters in the LLaMA-3 decoder are updated. Thus, changes between SFT and DPO cannot result from modified visual-encoder or alignment weights. This does not assume that the fixed visual representation is complete or that all hallucinations originate in the decoder; rather, it tests whether decoder-side adaptation is sufficient to improve grounding when useful evidence is already represented in the fixed multimodal input.

Our contributions are: **(i)** we formulate hard grounding in terms of evidence utilization, evidence calibration, and grounding consistency, complementing conventional task-level hallucination categories; **(ii)** we construct semantically close preference pairs that contrast visually supported responses with plausible but unsupported alternatives, targeting both unsupported overclaiming and unnecessary rejection; **(iii)** under a frozen visual encoder and alignment pathway, we show that decoder-side post-training can improve a meaningful subset of grounding failures using fixed multimodal representations; and **(iv)** we evaluate the resulting models on standard benchmarks, grounding-sensitive pairwise evaluation, and challenging qualitative cases, and release the HardVQA-DPO resource.

2 Methodology

2.1 Model Architecture and Framework Overview

We instantiate our framework on a large-scale MLLM consisting of a ViT visual encoder, a Perceiver-based cross-modal alignment module, and a LLaMA-3 70B language decoder. The vi-

sual encoder extracts image features, which are compressed by the Perceiver module and provided together with the textual input to the language decoder for autoregressive generation.

Our framework does not modify the model architecture. As illustrated in Figure 3, the ViT encoder and Perceiver alignment module remain frozen throughout post-training, while only LoRA parameters in the language decoder are updated. Post-training proceeds in two stages: large-scale multimodal SFT first establishes broad decoder-side multimodal question answering and grounding behavior, after which targeted DPO refines the decoder using hard grounding preference pairs. This fixed-pathway design provides the experimental control used throughout the paper: changes from SFT to DPO reflect decoder-side adaptation under the same visual feature extraction and cross-modal alignment pathway.

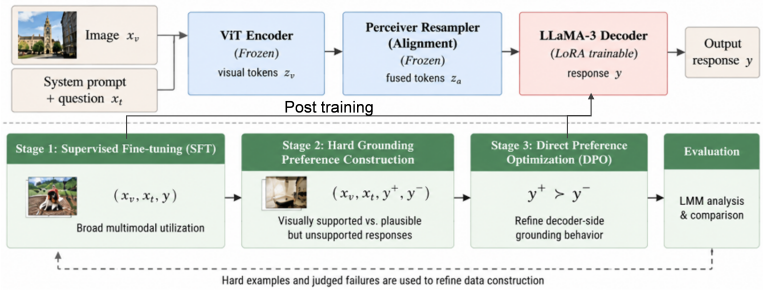


Figure 3: Fixed-visual-pathway architecture and stage-wise decoder post-training. The ViT encoder and Perceiver-based alignment module remain frozen throughout training, while Stage-1 SFT and Stage-2 DPO update only decoder-side LoRA parameters. Hard grounding preference pairs contrast visually supported responses with plausible but unsupported alternatives, targeting how the decoder utilizes fixed multimodal representations for grounded generation.

2.2 Hard Grounding Preference Construction

Let $x = (x_v, x_t)$ denote an image–question pair and y its grounded SFT response. From selected SFT examples, we construct harder preference instances (x', y^+, y^-) on the same image. A generator MLLM is provided with the original question–response pair and a target grounding behavior to generate a harder question and candidate responses. Candidates are subsequently adjudicated and, when necessary, manually refined. Thus, y^+ and y^- are often both fluent and plausible while differing primarily in their support from visual evidence. Generation and evaluation prompt templates are provided in the HardVQA-DPO repository.

Rather than defining hard examples solely by task category, we organize preference construction around three complementary grounding behaviors. **Evidence utilization** targets whether subtle but available evidence is used, including small, partially visible, crowded, or visually confusable objects. **Evidence calibration** targets whether response specificity matches the available evidence: y^+ rejects or reduces specificity when evidence is insufficient, but retains a specific answer when difficult yet sufficient evidence is present. **Grounding consistency** targets whether individual claims and reasoning remain consistent with the image, including cases in which y^+ and y^- reach the same final decision but differ in their visual justification. These behaviors are not mutually exclusive and span conventional tasks including counting, OCR, attributes, spatial relationships, object existence, and false-premise questions.

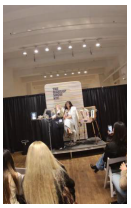
The central design principle is to keep substantial semantic content shared between y^+ and y^- while preserving a grounding-relevant distinction. A hard pair may differ only in a count, attribute, spatial relation, object-existence claim, or reasoning step. Importantly, the construction is bidirectional: insufficient evidence should lead to rejection or reduced specificity, whereas difficult but sufficient evidence should retain the supported prediction. Preference learning therefore cannot succeed merely by making the model uniformly more conservative.

Figure 4 illustrates how these grounding dimensions are instantiated across different visual tasks and conditions. When both responses are visually valid, preference is instead determined by question

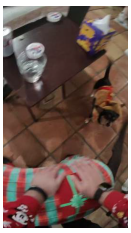
relevance and grounded informativeness, encouraging additional detail when that detail is supported by the image.



(a) Occluded evidence
Q: How many cards are currently in my hand?
+ **Preferred: 5.** Five cards are visible, including a partially occluded fifth card.
− **Rejected: 4.** Only four cards are identified as clearly visible.



(b) Crowded instances
Q: How many spotlights are in the row closest to the top?
+ **Preferred: 7.** Careful counting gives approximately seven fixtures despite fisheye distortion.
− **Rejected: 6.** The uppermost row is judged to contain exactly six visible fixtures.



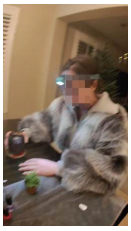
(c) Difficult but sufficient
Q: In the fifth photo-strip frame from the left, how many dogs can be seen?
+ **Preferred: 2.** The fifth frame shows two dogs standing together; preceding frames show one dog each.
− **Rejected: 1.** The rotated, small, partially obscured strip is judged too unclear to confirm two dogs.



(d) Insufficient evidence
Q: If I am walking North, the path turns East.
+ **Preferred: Unanswerable.** The tilted view provides no compass, map, or directional cue for determining East or West.
− **Rejected: No.** The path is asserted to curve left (West), implicitly treating image-left as West.



(e) False premise
Q: Is there a small animal, like a rat, visible peering out from the storm drain?
+ **Preferred: No.** No animal is visible near or inside the drain grate; only feet and pavement are visible.
− **Rejected: No.** It instead claims that no storm-drain grate is identifiable, failing to ground the rejection in the visible drain.



