When We Don't See the Same Picture: Aligning Agents with Divergent Visual Spaces

Muhammad Gul Zain Ali Khan [1] **Stephan Alaniz** [2 1] **Eric Schulz** [1] **Zeynep Akata** [2 1]

Abstract

An estimated 2.2 billion people worldwide live with some form of vision impairment. However, while modern models perform well for the general public, they often struggle to communicate effectively with agents under divergent visual perceptions. In this work, we seek to study how we can aligns under divergent visual space using Reinforcement Learning. Because simulating such interactions in the real world is costly, we adopt image reference games as a controlled testbed and design experiments with five distinct perceptual distortions inspired by real human visual impairments (e.g., cataract, color blindness). We evaluate in two settings, online and offline, with four post-training algorithms: SFT, DPO (offline) and KTO, GRPO (online), providing the first systematic study of alignment under divergent visual perceptions. Our results reveal that (i) offline adaptation can provide strong improvements, with DPO consistently outperforming other methods when supported by high-quality preference data; (ii) Online methods can provide a robust alternative in absence of preference dataset and among online adaptation methods, GRPO shows more consistent gains, and (iii) qualitative analysis shows that adapted agents align their descriptions toward perceptual features accessible to their conversation partners. We release our code and dataset.

1. Introduction

Effective communication between agents with differing perceptual experiences requires the adaptation of communication strategies to align with a partner's unique perceptual space. This challenge is particularly pronounced when interacting with individuals who have visual impairments such as color blindness or macular degeneration. Descriptions that are clear under typical vision may become ambiguous or ineffective under altered perception.

This challenge extends to multimodal AI systems, where assistive agents and collaborative multi-agent systems must often communicate across divergent perceptual channels. However, current multimodal post-training pipelines implicitly assume shared perceptual inputs. Despite their strong performance for the general population, modern models lack the robustness to communicate effectively under perceptual divergence. This limitation is particularly important given that an estimated 2.2 billion people worldwide live with some form of vision impairment or blindness according to the World Health Organization (WHO) (World Health Organization & The Lancet Global Health Commission on Global Eye Health, 2021). Enabling models to adapt their communication to perceptually diverse users is therefore an important step toward equitable and human-centric AI.

Studying such adaptive communication directly in the real world is prohibitively expensive. It would require either (i) collecting large-scale annotations from individuals with diverse perceptual profiles, or (ii) training agents through hundreds of interactive communication rounds for each objective function and training recipe. Since different optimization objectives can lead to substantially different behaviors (Pan et al., 2025), systematically evaluating them would necessitate repeated engagement with visually impaired users, an impractical and costly process. This creates a major bottleneck for principled development of alignment strategies under perceptual mismatch.

To address this challenge, we propose a controlled and reproducible testbed based on an *image reference game under perceptual divergence*. In our setting, a *speaker* observes undistorted images and generates a description of a target image, while a *listener* receives distorted versions of the same image pair and must identify the target. The asymmetry in visual input induces a perceptual gap that the speaker must learn to bridge. Training proceeds via reinforcement learning, where agreement between speaker and listener provides a sparse binary reward signal.

To ground this problem in realistic conditions, we sim-

[1]Helmholtz Munich [2]Technical University of Munich (TUM). Correspondence to: Muhammad Gul Zain Ali Khan <gulzainali@gmail.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

ulate five perceptual distortions inspired by common human visual impairments, spanning both spatial and pixel-level degradations. Within this unified framework, we systematically compare offline and online post-training strategies for adaptation. Offline methods include supervised fine-tuning (SFT) and direct preference optimization (DPO) (Rafailov et al., 2023). Online methods include Kahneman–Tversky Optimization (KTO) (Ethayarajh et al., 2024) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024). To enable preference-based learning, we construct a distortion-aware preference dataset using Qwen2.5-VL (Wang et al., 2024b), with positive preferences taken from the view of listener and negative preferences taken from the view of speaker.

Our experiments across synthetic and natural image benchmarks reveal several key findings. First, when the distortion type is known in advance, offline adaptation can yield strong performance improvements, with preference-based optimization providing particularly effective gains. Second, in settings where perceptual divergence must be handled dynamically, online reinforcement learning methods provide robust adaptation through interaction. Third, adaptation benefits are strongly influenced by the listener's baseline capability: improvements are substantial when the listener struggles under distortion, but more limited when the listener is already near ceiling performance. Qualitative analyses further show that adapted speakers systematically shift their descriptions toward distortion-consistent and structurally robust visual cues with DPO demonstrating robust alignment by prioritizing distortion consistent attributes and reducing distortion inconsistent attributes.

In summary, this work makes the following contributions: (i) we introduce a controlled framework for studying alignment between multimodal agents under divergent visual perception; (ii) we provide the first systematic comparison of offline and online post-training objectives in this setting; and (iii) we demonstrate how different objectives trade off peak performance, stability, and robustness across perceptual distortions.

2. Aligning Divergent Agents

Game Protocol. In each interaction, the speaker observes the undistorted ordered pair of images $(x_{\text{tgt}}, x_{\text{conf}})$ and generates a message describing the target image, $m \sim \pi_\theta(\cdot \mid x_{\text{tgt}}, x_{\text{conf}})$, where π_θ is the speaker policy. The **listener** receives the corresponding distorted pair $(\hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}})$ in random order and predicts the image referred to by message m, $a \sim \pi_\phi(\cdot \mid \hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}}, m)$, where $a \in \{\text{left}, \text{right}, \text{none of these}\}$. The episode succeeds when the listener selects the distorted target image.

Let $\mathcal{D}_\mathcal{T}$ denote the distribution over clean image pairs and their distorted counterparts. Each interaction yields a binary reward $r_i \in \{0, 1\}$, where $r_i = 1$ if the listener correctly identifies the target and $r_i = 0$ otherwise. We measure *agreement between agents* as $\mathcal{A}(\theta, \phi) = \mathbb{E}_{((x_{\text{tgt}}, x_{\text{conf}}), (\hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}})) \sim \mathcal{D}_\mathcal{T},\, m \sim \pi_\theta,\, a \sim \pi_\phi}[r_i]$.

Due to asymmetric observations, messages that are informative to the speaker may rely on visual cues that are degraded or absent for the listener. We therefore define perceptual alignment as learning a speaker policy that maximizes agreement under the listener's perceptual channel: $\theta^* = \arg\max_\theta \mathcal{A}(\theta, \phi)$. Achieving alignment requires the speaker to generate distortion-robust descriptions that remain discriminative in the listener's visual space.

2.1. Perceptual Distortions

We design five perceptual distortions inspired by real human visual impairments, applied in either the spatial or pixel domain. First, age-related macular degeneration (AMD), which leads to the deterioration of the central retina and hampers fine-detail recognition while leaving peripheral vision intact, is modeled by masking the central region of the image. Second, retinal detachment, where the retina separates and creates dark, irregular blind spots often starting at the periphery or lower visual field, is simulated by overlaying irregular dark patches on the lower half of the image. Third, color blindness, a deficiency in perceiving chromatic information, is represented by removing color channels and converting to luminance-only. Fourth, tunnel vision, characterized by the loss of peripheral vision, is simulated by masking out peripheral regions to retain only a central circular window. Lastly, cataract, a condition where the eye's lens becomes clouded and vision turns blurry with reduced contrast, is simulated via Gaussian blurring coupled with contrast reduction, after which random black artifacts are added to the image. An example of these perceptual distortions can be seen in Fig. 1.