(f) Grounding consistency
Q: Is the person holding a smartphone in their right hand?
+ **Preferred: No.** The person is holding a glass in their right hand, not a smartphone.
− **Rejected: No.** It identifies a cup/glass but incorrectly assigns it to the person's left hand and misinterprets the left–right relation.

Figure 4: **Representative hard grounding preference pairs constructed from WearVQA images.** Each example pairs a hard visual question with a preferred (+) and rejected (−) response whose distinction depends on visual grounding rather than fluency alone. **(a–b)** Fine-grained evidence utilization requires resolving partially occluded or crowded instances. **(c–d)** Evidence calibration is bidirectional: difficult but sufficient evidence should retain a specific prediction, whereas genuinely insufficient evidence should not induce an unsupported inference. **(e–f)** Grounding consistency distinguishes responses that may reach the same final decision but differ in whether their supporting claims are consistent with the image. The examples are illustrative rather than mutually exclusive and span the broader dimensions of evidence utilization, evidence calibration, and grounding consistency. All images are from WearVQA.

To facilitate inspection and reuse, we release **HardVQA-DPO**, a curated and growing resource containing more than 3K hard-grounding SFT examples and an initial set of DPO preference pairs. Additional construction examples are provided in Appendix A.

2.3 Hard False-Premise Construction

We construct false-premise examples by introducing an object, attribute, or relation that is plausible in context but unsupported by the image. Rather than using obviously absent concepts, we select premises that are visually confusable with nearby content because of appearance, occlusion, spatial proximity, or scene context. For example, Figure 1(d) asks about a bug although the visible object is not a bug, creating pressure to complete the plausible premise with a specific insect identity. Figure 4(e) shows a harder case in which both responses reject a proposed rat, but only the preferred response grounds the rejection in the visible drain. Thus, preference is assigned not merely for rejecting a false premise, but for matching the response and its justification to the available visual evidence.

2.4 Response Detail and Length Control

A potential confound in hallucination-oriented preference construction is response length: unsupported claims can be reduced trivially by producing shorter or less informative answers. Our objective instead is to suppress unsupported specificity while preserving supported detail. We therefore use two complementary controls. First, length-aware resampling applies exponential tilting, assigning a response of length l the weight $w(l) = \exp[\beta(l - \mu)]$, where μ is the mean response length and the weights are capped to avoid over-amplifying extreme long-tail examples. Second, for selected

SFT examples we expand short responses with additional visually supported attributes, spatial relationships, and contextual details. These controls provide richer supervision for constructing y^+ and y^- and reduce the likelihood that preference is determined merely by response length or level of detail.

2.5 Preference Data Quality Control

Because hard preference pairs can differ only in fine-grained details, we use multi-stage quality control to verify visual grounding and preference validity. Candidate pairs are generated using one LLM and independently evaluated by two LLM judges for whether (i) y^+ is supported by the image, (ii) y^- represents a valid lower-preference response, and (iii) the preference distinction is consistent with the intended grounding difficulty. An agentic metadata-processing stage consolidates the judgments and associated metadata. We additionally manually review approximately 3% of judge-agreement samples together with judge-disagreement cases. Only examples passing the resulting filtering process are retained for preference training.

2.6 Stage-wise Grounded Post-Training

Rather than constructing an independent preference dataset after SFT, we derive $\mathcal{D}_2$ primarily by extending selected grounded examples from the SFT dataset $\mathcal{D}_1$ into harder, semantically close preference boundaries. Stage 1 performs broad multimodal SFT, updating only decoder-side LoRA parameters while the visual encoder and alignment module remain frozen. Let π_{SFT} denote the resulting policy. Stage 2 initializes from π_{SFT} and applies DPO (Rafailov et al., 2023) to $\mathcal{D}_2$. For a preference triple (x, y^+, y^-), we optimize

$$\mathcal{L}_{\mathrm{DPO}} = -\log \sigma\left(\beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\mathrm{SFT}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\mathrm{SFT}}(y^- \mid x)}\right]\right), \tag{1}$$

where π_{SFT} is the fixed DPO reference policy and β controls the preference scale. The DPO objective itself follows standard DPO; our distinction lies in how preference data are coupled to the SFT stage. Selected grounded SFT cases are transformed into local contrasts between visually supported and plausible but unsupported continuations, concentrating second-stage supervision on difficult grounding boundaries while retaining the SFT checkpoint as both initialization and reference policy. The resulting procedure is **broad multimodal SFT $\rightarrow$ hard grounding preference construction $\rightarrow$ grounding-focused DPO**.

3 Experimental Results and Analysis

We evaluate the proposed approach along two complementary dimensions: (i) whether hard grounding preference supervision improves the Stage-1 SFT model on challenging multimodal responses, and (ii) whether this targeted decoder-side refinement preserves broader multimodal capabilities. For general capability, we evaluate DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), MMBench (Liu et al., 2024), and VQAv2 (Antol et al., 2015). For grounding-sensitive evaluation, we compare Stage-1 SFT and Stage-2 DPO responses on approximately 839 diverse multimodal examples using three independent LLM judges: GPT-4o, Gemini-2.5-Flash, and LLaMA-4-Maverick. The judges consider visual grounding, correctness, informativeness, and unsupported content (Zheng et al., 2023). We additionally compare against Qwen2.5-VL-7B and InternVL3-8B on standard benchmarks and use selected proprietary-model outputs only as qualitative cross-model case studies.

3.1 Quantitative Results

Performance on Standard Multimodal Benchmarks: Table 1 examines whether targeted preference refinement preserves the broader capabilities acquired during SFT. The controlled comparison of primary interest is between the Stage-1 SFT and Stage-2 DPO checkpoints, which share the same frozen visual pathway and differ only in decoder-side post-training.

Table 1: Performance comparison on standard multimodal benchmarks.

Model	DocVQA	TextVQA	MMBench	VQAv2
Qwen2.5-VL-7B	**87.08**	72.77	**82.60**	67.81
InternVL3-8B	75.02	63.35	82.49	61.86
LLaMA-3 MM SFT	81.5	80.5	78.0	81.5
LLaMA-3 MM DPO	83.0	**81.6**	80.1	**82.2**

Evaluation details: Validation splits are used because official test-set annotations are not publicly available. Evaluation sizes are DocVQA validation (5,349), TextVQA validation (5,000), MMBench development (dev, 4,329), and VQAv2 validation (47,206 questions). DocVQA is evaluated using ExactMatch-Norm for all models to maintain a consistent metric rather than ANLS. Complete evaluation scripts and benchmark implementations are provided in the anonymous repository.

The DPO checkpoint improves over its SFT initialization on all four benchmarks: 81.5→83.0 on DocVQA, 80.5→81.6 on TextVQA, 78.0→80.1 on MMBench, and 81.5→82.2 on VQAv2. Thus, refinement concentrated on difficult grounding examples does not appear to trade away broader multimodal capabilities. Qwen2.5-VL-7B remains stronger on DocVQA and MMBench, consistent with its strong document/OCR capabilities and MMBench's constrained multiple-choice evaluation, whereas our DPO checkpoint is stronger on the more open-ended TextVQA and VQAv2 under this protocol. Since the external models differ in architecture and training, these comparisons are contextual; the controlled result is the consistent DPO improvement over SFT.

Pairwise Grounding Evaluation: We next directly compare Stage-2 DPO with its Stage-1 SFT checkpoint on approximately 839 multimodal examples. For each image–query pair, responses from the two checkpoints are evaluated independently by GPT-4o, Gemini-2.5-Flash, and LLaMA-4-Maverick.

Table 2: Pairwise comparison of Stage-1 SFT and Stage-2 DPO responses across three independent LLM judges on approximately 839 multimodal examples.

Metric	GPT-4o	Gemini-2.5-Flash	LLaMA-4-Maverick
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 (pp)	+8.2	+7.2	+6.2

All three judges favor DPO more often than SFT, with DPO-minus-SFT win-rate differences of +8.2, +7.2, and +6.2 percentage points, respectively. The consistent direction across three evaluators reduces dependence on any single judge.

Human validation: We manually review approximately 3% of judge-agreement cases together with judge-disagreement cases. This audit checks the reliability of the automated pairwise judgments and should not be interpreted as an independent large-scale human evaluation.

Together, the benchmark and pairwise results provide complementary evidence that targeted decoder-side preference refinement improves judged grounding and response quality without degrading broader multimodal performance. These results use approximately 50k–120k targeted preference examples. Because preference difficulty and dataset size are not independently varied, the current experiments do not isolate the contribution of hardness from that of data scale.

Grounded Informativeness and Response Detail: Reducing hallucination should not come at the cost of uniformly shorter or less informative responses. We therefore apply length-aware resampling to increase the representation of informative longer responses. For a response of length l, we assign the sampling weight $w(l) = \exp[\beta(l - \mu)]$, where μ is the mean response length and β controls the strength of length-aware reweighting; the weights are capped to avoid over-amplifying extreme long-tail responses. Figure 5(0) shows the response-length distribution before and after this procedure. In addition, selected short responses are expanded with visually supported attributes, spatial relationships, and contextual details.

Figures 5(1–4) provide representative SFT→DPO comparisons. DPO can retain or add query-relevant detail when supported by the image, including fine-grained scene attributes, spatial relationships, and visible contextual information. Together, the distributional and qualitative results show that the observed reduction in unsupported specificity is not explained simply by shorter or uniformly more conservative generation: preference supervision can suppress unsupported claims while preserving or increasing supported detail.

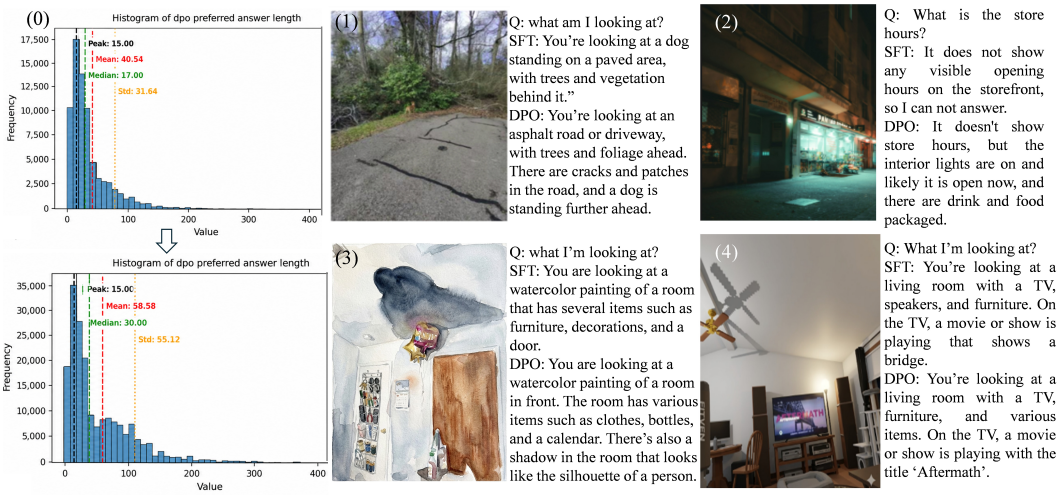


Figure 5: **Grounded informativeness and response-detail control. (0)** Preferred-response length distributions before and after length-aware resampling and response-detail expansion, which increase representation of informative longer responses rather than favoring uniformly short answers. **(1–4)** Representative SFT→DPO comparisons showing that preference optimization can retain or add visually supported, query-relevant details, including fine-grained scene attributes, spatial relationships, and contextual information. Together, these results complement the hallucination-reduction examples by showing that improved grounding does not arise merely from shortening responses or suppressing specificity indiscriminately. The first image was captured by the author; the second is from Krüger (2024, Unsplash); the third and fourth are AI-generated.

3.2 How Grounding Behavior Changes after DPO

The quantitative results establish an aggregate improvement; we next examine *how* behavior changes under the same frozen visual pathway in terms of evidence utilization, calibration, and grounding consistency. The desired behavior is bidirectional: suppress unsupported specificity when evidence is insufficient while preserving supported specificity when useful visual evidence is available.

Supported versus Unsupported Specificity: Figure 6 presents representative SFT→DPO changes selected from the larger qualitative analysis. When the image does not provide sufficient evidence—for example, when the total number of book pages or a fixed product price is not visible—the DPO model avoids specific unsupported values produced by SFT. The examples also extend beyond numerical claims: DPO corrects an unreliable visual interpretation in the painting example and avoids unsupported identity and biographical details in the memorial example. Together, these cases show that hard grounding preference optimization can reduce plausible-but-unsupported commitments across numerical, recognition, and semantic grounding failures rather than merely altering response length or style.

Remaining Failure Modes: Hard grounding preference supervision does not eliminate hallucination; residual errors include difficult spatial reasoning and occasional unsupported object insertion. Such errors do not by themselves reveal whether the limiting factor is the fixed representation or decoder use of that representation. Accordingly, our results show that decoder-side adaptation is sufficient to improve a meaningful subset of grounding failures, not that the decoder is the sole source of multimodal hallucination.