2.2. Preference Dataset Construction

To enable offline adaptation in the reference game with divergent visual spaces, we construct a preference dataset using Qwen2.5-VL-32B. The central idea is to create contrasting descriptions from conditioning on distorted versus undistorted image pairs (Fig. 1). Distorted-view descriptions serve as positive preferences, since they are aligned with the listener's perceptual limitations, while undistorted-view descriptions serve as negative preferences, as they often rely on features inaccessible under distortion. This preference framing explicitly encodes communicative success under mismatched perceptual inputs.

For each target image x_{tgt}, we randomly sample a confounding image x_{conf} from the same dataset to act as a distractor, without applying any post-processing or semantic filtering. Each pair is instantiated under both normal and distorted views, yielding the tuple

$$\big((x_{\text{tgt}}, x_{\text{conf}}), (\hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}})),\ \texttt{instruction},\ (m^+, m^-)\big),$$

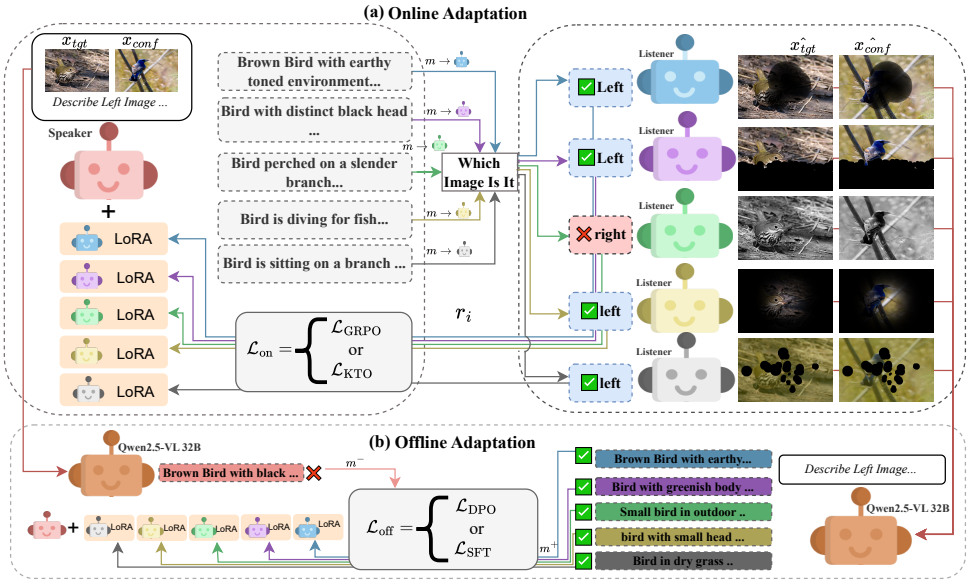


Figure 1. **Overview of the framework.** (a) The speaker, adapted with LoRA, generates descriptions for the target image and updates online via reinforcement learning (GRPO or KTO) based on listener feedback. Listeners receive the description with distorted images and predict the referred image, providing rewards to the speaker. (b) An offline data generation pipeline creates labeled descriptions for offline adaptation with supervised finetuning and direct preference optimization (Rafailov et al., 2023). LoRA modules are trained for each distortion to adapt the speaker.

where $(\hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}})$ denote their distorted counterparts.

We generate paired descriptions (m^+, m^-) with Qwen2.5-VL-32B where m^+ represents positive preference and m^- negative preference. The model is always presented with two images in order *Image 1* (target) and *Image 2* (confounding) and is instructed to describe only *Image 1*. For the distorted-view query, we provide $(\hat{x}_{\text{tgt}}, \hat{x}_{\text{conf}})$ and record the output as m^+; for the normal-view query, we provide $(x_{\text{tgt}}, x_{\text{conf}})$ and record the output as m^-. Qwen2.5-VL-32B is queried twice per instance—once with distorted inputs to produce m^+ and once with clean inputs for m^-—using identical text prompts and a fixed image ordering that always places the target as *Image 1*. Keeping the instruction prompt constant ensures that differences in responses arise solely from the change in visual input. By construction, m^+ captures distortion-consistent descriptions interpretable under impaired perception, while m^- encodes distortion-inconsistent content. This yields preference pairs labeled $m^+ \succ m^-$, providing the contrastive supervision required by preference-based alignment methods such as DPO.

The unified prompt template is:

> You see two images. Image 1: [IMG_1], Image 2: [IMG_2]. Focus on Image 1. Your goal is to describe Image 1 in a way that clearly distinguishes it from Image 2. Do not mention Image 2 in your description at all. Highlight unique features of Image 1 that differentiate it from Image 2. Ignore the occlusions in the image.

Under distortion, this template promotes robustness by encouraging features that remain interpretable despite perceptual noise, while strictly prohibiting reference to the distractor. The model thus produces paired descriptions (m^+, m^-) that encode a binary preference $m^+ \succ m^-$ consistent with distortion-aware communication, supporting preference-based optimization such as DPO.

2.3. Adapting Algorithms

We study offline and online post-training methods that are complementary in cost, stability, and feedback needs. Offline adaptation (SFT, DPO) is a cost-effective ante-hoc strategy when perceptual distortion is known in advance; SFT is a strong baseline and DPO (Rafailov et al., 2023) is a widely adopted preference-optimization method that avoids training a reward model while enforcing a KL constraint. Online adaptation (KTO (Ethayarajh et al., 2024), GRPO (Shao et al., 2024)) captures in situ personalization when adaptation to individual listeners is desired and the visual divergence is unknown. All methods are applied to the speaker policy π_θ via LoRA (Hu et al., 2022b).

For offline adaptation, we train on preference triples $\{(\mathbf{x}, m^+, m^-)\}$ where $\mathbf{x} = (x_{\text{tgt}}, x_{\text{conf}})$ is the undistorted input pair, m^+ is the positive preference, and m^- is the negative preference. All algorithms condition on $\mathbf{x}$. SFT fine-tunes the speaker directly on positive responses m^+,

providing a simple baseline for offline adaptation. DPO leverages paired preferences (m^+, m^-) to optimize directly without learning a reward model, offering stability and efficiency for offline adaptation. For online adaptation, we employ KTO and GRPO. KTO incorporates principles from behavioral economics, capturing asymmetries such as loss aversion, which makes it well-suited for online adaptation with noisy or uneven feedback. GRPO is a critic-free reinforcement learning method that normalizes rewards across sampled groups, enabling stable and scalable online alignment in interactive settings. In summary, we study two offline objectives $\mathcal{L}_{\text{off}} = \{\mathcal{L}_{\text{SFT}}, \mathcal{L}_{\text{DPO}}\}$ and two online objectives $\mathcal{L}_{\text{on}} = \{\mathcal{L}_{\text{KTO}}, \mathcal{L}_{\text{GRPO}}\}$, as depicted in Fig. 1. Details of algorithms are in Appendix.