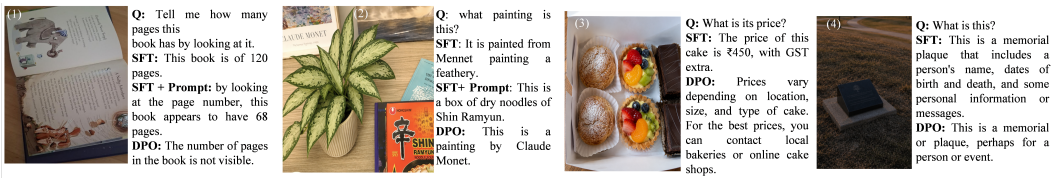


Figure 6: Representative SFT→DPO changes in visually grounded generation. Across the four examples, DPO reduces unsupported commitments made by the Stage-1 SFT model, including an unverifiable book-page count, an unreliable visual identification, an unsupported product price, and unsupported details about a memorial. These examples illustrate improvements across numerical, recognition, and semantic grounding failures. The first image was captured by the author; the second is from Krüger (2024, Unsplash); the third and fourth are AI-generated.

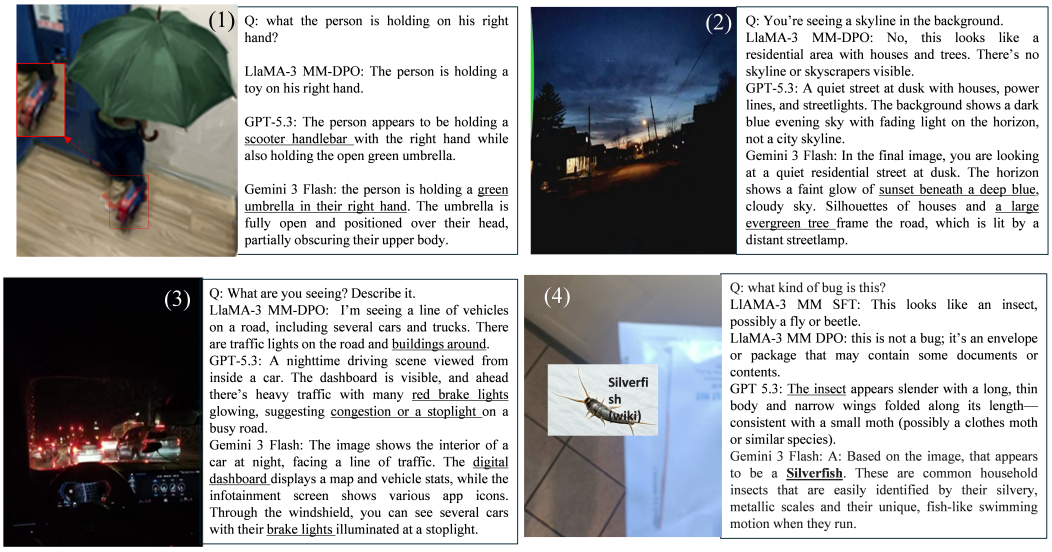


Figure 7: Cross-model qualitative examples spanning complementary grounding challenges: **(1)** physical and spatial grounding under partial visibility, **(2)** interpretation under severely limited visual evidence, **(3)** fine-grained grounding in a crowded traffic scene, and **(4)** false-premise grounding. Across these examples, the models differ in their use of available visual evidence and in the extent to which plausible but unsupported interpretations are introduced or elaborated. These cases illustrate grounding challenges involving evidence utilization, evidence calibration, spatial reasoning, and unsupported generation; the comparison is diagnostic rather than a ranking of models.

4 Conclusion

We study multimodal hallucination through hard grounding preference supervision, where preferred and rejected responses are semantically close but differ in their support from visual evidence. Under a frozen visual encoder and alignment pathway, SFT establishes broad decoder-side multimodal behavior, while targeted DPO refines evidence utilization, calibration, and grounding consistency. This design targets both unsupported specificity and unnecessary rejection rather than simply encouraging more conservative generation.

The DPO model improves over its SFT initialization across DocVQA, TextVQA, MMBench, and VQAv2 and achieves 6.2–8.2 percentage-point higher pairwise win rates across three independent judges on approximately 839 additional multimodal examples. These results show that decoder-side adaptation alone can improve a meaningful subset of grounding failures when useful visual evidence is available through the fixed multimodal representation. We release HardVQA-DPO to support further study of hard grounding supervision and visually grounded multimodal generation.

GENERATIVE AI USAGE DISCLOSURE

Generative AI tools were used in this work for synthetic data construction and model-based evaluation, as described in the corresponding methodology and evaluation sections. Generative AI tools were also used to assist with literature discovery and to partially draft, edit, and improve the clarity of portions of the manuscript. All AI-assisted content and references were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Eun Chang, Zhuangqun Huang, Yiwei Liao, Sagar Bhavsar, Amogh Param, Tammy Stark, Adel Ahmadyan, Xiao Yang, Jiaqi Wang, Ahsan Abdullah, Giang Nguyen, Akil Iyer, David Hall, Elissa Li, Nicolas Scheffer, Ahmed Kirmani, Babak Damavandi, Rakesh Wanga, Anuj Kumar, Rohit Patel, Seungwhan Moon, and Xin Luna Dong. WearVQA: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307, 2017.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. doi: 10.1109/CVPR52733.2024.01356.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, volume 15064, pp. 216–233. Springer, 2024.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1314–1332. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.98.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Fatemeh Pesaran Zadeh, Yoojin Oh, and Gunhee Kim. LPOI: Listwise preference optimization for vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26830–26844, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

APPENDIX

A HARDVQA-DPO CONSTRUCTION EXAMPLES

The following examples provide representative end-to-end illustrations of the HardVQA-DPO construction process, including the original SFT question and response, the harder grounding-targeted DPO question, the corresponding preferred and rejected responses with rationales, and human-review metadata. They cover multiple grounding behaviors discussed in the main paper, including under-utilization of subtle visual evidence, unnecessary rejection, unsupported inference, false-premise handling, and grounding-inconsistent reasoning. These behaviors are not mutually exclusive and occur across conventional tasks such as counting, object existence, attributes, and spatial reasoning.

A subset of these examples is also presented in condensed form in Figure 4 of the main paper, where only the grounding-relevant DPO preference pairs are shown for clarity. We retain the complete examples here to document the original SFT question–answer pairs, full preference rationales, and associated human-review metadata, providing additional transparency into the construction and quality-control process.

All images shown in this appendix are sourced from the publicly available WearVQA dataset (Chang et al., 2025), hosted on Hugging Face under the Apache-2.0 license. WearVQA contains egocentric real-world images designed for visual question answering in wearable-device scenarios.

#1
dpo_images/pix_4aff0bc68c36ea977316d3c0d76481e4_400.jpeg



category: object_counting

sft_question: What is the person in the foreground holding?

sft_answer: A fan of playing cards.

dpo_question: How many cards are currently in my hand?

preferred: 5

The foreground player's hand shows a fan of cards with visible numbers including 12, 5, 6, 9, and at least one more card, totaling approximately 5 cards.

rejected: 4

Four cards are clearly visible in the foreground hand: 19, 12, 7, and 6.

human review: the 5th one is occluded by another card in the foreground

#2
dpo_images/30600b4e48ddb3228b9467d38d76d308_456.png



category: object_counting

sft_question: What stands out in the foreground at the bottom of the frame?

sft_answer: A white longhorn or bull skull.

dpo_question: Looking at the bottom of the frame, how many animal skulls are visible in the foreground?

preferred: 2

Two animal skulls are clearly visible in the foreground: one is centered and the other is partially visible on the left edge.

rejected: 1

There is clearly one white longhorn/bull skull centered prominently in the foreground. On the far left edge, there appears to be a partial horn or edge of something, but it is not clearly a second distinct skull - it looks more like the same skull's horn extending to the left, or a decorative element. Only one complete animal skull is definitively visible.

human review: the left skull is partially visible

#3
dpo_images/pix_pix_fbeede6796dad86460709ce8d93d9842_115.jpeg



category: object_attribute

sft_question: What is the metal grate at the bottom center of the sidewalk for?

sft_answer: It is a storm drain for water management.

dpo_question: The metal grate at the bottom center of my view is designed to provide light to an underground basement.

preferred: no

The grate at the bottom center is a storm drain, typical for water management, not a light well; its design and placement on the sidewalk support this.

rejected: no

The bottom center of the image shows a sidewalk/street surface with people's feet and a stroller, but no clearly visible metal grate is identifiable. The image is blurry and crowded with pedestrians; no storm drain or grate is discernible at the bottom center.

human review: The metal gate is in a hidden position on the ground among crowds but still partially visible.

#4
dpo_images/pix_pix_fbeede6796dad86460709ce8d93d9842_115.jpeg



category: nonexistent_object

sft_question: Describe what is visible on the sidewalk in front of me.

sft_answer: People in winter coats standing on a crowded urban sidewalk.

dpo_question: Is there a small animal, like a rat, visible peering out from the storm drain on the ground in front of me?

preferred: no

No animal is visible near or inside the storm drain grate; only people's feet and pavement are seen.

rejected: no

The image shows a crowded urban sidewalk with people in winter coats. No storm drain grate is clearly visible in the image - the ground is largely obscured by people's feet and legs, and no grate or drain is identifiable.

human review: Question is advereral to ask an non-exit object and rejeted answer is not direclty related to the question

#5
dpo_images/b377df0944529f1ab950e2a59244e1bb_450.png



category: spatial_relationship

sft_question: What kind of path is visible in this view?

sft_answer: A cobblestone path.

dpo_question: If I am walking North, the path turns East.

preferred: unanswerable

The image is taken from a tilted egocentric perspective (rotated roughly 90 degrees), making it impossible to reliably determine cardinal directions or the curvature of the path. The cobblestone path and walkway are visible, but no compass reference, map, or clear directional cue exists to confirm whether the path turns East or West when walking North.

rejected: no

The path curves to the left (West) from the perspective of someone walking North, not to the right (East).

human review: rejected answer assumes the left is west and the right is east without giving that the person is facing the north, which can not be told

#6
dpo_images/ted_sn_pb_47.jpeg



category: object_counting

sft_question: What am I holding?

sft_answer: A photo strip showing dogs.

dpo_question: In the strip of photos I am holding, how many dogs can be seen in the fifth frame from the left?

preferred: 2

The fifth frame from the left in the strip of photos clearly shows two dogs standing together, while the previous frames each show one dog.

rejected: 1

The image is rotated/upside-down and the photo strip is small and partially obscured, making it very difficult to count dogs per frame precisely. The strip appears to show approximately 5-6 frames with dogs, but the fifth frame from the left (rightmost visible area) appears to show a single dog or is too unclear to confirm two dogs.

human review: Two dogs can be seen on the 5th image though all other images have only one dog

#7
dpo_images/pix_2f779fe43d9f4b68a7b5e77255a7b06b_1400.jpeg



category: object_counting

sft_question: What lighting fixtures are mounted on the ceiling?

sft_answer: Track-mounted spotlights.

dpo_question: How many spotlights are in the row closest to the top of the image?

preferred: 7

The uppermost row of track-mounted spotlights on the ceiling shows approximately 7 visible light fixtures when counted carefully across the image, though the fisheye distortion makes an exact count difficult.

rejected: 6

The uppermost row of spotlights on the ceiling contains exactly six visible light fixtures.

human review: the 7th light on the right end is vague but still accountable given the rational

#8
dpo_images/b9c7695319ace6cc.jpg



category: object_counting

sft_question: What package is inside the open holiday gift box?

sft_answer: A red and white K. Minamoto package with strawberry illustrations.

dpo_question: In the open holiday gift box in front of you, how many whole strawberries are illustrated on the red and white rectangular package (labeled K. Minamoto)?

preferred: 5

The red and white rectangular package labeled 'K. Minamoto' clearly shows 5 whole strawberries illustrated on its surface.

rejected: 3

The red and white rectangular package labeled K. Minamoto inside the open holiday gift box shows strawberry illustrations. Examining the visible package, approximately 3 whole strawberries are depicted on the front face. The image resolution makes an exact count difficult, but 3 is more consistent with what is visible.

human review: 5 straberies can been including one harder to tell due to light reflection effect

#9
dpo_images/268bd0fd2dcb383f6640c604ed343734_40.png



category: object_counting

sft_question: What trash bins can you see along the sidewalk?

sft_answer: A silver trash bin on the right sidewalk and another bin near the crosswalk.

dpo_question: There is only one trash bin clearly visible on the sidewalk.

preferred: no

Two trash bins are visible: one silver bin on the right sidewalk and another bin further down near the crosswalk.

rejected: yes

One prominent silver/white cylindrical trash bin is clearly visible on the right sidewalk. A possible second dark object near the crosswalk area is too ambiguous to confidently identify as a trash bin.

human review: two trash bins 'blended' toether but can be distinguished from the spatial relationship

#10
dpo_images/31b4dd4b5b81f16cab92ea1d1bb14e24_820.png



category: object_counting

sft_question: What is set up along the walkway ahead?

sft_answer: Scaffolding.

dpo_question: Looking ahead under the scaffolding, how many fire hydrants are visible to your left?

preferred: 1

A single fire hydrant is clearly visible on the left side near the planter; no others are visible along the walkway.

rejected: 0

Examining the left side of the image carefully, there is a planter with vegetation, parked cars, and scaffolding poles, but no fire hydrant is clearly visible.

human review: A grey fire hydrant can be found on the left though it is partially hiden by the first metal column and blended into the car behind

#11
dpo_images/388d85b8dc5899d8d7ba1005232ee06d_400.jpeg



category: object_counting

sft_question: What is attached to the front of the box on the right?