3. Related Work

A common assumption in prior work is that agents share a consistent world-view (Ju & Aral, 2025; Hsu et al., 2025; Zhang et al., 2024). However, Human communication often requires adapting to partners with different perceptual experiences, for example in cases of impaired or limited vision.Inspired by this, recent work has introduced multimodal reference games as testbeds for studying alignment (Corona et al., 2019; Takmaz et al., 2023) between communicating agents. However, most approaches still assume identical perceptual inputs across agents, overlooking the mismatches that frequently arise in practice due to sensory limitations or hardware degradation.

Personalization in language models has long been studied, particularly in dialogue systems (Serban et al., 2015; Song et al., 2019; Zhang et al., 2019). Some Theory of mind (ToM)-based approaches (Ma et al., 2023) propose plug-and-play modules that update weights dynamically (Takmaz et al., 2023) or model listener behavior internally (Raileanu et al., 2018). While prior work has explored speaker–listener adaptation in text-only settings (Wang et al., 2024a), we extend this line to multimodal image reference games. Other personalization methods learn on large scale user dialogue histories (Ma et al., 2021; Zhong et al., 2022), whereas our focus is on both online and offline adaptation strategies.

Parameter-efficient fine-tuning enables scalable adaptation of multimodal models. LoRA modules (Hu et al., 2022a) add trainable low-rank adapters on top of frozen backbones and have inspired many extensions (Zhang et al., 2023; Lialin et al., 2023; Liu et al., 2023; Wu et al., 2024; Sheng et al., 2023; yang Liu et al., 2024). Alternative methods adapt only small subsets of weights (Ben Zaken et al., 2022; Ansell et al., 2021) or use adapters embedded within or alongside network layers (Pfeiffer et al., 2020; Sung et al., 2022; Mercea et al., 2024). In this work, we adopt widely used LoRA for adapting MLLMs.

Reinforcement learning provides another route for adaptation, especially in online settings (Snell et al., 2023; Ziegler et al., 2019; Ramamurthy et al., 2023). GRPO (Shao et al., 2024) stabilizes training by normalizing rewards within groups, avoiding the need for a critic and reducing variance. Preference-based methods such as KTO (Ethayarajh et al., 2024) and DPO (Rafailov et al., 2024) optimize directly from positive/negative feedback pairs. Related efforts (Guo et al., 2024; Liu et al., 2024) explore online adaptation via model feedback, but our work focuses specifically on aligning communication under mismatched perceptual conditions.

Visual reference identification has been studied in multimodal dialogue (de Vries et al., 2016; Ni et al., 2021; Alaniz et al., 2021; Das et al., 2016). While Corona et al. (Corona et al., 2019) examined attribute-based reference games, we broaden the scope to free-form description generation in the presence of perceptual impairments. One of our contribution is to systematically compare online and offline adaptation strategies in this setting.

4. Results

We report results on CLEVR (Johnson et al., 2017), CUB (Wah et al., 2011) and ImageNet(Deng et al., 2009) in Tab. 1, Tab. 2 and Tab.3 for maximum test accuracy and test accuracy of validation-set-picked checkpoint under greedy sampling. Test-pick accuracy represents maximum learnability of the algorithm while val-pick accuracy represents more realistic setting where the correct model needs to be picked. We report all results with greedy sampling.

On CLEVR, all adaptation methods improve over the ZS baseline. KTO (Ethayarajh et al., 2024) shows the strongest overall improvements, with a remarkable +42% gain under grayscale (maximum) and a stable +6.4% average improvement under validation-set-picked checkpoints. GRPO (Shao et al., 2024) produces smaller peak gains (+13% on grayscale) and shows modest validation-set-picked checkpoint's performance (+2.8% average), making it conservative but reliable. Among offline methods, DPO (Rafailov et al., 2023) achieves the highest average maximum gain (+10.8%) but collapses to only +0.8% under validation-set-picked checkpoint's evaluation, revealing instability. SFT attains strong peaks on AMD (+15%) and tunnel vision (+11%), but its validation-set-picked checkpoint's scores collapse on cataract (–11%), dragging its average below baseline (–1.2%).

On CUB, improvements are smaller. KTO again proves robust, yielding strong gains on cataract (+18.0 maximum, +13.0 for validation-set-picked checkpoint) and detached retina (+5.0), leading to average improvements of +6.4% (maximum) and +5.0% (validation-set-picked checkpoint). GRPO is highly competitive, achieving the highest overall averages (+7.0% maximum, +4.6% validation-set-picked checkpoint), with especially strong performance on cataract (+19.0 / +15.0). Among offline methods, DPO achieves the

Method	AMD	Cataract	Grayscale	Tunnel Vision	Detached Retina	Avg.
ZS	76.0	73.0	36.0	77.0	82.0	68.8
KTO	79.0/75.0 (+3.0/-1.0)	75.0/72.0 (+2.0/-1.0)	**78.0**/73.0 (+42.0/+37.0)	78.0/72.0 (+1.0/-5.0)	84.0/84.0 (+2.0/+2.0)	78.8/75.2 (+10.0/+6.4)
GRPO	77.0/75.0 (+1.0/-1.0)	73.0/73.0 (+0.0/+0.0)	49.0/49.0 (+13.0/+13.0)	79.0/76.0 (+2.0/-1.0)	85.0/85.0 (+3.0/+3.0)	72.6/71.6 (+3.8/+2.8)
DPO	81.0/76.0 (+5.0/+0.0)	**77.0**/75.0 (+4.0/+2.0)	75.0/40.0 (+39.0/+4.0)	79.0/72.0 (+2.0/-5.0)	**86.0**/85.0 (+4.0/+3.0)	**79.6**/69.6 (+10.8/+0.8)
SFT	**91.0**/76.0 (+15.0/+0.0)	73.0/62.0 (+0.0/-11.0)	37.0/37.0 (+1.0/+1.0)	**88.0**/78.0 (+11.0/+1.0)	85.0/85.0 (+3.0/+3.0)	74.8/67.6 (+6.0/-1.2)

Table 1. Adaptation performance under different visual distortions for **CLEVR** (Johnson et al., 2017). We use Qwen2.5-VL 7B as speaker and listener. For each method we report two test accuracies: Maximum Test Accuracy / Test Accuracy of validation-set–picked checkpoint, with gains relative to **ZS**. Positive gains are in green, negatives in red. Best numbers per row are in **bold**. The last column reports average accuracies across distortions for both metrics, with gains relative to the ZS average.

Method	AMD	Cataract	Grayscale	Tunnel Vision	Detached Retina	Avg.
ZS	90.0	63.0	76.0	90.0	88.0	81.4
KTO	**94.0**/92.0 (+4.0/+2.0)	81.0/76.0 (+18.0/+13.0)	79.0/79.0 (+3.0/+3.0)	92.0/92.0 (+2.0/+2.0)	93.0/93.0 (+5.0/+5.0)	87.8/86.4 (+6.4/+5.0)
GRPO	93.0/91.0 (+3.0/+1.0)	82.0/78.0 (+19.0/+15.0)	81.0/77.0 (+5.0/+1.0)	93.0/91.0 (+3.0/+1.0)	93.0/93.0 (+5.0/+5.0)	**88.4**/86.0 (+7.0/+4.6)
DPO	89.0/80.0 (-1.0/-10.0)	**89.0**/73.0 (+26.0/+10.0)	**85.0**/77.0 (+9.0/+1.0)	**96.0/96.0** (+6.0/+6.0)	92.0/90.0 (+4.0/+2.0)	90.2/83.2 (+8.8/+1.8)
SFT	85.0/78.0 (-5.0/-12.0)	77.0/74.0 (+14.0/+11.0)	81.0/80.0 (+5.0/+4.0)	**96.0/96.0** (+6.0/+6.0)	**96.0/96.0** (+8.0/+8.0)	87.0/84.8 (+5.6/+3.4)

Table 2. Adaptation performance under different visual distortions for **CUB** (Wah et al., 2011). We use Qwen2.5-VL 7B as speaker and listener. For each method we report two test accuracies Maximum Test Accuracy / Test Accuracy of validation-set–picked checkpoint and gains relative to **ZS**. Positive gains are in green, negatives in red. Best numbers per row are in **bold**. The last column reports average accuracies across distortions for both metrics, with gains relative to the ZS average.

largest maximum gains (+8.8% average), driven by cataract (+26.0) and grayscale (+9.0), while SFT excels under tunnel vision and detached retina (+6.0 to +8.0). However, both offline methods drop significantly when evaluated with validation-set-picked checkpoint checkpoints (–7.0 to –8.0), again highlighting instability under practical selection.