sft_answer: A key protruding from its lock.

dpo_question: How many keys are sticking out of the front of the box on the right?

preferred: 1

The box on the right has a single key visibly protruding from its front lock; no second key is present.

rejected: unanswerable

The decorative box on the right side of the display case appears to be an ornate Chinese cloisonné or enamel box with handles on the sides. The image resolution and angle make it very difficult to confirm any key or lock mechanism protruding from the front, No key is clearly visible

human review: One key can be identified on the front of the box on the right though it is not easy to tell

#12
dpo_images/ef37bff6a21da50e.jpg



category: object_counting

sft_question: What brand of cooling products is on the shelves in front of me?

sft_answer: Corsair.

dpo_question: How many white Corsair radiators are visible on the shelves in front of me?

preferred: 3

I can identify white Corsair radiators in the image: one on the top shelf (right side), one in the middle shelf (the Corsair H100i RGB PLATINUM SE white box with a clearly white radiator visible), and one on the bottom shelf area (Corsair Hydro Series H100i RGB PLATINUM SE on the right side of the lower shelf also appears to show a white/light radiator). The H100i RGB PLATINUM on the lower left appears to show a darker/black radiator. Counting is difficult due to image angle and partial visibility, but at least 2-3 white radiators are visible.

rejected: 2

There is one white Corsair radiator on the top shelf and one on the middle shelf; all other visible Corsair radiators are black.

human review: 2 objects are apparent but the 3rd one can still be judged given the reasonable rationals

#13
dpo_images/up-75211492-3fcf-4bfb-a54e-ee7c585dae2c_1003.png



category: nonexistent_object

sft_question: What is sitting on the table in front of me?

sft_answer: A clear glass vase with flowers.

dpo_question: Is there a blue coffee mug sitting behind the green candle in front of me?

preferred: no

Behind the green candle is a clear vase with flowers, no blue coffee mug is visible.

rejected: no

The image shows a clear glass vase with yellow, pink/purple flowers and branches on a desk/table. There is no green candle visible anywhere in the image.

human review: Both DPO answers said No, but the rejected answer didn't recognize the candle

#14
dpo_images/pix_af33b166f00dfd9a301c7e3fceeb9f5d_212.jpeg



category: object_counting

sft_question: What seating is visible to my right?

sft_answer: Benches.

dpo_question: Each bench to my right is supported by exactly four legs.

preferred: no

The benches to the right show three visible legs on the front side, indicating more than four legs per bench and likely six

rejected: unanswerable

The benches on the right side of the image are visible but partially obscured by darkness and fairy lights. While multiple benches are present, the leg configuration is not clearly discernible from the image - the lighting conditions and angle make it impossible to confidently count the legs and verify whether they have four, six, or any other number of legs.

human review: The truth is six legs and rejected answer is uncertain which is less aligned with the human perception and physical grounding

#15
dpo_images/d34a2711da5038d0abd4b79fcf2366a0_300.png



category: object_counting

sft_question: Describe the dog on the floor in front of me.

sft_answer: A small dog wearing a reindeer-like costume.

dpo_question: How many dogs are visible in front of me?

preferred: 2

One dog is being petted in the foreground, and a second smaller dog is standing on the floor near the table.

rejected: 1

The image shows one small dog wearing a costume (appears to be a reindeer or similar outfit) standing on the tile floor near the table. The wearer's hands are visible in the foreground but there is no clear second dog - only one dog is identifiable in the scene.

human review: The dog in the front is very hard to tell due to the paper cover on its back and however the exposed feets with hairs help judtify

#16
dpo_images/89d89e776a3ae6d5.jpg



category: object_counting

sft_question: What product box is hanging on the center hook in the middle row?

sft_answer: An Anker PowerDrive PD+ 2 box.

dpo_question: How many Anker PowerDrive PD+ 2 boxes are hanging on the hook in the middle row, center column?

preferred: 4

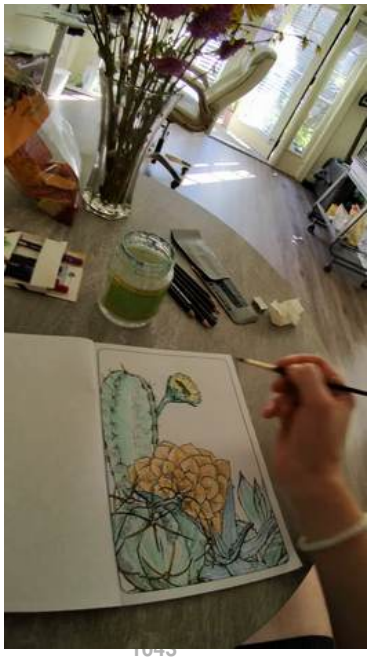
There are four distinct Anker PowerDrive PD+ 2 boxes visible stacked on the center hook in the middle row.

rejected: 3

The center hook in the middle row shows the Anker PowerDrive PD+ 2 box (blue, labeled '3X FASTER FOR THE NEW IPHONE' with PD badge). From the side profile visible in the image, approximately 3 boxes appear to be stacked on that hook, not 4. The stack depth is difficult to determine precisely but 3 is the best estimate from visible evidence.

human review: 4 boxes can be estiamted give two in front, and anoher two stacked like one but according to its depth/thickness it shall be total 4

#17
dpo_images/up-75211492-3fcf-4bfb-a54e-ee7c585dae2c_1239.png



category: object_counting

sft_question: What objects are on the table around the writing supplies?

sft_answer: A green-labeled jar, a grey pencil pouch, and dark pencils between them.

dpo_question: Looking at the table in front of me, how many dark pencils are lying between the jar and the grey pouch?

preferred: 5

Between the green-labeled jar and the grey pencil pouch, there appear to be approximately 5 dark pencils grouped together on the table surface. The exact count is somewhat difficult due to image angle, but 5 is more consistent with what is visible than either 3 or 4.

rejected: 4

There are four black pencils clearly visible between the jar and the grey pouch on the table.

human review: the pensiles are crowed and blended togehter but 5 pensiles can still be identified

#18
dpo_images/up-fa7a37f5-2583-4689-80b8-b057e6cc33e1_328.png



category: spatial_relationship

sft_question: Describe the outdoor scene with the dog and the building.

sft_answer: A dog is running on grass in front of a two-story red/brown apartment building.

dpo_question: True or false: The parking structure is located between the dog and the red building.

preferred: false

The image shows a two-story red/brown apartment building, a dog running on green grass in the foreground, a palm tree, and what appears to be a couple of parked cars to the right of the building. There is no clearly identifiable parking structure (multi-level garage or dedicated parking facility) visible - only a few parked cars in a surface lot area.

rejected: false

The parking structure is visible to the right of the red building, not between the dog and the building. The dog is on the grass in front of the building.

human review: Both answers are No, but there is no solid evidence to demostrate that there is a parking lot, the rejected answer is too deterministic

#19
dpo_images/D1ADA1FD-6DF6-445D-A173-CFA3815C0E2E_1_105_c.jpeg



category: object_attribute

sft_question: What is inside the jar with the red lid?

sft_answer: Pieces of tofu.

dpo_question: Was the tofu in the red-lidded jar fermented as a single whole block before being sliced into the pieces seen inside?

preferred: no

The jar with the red lid contains visible tofu pieces, not a single block, consistent with typical fermented tofu production where tofu is cut before fermentation.

rejected: It cannot be determined from visual inspection alone - the internal manufacturing process is not visually verifiable.

human review: the preferred answer is accurate by providing the truth which is not revealed from the image itself, while the rejected answer is not wrong given the visual information alone

#20
dpo_images/pix_10b90cdad5a146f2aa513795ea7ba4f2_290.jpeg



category: nonexistent_object

sft_question: What is the person holding in their right hand?

sft_answer: A glass.

dpo_question: Is the person holding a smartphone in their right hand?

preferred: no

The person is holding a glass in their right hand, not a smartphone.

rejected: no

The image is blurry but the person appears to be holding a dark rounded object (possibly a cup/glass) in their left hand (from viewer's perspective, the hand on the left side of the frame). The right hand rests near a small cactus figurine on the table. The object in the left hand does not resemble a smartphone.

human review: The rejected asnwer seems not wrong but not plausible, e.g. it can not tell the object relationship such as the left or right of the hand related to the human

B Additional Cross-Model Qualitative Analysis

This appendix provides additional analysis of the cross-model examples introduced in the main text. The purpose is not to establish a statistically representative ranking of model families, but to examine how different systems respond when visual evidence is ambiguous, incomplete, or insufficient for a unique answer. These cases complement the controlled SFT–DPO comparison by illustrating how similar grounding difficulties can appear across models with different architectures and training procedures.

B.1 Detailed Analysis of the False-Premise Example

We first revisit the false-premise bug example discussed in the main text. There, the example illustrates a general cross-model behavior: when a question introduces an unsupported premise, different MLLMs can make different plausible commitments and subsequently elaborate on them. Here we examine the individual responses in greater detail from the perspectives of evidence utilization, response specificity, and premise following.

The responses reveal two related effects. First, different models assign different specific interpretations to the same ambiguous visual evidence. The disagreement itself does not identify the source of the error, but it is consistent with model-specific linguistic and contextual priors influencing generation when the image does not strongly determine the answer. Second, after making an initial identification, a model can generate additional details that are coherent with that identification even when the initial premise is unsupported. Thus, the problem is not simply verbosity: an autoregressive continuation can be internally coherent while remaining grounded in an unsupported initial commitment.

The DPO-trained model instead rejects the assumed object in this example. Because its ViT encoder and Perceiver alignment module are unchanged from the SFT checkpoint, the SFT–DPO behavioral difference is consistent with decoder-side post-training changing how the fixed multimodal evidence is used. This example does not establish that the original failure arose exclusively in the decoder, and the responses of external models cannot localize their failures to a particular component because their multimodal architectures differ.

B.2 Detailed Cross-Model Analysis of Grounding Behavior

The remaining examples introduced in the main text span unsupported object inference, spatial reasoning under partial occlusion, ambiguous visual conditions, OCR and fine-grained recognition, and contextual interpretations that extend beyond directly observable evidence. Here we examine these cases through the three complementary grounding dimensions used throughout the paper: evidence utilization, evidence calibration, and grounding consistency.

Evidence utilization: When relevant evidence is small, partially visible, crowded, or otherwise difficult to perceive, responses can fail by under-utilizing information that is nevertheless present in the image. This can appear as an incorrect count, a missed object or attribute, or an unnecessarily generic answer. Importantly, the opposite behavior is also possible: a model may reject a question even when subtle but sufficient evidence supports a specific answer. These cases motivate preference pairs that distinguish difficult evidence from genuinely insufficient evidence.

Evidence calibration: When the image is underdetermined, models differ in how much specificity they introduce. Some responses provide plausible prices, object identities, attributes, spatial relations, or contextual interpretations that cannot be verified from the image. Other responses appropriately reduce specificity or state that the required information is unavailable. Greater detail is not itself a grounding failure: several examples contain richer responses that remain consistent with the image. The relevant distinction is whether the specificity of the response is supported by the available evidence.

Grounding consistency: Grounding errors can also occur within an otherwise plausible explanation. In some cases, two responses reach the same final decision while differing in whether the supporting rationale is visually correct. This motivates evaluating grounding at the level of individual claims and reasoning steps, rather than relying only on final-answer accuracy.

These behaviors should be interpreted as qualitative case studies rather than model-level performance estimates. They illustrate that the grounding–informativeness tension motivating HardVQA-DPO is not restricted to a single task category or a single model family.

B.3 Summary of Observed Grounding Behaviors

Table 3 summarizes the grounding behaviors represented by the selected examples. The examples are not sampled as a statistically representative benchmark and therefore should not be interpreted as estimates of overall model performance.

Table 3: Grounding behaviors represented in selected qualitative examples.

Behavior	**Visual condition**	**Representative response difference**
Evidence under-utilization	Small, occluded, crowded, or fine-grained evidence	Available visual evidence is missed, producing an incorrect or under-specified response
Unnecessary rejection	Difficult but sufficient evidence	The response declines to answer even though the image supports a specific conclusion
Unsupported inference	Ambiguous or insufficient evidence	A plausible object, attribute, count, price, relation, or interpretation is introduced without sufficient visual support
False-premise following	Query assumes an unsupported object, attribute, or relation	The premise is accepted and subsequently elaborated rather than rejected or qualified
Grounding inconsistency	Multi-step or explanatory response	The final decision can be correct while intermediate claims or justification conflict with visible evidence

B.4 Cross-Model Qualitative Examples and Analysis

The complete set of qualitative cross-model examples is provided on the following pages. Comparisons involving proprietary models reflect the API versions available at the time of evaluation and may change as these systems are updated. No external OCR or computer-vision toolkits were added to the evaluated systems during testing.

Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3	Analysis
 Q: What you see?	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?	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?	SFT: It is painted from Mennet painting a feathery. SFT+ Prompt: This is a box of Goodles Down the Hatch, which is a brand that makes creative and fun macaroni and cheese recipes. 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.
 (Courtesy of Amazon) Q: How is this game played?	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.