Finally, on ImageNet (Tab. 3), the high ZS baseline of 83.0% leads to more modest but consistent gains across methods. DPO achieves the highest average maximum accuracy (88.0%) and gain (+5.0%), with strong performance on Grayscale (+5.0%) and Tunnel Vision (+7.0%). GRPO stands out for its reliability, showing the most stable performance with almost no drop between maximum and validation-picked scores (87.6% vs. 87.4%) and delivering a solid average gain of +4.4%. SFT shows strong peak performance on specific distortions like Cataract (+9.0%) and achieves the best scores on AMD and Detached Retina, but its zero gain on Grayscale and Tunnel Vision limits its average improvement. KTO provides a modest but stable average gain (+1.0% for validation-picked) but is uniquely penalized by the Grayscale distortion, where it underperforms the ZS baseline by -4.0%. Overall, GRPO and DPO show the most promising balance of performance.

5. Robustness Analysis

5.1. Cross Dataset Transfer

We next evaluate the cross-dataset generalization of adaptation methods by training on one dataset and testing on the other. Specifically, we first train on CLEVR and evaluate on CUB (Table 4). We then mirror this procedure by training on CUB and evaluating on CLEVR (Table 5). For evaluation, we report *test-pick* accuracy as described in Section 4. In each case, we report results for KTO, GRPO, DPO, and SFT under three representative distortions (cataract, grayscale, tunnel vision).

In the CLEVR→CUB setting, adaptation improves over ZS for all methods in terms of average accuracy. KTO yields the strongest overall transfer, achieving the best average of 84.7 (+8.4 over ZS), driven primarily by a large gain under cataract (+18.0) and consistent improvements on grayscale (+3.0) and tunnel vision (+4.0). GRPO and DPO show similar average behavior (both 82.0, +5.7), with GRPO exhibiting moderate gains across all distortions (+12.0 cataract, +4.0 grayscale, +1.0 tunnel vision), while DPO trades slightly smaller cataract gains (+10.0) for a stronger grayscale boost (+5.0). SFT also transfers strongly with an average of 82.7 (+6.4), matching KTO on cataract (+18.0) and achieving the best tunnel-vision performance (95.0, +5.0), but it degrades under grayscale (72.0, -4.0), which reduces its overall average. Overall, we observe that transfer from the compositional CLEVR domain to the natural-image CUB domain is generally successful: all methods improve on average, with KTO providing the most consistent gains.

In contrast, the CUB→CLEVR transfer is more challenging and exhibits substantial variability across distortions. KTO attains the strongest cataract performance among adapted methods (65.0), but it still underperforms relative to ZS on cataract (-8.0) and tunnel vision (-2.0), resulting in an overall drop in average accuracy (60.3, -1.7). GRPO follows a similar trend with a reduced average (60.3, -1.7), despite achieving the best tunnel-vision score (80.0, +3.0). SFT is the least stable in this direction, decreasing on cataract (-

Method	AMD	Cataract	Grayscale	Tunnel Vision	Detached Retina	Avg.
ZS	89.0	75.0	79.0	85.0	87.0	83.0
KTO	92.0 / 92.0 (+3.0 / +3.0)	82.0 / 74.0 (+7.0 / -1.0)	75.0 / 75.0 (-4.0 / -4.0)	89.0 / 89.0 (+4.0 / +4.0)	90.0 / 90.0 (+3.0 / +3.0)	85.6 / 84.0 (+2.6 / +1.0)
GRPO	**95.0 / 95.0 (+6.0 / +6.0)**	79.0 / 78.0 (+4.0 / +3.0)	83.0 / 83.0 (+4.0 / +4.0)	91.0 / 91.0 (+6.0 / +6.0)	90.0 / 90.0 (+3.0 / +3.0)	87.6 / 87.4 (+4.6 / +4.4)
DPO	94.0 / 94.0 (+5.0 / +5.0)	82.0 / 76.0 (+7.0 / +1.0)	**84.0** / 81.0 (+5.0 / +2.0)	**92.0** / 90.0 (+7.0 / +5.0)	88.0 / 88.0 (+1.0 / +1.0)	**88.0** / 85.8 (+5.0 / +2.8)
SFT	**95.0 / 95.0 (+6.0 / +6.0)**	**84.0** / 75.0 (+9.0 / +0.0)	79.0 / 79.0 (+0.0 / +0.0)	85.0 / 85.0 (+0.0 / +0.0)	**92.0** / 92.0 (+5.0 / +5.0)	87.0 / 85.2 (+4.0 / +2.2)

Table 3. Adaptation performance under different visual distortions for **ImageNet** (Deng et al., 2009). We use Qwen2.5-VL 7B as speaker and listener. For each method we report two test accuracies: Maximum Test Accuracy / Test Accuracy of validation-set–picked checkpoint, and gains relative to **ZS**. Positive gains are in green, negatives in red. Best numbers per column are in **bold**. The last column reports average accuracies across distortions for both metrics, with gains relative to the ZS average.

Method	Cataract	Grayscale	Tunnel Vision	Avg.
ZS	63.0	76.0	90.0	76.3
KTO	**81.0** (+18.0)	79.0 (+3.0)	94.0 (+4.0)	**84.7** (+8.4)
GRPO	75.0 (+12.0)	80.0 (+4.0)	91.0 (+1.0)	82.0 (+5.7)
DPO	73.0 (+10.0)	**81.0** (+5.0)	92.0 (+2.0)	82.0 (+5.7)
SFT	81.0 (+18.0)	72.0 (-4.0)	**95.0** (+5.0)	82.7 (+6.4)

Table 4. Adaptation performance for **CLEVR→CUB**. We use Qwen2.5-VL 7B as speaker and listener. We report ZS accuracy and Test-Pick accuracy (gains in green/red). Best results per distortion are in **bold**. The last column shows the row-wise average across distortions.

Method	Cataract	Grayscale	Tunnel Vision	Avg.
ZS	73.0	36.0	77.0	62.0
KTO	**65.0** (-8.0)	41.0 (+5.0)	75.0 (-2.0)	60.3 (-1.7)
GRPO	62.0 (-11.0)	39.0 (+3.0)	**80.0** (+3.0)	60.3 (-1.7)
DPO	58.0 (-15.0)	**54.0** (+18.0)	74.0 (-3.0)	**62.0** (+0.0)
SFT	61.0 (-12.0)	43.0 (+7.0)	72.0 (-5.0)	58.7 (-3.3)

Table 5. Adaptation performance for **CUB→CLEVR**. We use Qwen2.5-VL 7B as speaker and listener. We report ZS accuracy and Test-Pick accuracy (gains in green/red). Best results per distortion are in **bold**. The last column shows the row-wise average across distortions.