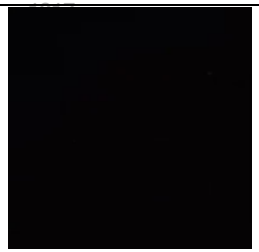
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3	Analysis
 Q: Tell me how many pages this book has by looking at it.	SFT: This book is of 240 pages. SFT + Prompt: by looking at the page number, this book appears to have 128 pages. DPO: The number of pages in the book is not visible.	Based on the image provided, the bottom right corner of the open page explicitly shows the number 140. While I can't see the very last page, it's a safe bet the book has at least that many!	It's not possible to determine the total number of pages from this image alone. The book is partially open, showing only one page and part of the cover, without any visible page numbers or thickness indicators.	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?	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 Input: 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?	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.	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	Analysis
 Q: What is its price?	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.
 (AI generated) Q: Is this teacher a good instructor?	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.	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	Analysis
 Q: You're seeing a skyline in the background.	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.	In the final image, you are looking at a quiet residential street at dusk. The horizon shows a faint glow of sunset beneath a deep blue, cloudy sky. Silhouettes of houses and a large evergreen tree frame the road, which is lit by a distant streetlamp.	A quiet street at dusk with houses, power lines, and streetlights. The background shows a dark blue evening sky with fading light on the horizon, not a city skyline.	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).
 (AI augmented) Q: What am I looking at?	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.
 (AI generated) Q: What am I looking at?	SFT: You're 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.

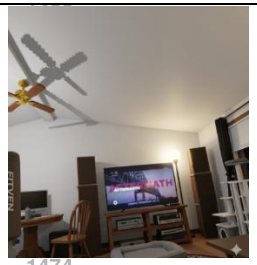
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3	Analysis
 Q: What are you seeing? Describe it	SFT: I'm seeing several cars on a road, along with traffic lights and other vehicles. DPO: I'm seeing a line of vehicles on a road, including several cars and trucks. There are traffic lights on the road and buildings 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.
 (AI generated) Q: What am I looking at?	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.
 (AI generated) Q: What am I looking at?	SFT: You're looking at a room that has several items such as furniture, decorations, and a door. DPO: It is a water color painting. You're looking at a room with 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.

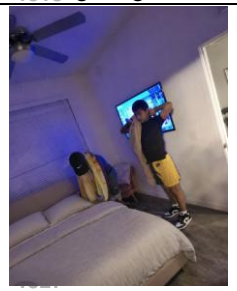
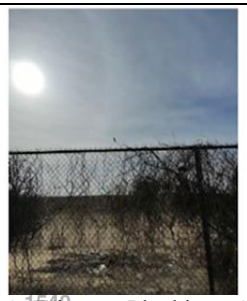
Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3	Analysis
 (AI generated) Q: Tell me what I'm looking at right now?	SFT: You're looking at a room with wooden flooring, curtains, and various furniture. DPO: You're looking a painting of room. The room with wooden flooring and curtains on the walls. The room also 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 least detailed.
 (AI augmented) Q: What I am seeing?	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 three models exhibit errors in spatial reasoning. LLaMA-3 MM (DPO) and GPT-5.3 infer that the right person is facing inward or watching the TV, which is inconsistent with the physical orientation of the body (e.g., the feet are directed outward). Gemini 3 Flash instead describes the left person as "sitting" on the bed, which also conflicts with the visual evidence, as the person appears positioned beside the bed and oriented downward toward it.
 Q: What am I looking at?	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 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.

Image	LLaMA3-MM	Gemini 3 Flash	GPT-5.3	Analysis
 (courtesy of Operation Homefront) Q: read the text	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) Underlined text indicates hallucinated content. (2) Because some models tend to generate excessively long responses, evaluation prompts constrain responses to approximately 50--60 words or fewer for more comparable outputs. (3) Images were captured by the authors or AI-generated, as indicated. (4) Due to stochastic generation, differences in image processing and resolution, and other nondeterministic inference factors, exact model responses may vary across runs.

C Additional Analysis

C.1 Data Scale and Preference Refinement

Stage-2 training uses approximately 50k–120k targeted preference examples, substantially fewer than the supervised examples used in Stage 1. Across the evaluated settings, the DPO-trained model consistently receives higher pairwise preference rates than its SFT initialization.

The preference data concentrate on difficult grounding boundaries rather than a single task category. They include under-utilization of subtle visual evidence, unnecessary rejection, unsupported inference, false-premise following, and grounding-inconsistent reasoning across counting, OCR, attributes, spatial relations, and object-existence questions. Preferred and rejected responses are constructed to share substantial semantic content so that their distinction depends primarily on fine-grained grounding rather than easily separable differences in linguistic quality.

The experiments show that targeted preference refinement provides an effective training signal at this scale. However, preference difficulty and dataset size are not independently varied. We therefore do not attribute the observed gains specifically to hard-example selection rather than additional targeted preference data more generally. A matched-scale comparison across preference-construction strategies would be required to isolate these effects.

C.2 Supported and Unsupported Specificity

A central objective of our construction is to avoid conflating improved grounding with generic conservatism. More descriptive responses contain more objects, attributes, relations, and contextual claims and therefore create additional opportunities for unsupported generation. However, simply shortening responses or rejecting difficult questions would also discard useful information.

Our preference construction consequently operates in both directions. When the image is insufficient to support a specific conclusion, the preferred response avoids or qualifies unsupported detail. When subtle but sufficient evidence is available, the preferred response retains the specific answer and can include additional supported information. In the qualitative examples, DPO suppresses unsupported page counts, prices, and premise-driven interpretations while recovering additional supported OCR content and visual details in other cases.

These opposite-direction changes are important because they are inconsistent with a simple explanation in which DPO reduces hallucination merely by making responses uniformly shorter or more conservative. Instead, the observed behavior is consistent with better matching response specificity to the available visual evidence. We do not interpret these selected examples as a formal measurement of the detail–grounding trade-off.

C.3 Limitations and Scope of Decoder-Side Adaptation

Hard grounding preference supervision reduces but does not eliminate unsupported generation. Figure 8 provides one illustrative residual case in which the DPO-trained model introduces a bird sitting on the fence despite insufficient visual support for that object. This example is not intended to characterize the full space of residual errors, but demonstrates that plausible unsupported content can persist after preference optimization.

Because the ViT encoder and Perceiver alignment module remain frozen, decoder-side post-training can improve grounding when task-relevant information is available through the fixed multimodal representation but is under-utilized or incorrectly interpreted by the decoder. If required visual information is absent from that representation, decoder-side adaptation cannot reliably recover it. The residual example does not distinguish these cases; our controlled SFT–DPO comparison instead establishes decoder-side sufficiency for correcting a subset of grounding failures.



Figure 8: **Illustrative residual hallucination after DPO.** The DPO-trained model introduces an unsupported bird on the fence, demonstrating that hard grounding preference supervision reduces but does not eliminate plausible unsupported generation.