12.0) and tunnel vision (-5.0) and producing the lowest average (58.7, -3.3), even though it improves on grayscale (+7.0). DPO presents the clearest trade-off: it achieves a large improvement on grayscale (54.0, +18.0) but drops sharply on cataract (-15.0) and slightly on tunnel vision (-3.0), leading to an average that matches ZS (62.0, +0.0). Taken together, these results suggest that while CLEVR→CUB transfer is broadly effective, the reverse direction remains difficult. Additionally, we confirm that the strategies learnt by the models are generalizable.

5.2. Adapting to different agent under divergent visual space

Table 6 reports adaptation results on the CUB dataset using Qwen2.5-VL-7B as the speaker and InternVL3.5 (Wang et al., 2025) as the listener. The baseline listener is already strong, with near-ceiling accuracy on Tunnel Vision (96.0) and Detached Retina (97.0), but weaker under Cataract (82.0), Grayscale (72.0), and AMD (95.0). In this high-performing regime, adaptation provides only modest gains. KTO improves slightly on Grayscale (+1.0 Test-Pick), and provides small improvements on Tunnel Vision (+1.0) and Detached Retina (+3.0), while matching the baseline on Cataract and AMD, averaging 89.4/88.2 (+1.0/–0.2). GRPO achieves the strongest boost on Grayscale (+5.0 Test-Pick) but shows limited improvements elsewhere, matching the baseline on Cataract, Tunnel Vision, and AMD, and improving modestly on Detached Retina (+1.0), yielding averages of 89.8/86.4 (+1.4/–2.0). Overall, gains are limited, reflecting that adaptation has less room to help when the baseline listener already performs well. Furthermore, the weaker results on Cataract highlight the challenges of adapting across heterogeneous agents.

Table 7 presents results on CLEVR with the same speaker–listener pairing. Here the baseline is much weaker overall (52.4 average), and especially poor on Grayscale (28.0). In this lower-performing regime, adaptation produces substantial improvements. KTO provides the most consistent and significant gains, especially on Grayscale (+42.0/+36.0), with additional improvements on Cataract (+4.0/+3.0) and Detached Retina (+4.0/+4.0), raising the averages to 62.6/60.2 (+10.2/+7.8). GRPO offers a smaller benefit, with improvements under Cataract (+2.0/+2.0), Tunnel Vision (+2.0/–2.0), Detached Retina (+4.0/+4.0), and AMD (+1.0/+1.0), but limited gains on Grayscale (+2.0/–3.0), resulting in averages of 54.6/52.8 (+2.2/+0.4).

Taken together, these findings reveal a sharp asymmetry. When the baseline listener is already strong, as in CUB, adaptation provides only minor and uneven gains. By contrast, when the listener is weaker, as in CLEVR, adaptation, especially with KTO, yields large and reliable improvements. Compared to earlier results with Qwen2.5-VL-7B as both speaker and listener, these results suggest that adapting across heterogeneous agents is more challenging, with smaller benefits for strong listeners but clear advantages when the listener struggles under perceptual divergence

6. Qualitative Results

We present qualitative examples in Fig. 3 to highlight key observations. In each case, the left column shows the original images x, while the rightmost column shows its distorted

Method	Cataract	Grayscale	Tunnel Vision	Detached Retina	AMD	Avg.
ZS	82.0	72.0	96.0	97.0	95.0	88.4
KTO	**82.0**/80.0 (+0.0/-2.0)	**73.0**/72.0 (+1.0/+0.0)	**97.0/97.0** (+1.0/+1.2)	**100.0**/98.0 (+3.0/+1.0)	**95.0/94.0** (+0.0/-1.0)	89.4/88.2 (+1.0/-0.2)
GRPO	82.0/72.9 (+0.0/-9.1)	**77.0**/71.0 (+5.0/-1.0)	**97.0/97.0** (+1.0/+1.0)	98.0/97.0 (+1.0/+0.0)	**95.0**/94.0 (+0.0/-1.0)	**89.8**/86.4 (+1.4/-2.0)

Table 6. Adaptation performance under different visual distortions with InternVL3.5 (Wang et al., 2025) as listener for **CUB** (Wah et al., 2011). We use Qwen2.5-VL 7B as speaker. We report *Test-Pick/Val-Pick* accuracies with gains relative to **ZS**. Positive gains are in green, negatives in red. Best adapted values per column are in **bold**. The last column reports row-wise averages (and gains) across distortions.

Method	Cataract	Grayscale	Tunnel Vision	Detached Retina	AMD	Avg.
ZS	53.0	28.0	60.0	60.0	61.0	52.4
KTO	**57.0**/56.0 (+4.0/+3.0)	**70.0**/64.0 (+42.0/+36.0)	61.0/56.0 (+1.0/-4.0)	**64.0**/64.0 (+4.0/+4.0)	61.0/61.0 (+0.0/+0.0)	**62.6**/60.2 (+10.2/+7.8)
GRPO	55.0/55.0 (+2.0/+2.0)	30.0/25.0 (+2.0/-3.0)	**62.0**/58.0 (+2.0/-2.0)	64.0/64.0 (+4.0/+4.0)	**62.0**/62.0 (+1.0/+1.0)	54.6/52.8 (+2.2/+0.4)

Table 7. Adaptation performance under different visual distortions with InternVL3.5 (Wang et al., 2025) as listener for CLEVR (Johnson et al., 2017) dataset. We use Qwen2.5-VL 7B as speaker. We report Test-Pick/Val-Pick accuracies (with gains in green for improvements and red for drops). Best adapted values per column are in **bold**. The last column reports row-wise averages with gains relative to the ZS average.

counterparts $\hat{x}$. For both, we report the base model's description and the adapted model's description. Method names are shown in blue, text in **red** indicates changes relative to the base description, and **green text** marks newly introduced information. Adapted models generally produce descriptions that are more consistent with the distorted view, leading to improved speaker–listener communication. For AMD distortion, adapted model enriches the description by emphasizing fine details such as the bird's beak and the rocky surface. Under Detached Retina, adapted model incorporates contextual background cues, describing the blurred foliage, which provides the listener with disambiguating information. In the Grayscale case, the adapted model avoids reliance on color-based cues and instead introduces structure-sensitive descriptors such as "a thin, bare branch," which align more closely with the listener's visual perception. For Tunnel Vision, the adapted model improves the base description "just caught a fish" to the more faithful "dives towards the water," aligning with what the listener actually sees, where the fish is not clearly visible. Finally, in the Cataract setting, adapted model strengthens the description of the target bird while adding background details about vegetation and posture, helping the listener distinguish the target from confounding images. These qualitative findings are consistent with our quantitative results, showing that adaptation improves interpretability and fidelity as well.

6.1. Analyzing Generations

Figure 4 analyzes how different training objectives modify the lexical content of generated descriptions under grayscale distortion on the CUB dataset. We measure *lexical richness* using a lightweight attribute lexicon: generated captions are parsed and the number of terms belonging to predefined semantic categories (e.g., color, shape, background, parts) is counted per sentence. Higher values indicate that the model explicitly mentions more visual attributes in its description. We report mean scores per sentence. In absolute terms, all fine-tuning methods increase lexical richness compared to the base model. SFT exhibits the largest gains, particularly in background, parts, and shape, suggesting substantial lexical expansion. GRPO and KTO provide more moderate yet consistent improvements across structural categories such as position and posture. DPO shows better alignment by improving structural attributes. Taken together, SFT wins by increasing the amount of information across all attributes while RL methods tend to be more conservative with non-structural attributes like colors with DPO placing most focus on structural attributes.

We extend the grayscale lexicon to include new categories of clarity and central peripheral words. Clarity include terms like 'blur', 'fog' and 'clear' while central peripheral include terms like 'central', 'periphery' and 'focal' highlighting the focus in image. Figures 2 summarize these cumulative counts across CLEVR and CUB datasets. SFT consistently produces the largest counts across nearly all attribute categories caused by verbosity. It also indicates a tendency toward over-specification.

In contrast, RL methods exhibit more controlled lexical behavior. GRPO and KTO produce moderate yet consistent increases across structural attributes such as position, posture, and parts. DPO shows a more selective pattern, concentrating improvements on structurally informative attributes. This behavior likely stems from DPO's use of negative supervision, which explicitly penalizes undesirable attributes. KTO and GRPO shows different patterns across different datasets. This demonstrates that objective functions behave different based on the task. Overall, these results suggest that training objectives strongly influence how models allocate semantic emphasis. We provide details of the lexicons and further analysis in the supplementary.

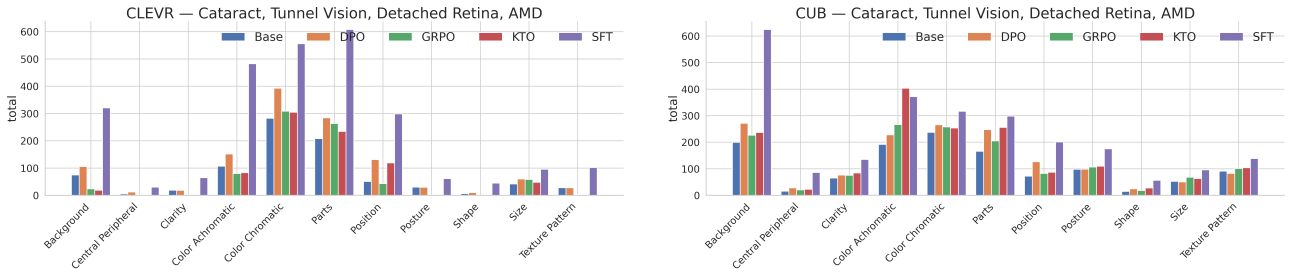


Figure 2. Cumulative attribute counts across all distortions for CLEVR (left) and CUB (right). SFT shows the largest increases across nearly all categories caused by verbosity. RL objectives exhibit more controlled behavior.

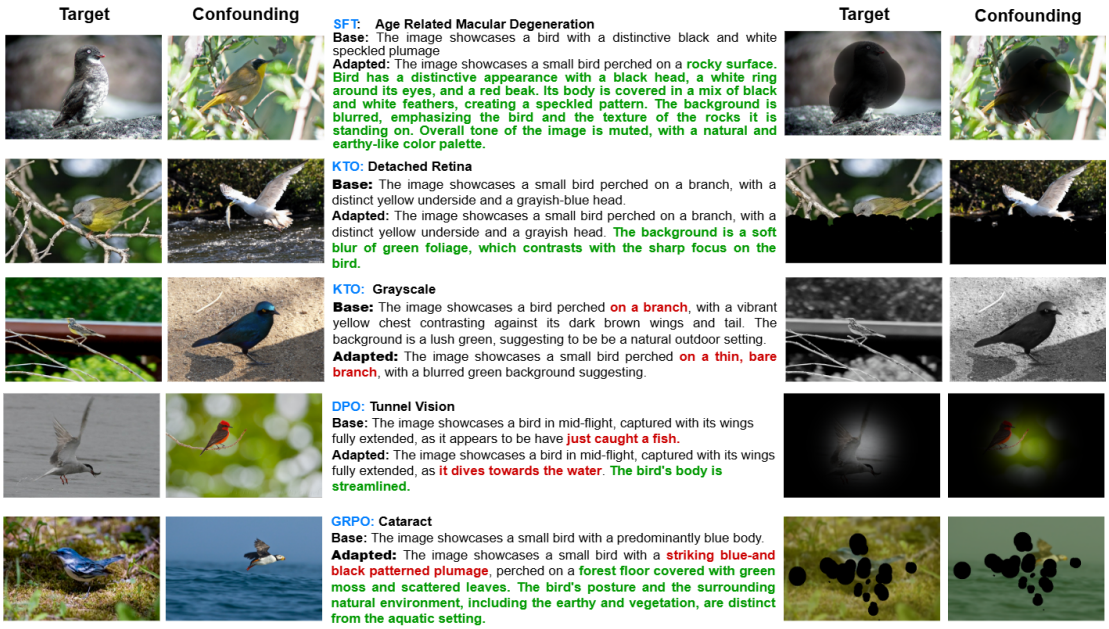


Figure 3. **Qualitative results** for Qwen2.5-VL 7B as speaker. The left image is being described in each description. On left we share original images and on right we share distorted images. In each description text with color blue represents the method name. **Red text** represent change of information from original description while **green text** represents new information.

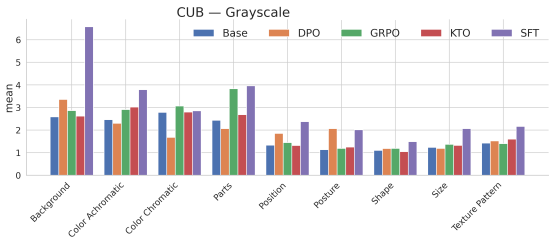


Figure 4. Category-wise lexical analysis under grayscale distortion on CUB. We show the mean number of terms per sentence.

7. Conclusion

We studied aligning multimodal agents that inhabit *divergent visual spaces*. Using an image reference game with five distortion types, we compared offline (SFT, DPO) and online (KTO, GRPO) post-training strategies. Our experiments show that (i) Offline adaptation is the best if the impairment is known with DPO consistently outperforming; (ii) online adaptation with KTO and GRPO delivers strong improvements across distortions by learning directly from interaction feedback and KTO can show strong improvements in certain tasks through high peak performance; and (iii) qualitative analyses confirm that adapted speakers prioritize distortion-consistent cues with DPO demonstrating highest alignment with listener's view. Furthermore, we show that choice of objective function strongly influences how models allocate semantic emphasis. Cross-dataset studies further indicate that compositional nature of CLEVR dataset makes it easier to transfer to natural images (CUB). Adapting to a different listener family yields smaller gains, underscoring the challenge of alignment across heterogeneous model representations.

References

Alaniz, S., Marcos, D., and Akata, Z. Learning decision trees recurrently through communication. In *CVPR*, 2021.

Ansell, A., Ponti, E., Korhonen, A., and Vulic, I. Composable sparse fine-tuning for cross-lingual transfer. In *ACL*, 2021.

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*, 2022.

Corona, R., Alaniz, S., and Akata, Z. Modeling conceptual understanding in image reference games. In *NeurIPS*, 2019.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. Visual dialog. In *CVPR*, 2016.

de Vries, H., Strub, F., Chandar, A. P. S., Pietquin, O., Larochelle, H., and Courville, A. C. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. In *arXiv preprint arXiv:2402.01306*, 2024.

Guo, S., Zhang, B., Liu, T., Liu, T., Khalman, M., Llinares-López, F., Ramé, A., Mesnard, T., Zhao, Y., Piot, B., Ferret, J., and Blondel, M. Direct language model alignment from online ai feedback. In *arXiv preprint arXiv:2402.04792*, 2024.

Hsu, C., Buffelli, D., McGowan, J., Liao, F., Chen, Y., Vakili, S., and Shiu, D. Group think: Multiple concurrent reasoning agents collaborating at token level granularity. *CoRR*, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022a.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022b.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Ju, H. and Aral, S. Collaborating with AI agents: Field experiments on teamwork, productivity, and performance. *CoRR*, 2025.

Lialin, V., Shivagunde, N., Muckatira, S., and Rumshisky, A. Relora: High-rank training through low-rank updates. In *NeurIPS workshops*, 2023.

Liu, Liu, J., Koike-Akino, T., Wang, P., Brand, M., Wang, Y., and Parsons, K. Loda: Low-dimensional adaptation of large language models. In *NeurIPS workshops*, 2023.

Liu, A., Bai, H., Lu, Z., Kong, X., Wang, S., Shan, J., Cao, M., and Wen, L. Direct large language model alignment through self-rewarding contrastive prompt distillation. In *ACL*, 2024.

Ma, Z., Dou, Z., Zhu, Y., Zhong, H., and rong Wen, J. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *SIGIR*, 2021.

Ma, Z., Sansom, J., Peng, R., and Chai, J. Towards a holistic landscape of situated theory of mind in large language models. In *EMNLP*, 2023.

Mercea, O.-B., Gritsenko, A., Schmid, C., and Arnab, A. Time-memory-and parameter-efficient visual adaptation. In *CVPR*, 2024.

Ni, J., Young, T., Pandelea, V., Xue, F., Adiga, V. V., and Cambria, E. Recent advances in deep learning based dialogue systems: a systematic survey. In *Artificial Intelligence Review*, 2021.

Pan, Z., Zhang, Y., Zhang, J., Lu, H., Luo, H., Han, Y., Yu, P. S., Li, M., and Liu, H. Fairreason: Balancing reasoning and social bias in mllms. *arXiv preprint arXiv:2507.23067*, 2025.

Pfeiffer, J., Vulic, I., Gurevych, I., and Ruder, S. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *EMNLP*, 2020.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.

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

Raileanu, R., Denton, E. L., Szlam, A., and Fergus, R. Modeling others using oneself in multi-agent reinforcement learning. In *ICML*, 2018.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *ICLR*, 2023.

Serban, I., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 2015.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sheng, Y., Cao, S., Li, D., Hooper, C., Lee, N., Yang, S., Chou, C., Zhu, B., Zheng, L., Keutzer, K., Gonzalez, J. E., and Stoica, I. S-lora: Serving thousands of concurrent lora adapters. In *arXiv preprint arXiv:2311.03285*, 2023.

Snell, C. B., Kostrikov, I., Su, Y., Yang, M., and Levine, S. Offline rl for natural language generation with implicit language q learning. In *ICLR*, 2023.

Song, H., Zhang, W., Hu, J., and Liu, T. Generating persona consistent dialogues by exploiting natural language inference. In *AAAI*, 2019.

Sung, Y.-L., Cho, J., and Bansal, M. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *NeurIPS*, 2022.

Takmaz, E., Brandizzi, N., Giulianelli, M., Pezzelle, S., and Fern'andez, R. Speaking the language of your listener: Audience-aware adaptation via plug-and-play theory of mind. In *ACL*, 2023.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Wang, J., Leong, C. T., Wang, J., Lin, D., Li, W., and Wei, X.-Y. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *ACL*, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

World Health Organization and The Lancet Global Health Commission on Global Eye Health. World Report on Vision. *World Health Organization*, 2021. URL https://www.who.int/publications/i/item/9789241517402.

Wu, Y., Xiang, Y., Huo, S., Gong, Y., and Liang, P. Lora-sp: Streamlined partial parameter adaptation for resource-efficient fine-tuning of large language models. In *arXiv preprint arXiv:2403.08822*, 2024.

yang Liu, S., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. Dora: Weight-decomposed low-rank adaptation. In *ICML*, 2024.

Zhang, L., Zhang, L., Shi, S., Chu, X., and Li, B. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. In *arXiv preprint arXiv:2308.03303*, 2023.

Zhang, Y., Gao, X., Lee, S., Brockett, C., Galley, M., Gao, J., and Dolan, W. B. Consistent dialogue generation with self-supervised feature learning. In *arXiv preprint arXiv:1903.05759*, 2019.

Zhang, Y., Sun, R., Chen, Y., Pfister, T., Zhang, R., and Arik, S. Ö. Chain of agents: Large language models collaborating on long-context tasks. In *NeurIPS*, 2024.

Zhong, H., Dou, Z., Zhu, Y., Qian, H., and rong Wen, J. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *NAACL-HLT*, 2022.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. In *arXiv preprint arXiv:1909.08593*, 2019.

Appendix Contents

Supplementary Material

A. Lexicon-Based Analysis for Grayscale Evaluations

To better understand how models adapt their language when visual color information is removed, we perform a lexicon-based analysis of the generated textual descriptions under grayscale image distortions. The goal of this analysis is to quantify which types of visual attributes the models rely on when color cues are unavailable. This analysis is provided in Section 6.1 of main manuscript.

A.1. Lexicon Categories

We construct a vocabulary of attribute words grouped into semantic categories corresponding to common visual properties. These categories capture different types of visual reasoning signals that may appear in generated descriptions.

Chromatic Color Terms These terms describe colors that cannot be directly inferred from grayscale images. Their presence may indicate hallucination or reliance on prior knowledge rather than visual evidence.

red, blue, green, yellow, orange, purple, pink, brown, cyan, magenta, gold, golden, silver, crimson, scarlet, azure, emerald, amber, violet, maroon, teal, turquoise, lavender, coral, indigo, beige, tan, olive, burgundy, multicolor, multicoloured, hue, hues, chromatic, saturation, bright, vivid, vibrant, colored, colour, colorful, colourful

Achromatic Color Terms These terms describe intensity-based color properties that remain visible in grayscale images.

black, white, gray, grey, greyish, grayish, dark, light, monochrome, grayscale, greyscale, achromatic, shade, shades, tone, tones, contrast, contrasting, pale, dull, brightness, shadow, shadows, highlight, highlights, silhouette

Spatial Position Words describing relative spatial locations within the image.

left, right, center, centre, top, bottom, middle, above, below, behind, front, beside, between, near, upper, lower, foreground, background, side, sides, corner, edge, edges, horizon, vertical, horizontal

Posture and Pose Terms describing the orientation or activity of objects.

standing, sitting, perching, perched, flying, lying, walking, running, resting, posed, pose, facing, upright, bent, stretched, crouching, leaning, spread, spreading, extended, folded, raised

Background and Environment Words describing environmental context surrounding the main object.

sky, grass, branch, branches, ground, water, tree, trees, leaves, foliage, rock, rocks, sand, soil, wall, floor, surface, outdoor, indoor, natural, habitat, environment, setting, scene, landscape

Shape Terms describing geometric form or outline.

round, oval, square, triangular, circular, rectangular, curved, straight, angular, pointed, rounded, elongated, symmetrical, asymmetric, shape, shaped, form, outline

Texture and Pattern Words describing surface appearance or repeating structures.

striped, stripes, spotted, spots, smooth, rough, pattern, patterns, texture, textured, mottled, speckled, flecked, dotted, lined, scaled, feathered, plumage

Size Terms describing object scale or relative dimensions.

large, small, medium, tiny, big, little, sized, length, width, tall, short, long, wide, narrow, thick, thin, compact, prominent, dominant

Object Parts and Anatomy Words describing object components, including bird anatomy (for CUB) and object primitives (for CLEVR).

> head, beak, bill, wing, wings, tail, body, breast, throat, crest, crown, back, belly, leg, legs, foot, eye, eyes, neck, chest, object, objects, cube, sphere, cylinder, metal, rubber, material

A.2. Quantification Procedure

For each generated response, we count the occurrences of words belonging to each lexicon category. This produces a per-sample vector of category frequencies indicating which types of visual attributes are mentioned in the description. We then report mean of these occurrences by dividing it by number of successful generations.

This analysis is performed separately for the base (zero-shot) model and the fine-tuned model, enabling direct comparison of how training influences the types of visual attributes used in descriptions.

A.3. Extended Lexicon for Spatial Distortion Analysis

To analyze model behavior under spatially localized distortions (AMD, Cataract, Tunnel Vision, Detached Retina), we extend the lexicon introduced in Sec. A with additional categories that capture spatial visibility and image clarity. These terms are designed to reflect language that models may use when reasoning about partially visible or degraded visual content.

Central vs. Peripheral Visibility This category captures references to spatial regions within the visual field. These terms are particularly relevant for distortions that affect central or peripheral vision (e.g., macular degeneration or tunnel vision).

> central, peripheral, periphery, peripherally, focal, foveal, visible, visibility, obscured, occluded, blocked, hidden, view, field, region, area, zone, spot, patch

Clarity and Blur These terms describe visual clarity and are intended to capture language related to blur or degraded image quality, which is characteristic of distortions such as cataracts or defocus.

> blur, blurred, blurry, hazy, haze, foggy, fog, clear, clearly, sharp, sharply, distorted, distortion, fuzzy, unclear, indistinct, cloudy, opaque

The remaining lexicon categories (color, spatial position, posture, background, shape, texture/pattern, size, and object parts) remain unchanged from the grayscale analysis. Together, the full vocabulary enables quantification of how models describe spatial structure, visibility, and visual clarity when reasoning about distorted inputs.

The same counting procedure described in Sec. A is applied to these additional categories. However, for aggregation, we simply report the sum of occurences in Figure 2 of main manuscript.

A.4. Validating Lexicon Analysis through Attribute Verification under Visual Distortions using MLLM

To analyze how visual distortions affect the perceptual grounding of generated descriptions, we measure how many attributes described by a speaker are verifiable in the distorted image.

For each generated description, we first extract a set of visual attributes using a separate Qwen2.5-VL 32B attribute extraction step. The extractor converts the generated description into a list of atomic, verifiable attributes (e.g., "has a red crest", "is perched on a branch"). Each attribute is then independently verified against the corresponding distorted image using the same MLLM acting as a visual verifier. The verifier receives the distorted image and a single attribute statement and produces a binary decision indicating whether the attribute is present in the image.

Given a description with k extracted attributes, let $v_i \in \{0, 1\}$ denote whether attribute i is verified by the visual verifier. The attribute verification score for a sample is computed as

$$\text{Score} = \sum_{i=1}^{k} v_i.$$

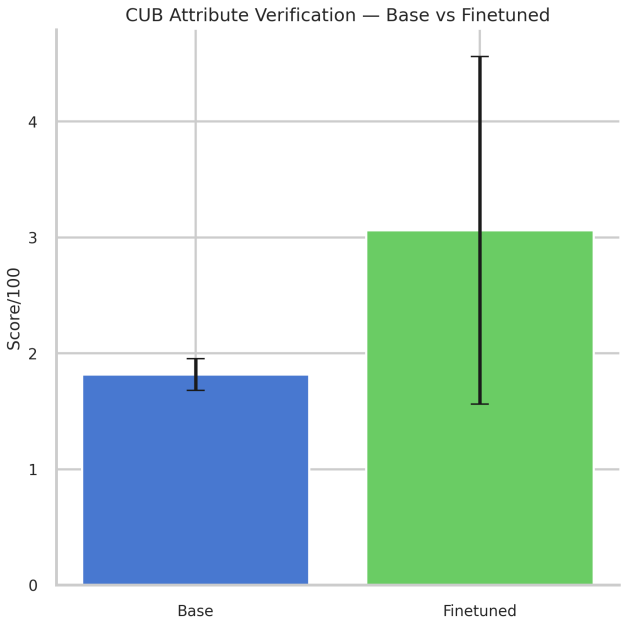


Figure 5. **Attribute verification under perceptual distortions.** For each generated description, we extract a set of visual attributes and verify whether each attribute is supported by the distorted image using an MLLM-based verifier. **Base** corresponds to the zero-shot model, while **Finetuned** reports the mean performance across all adapted speakers. Higher values indicate that generated descriptions remain better grounded in visually observable attributes despite the applied distortions.

This score measures the number of generated attributes that are visually supported by the distorted input image. Furthermore, we divide this score by 100 (Score/100) to get the final score we report in the figure below.

For CUB and distortion type, we evaluate multiple adaptation algorithms. For every configuration we compute the mean attribute verification score across evaluation samples. To simplify comparison, we report two aggregate groups in the main analysis: **Base**, corresponding to the zero-shot model (`checkpoint-0`), and **Finetuned**, which represents the average score across all adapted models.

The resulting metric reflects how well the generated descriptions remain grounded in visually observable attributes when the input image is degraded by perceptual distortions. The corresponding results are visualized in Figure 5. This is also aligned with our lexicon based analysis which consistently finds increasing attributes that are aligned with the distortions.

A.5. Word Count Analysis

Figure 6 shows the average words per sentence on the right. It shows show that KTO produces the longest sentences, indicating higher sentence-level verbosity. The base model, DPO, and GRPO remain relatively close, reflecting moderate syntactic expansion. In contrast, SFT generates shorter sentences on average, suggesting more concise sentence structuring despite other verbosity changes. On the other hand, the total word count in left figure in Figure 6 reveals that SFT produces substantially longer responses overall, demonstrating the largest increase in global verbosity. DPO, GRPO, and KTO yield moderate increases compared to the base model. This suggests that RL-based methods lead to controlled verbosity growth, whereas SFT primarily increases verbosity through longer overall outputs. Furthermore, Figure 4 shows that SFT increases all the attributes and majorly wins by mentioning everything instead of aligning strictly with the listener. SFT only focuses on positive preferences which misses out on important signal DPO gets by contrasting positive preferences with negative.

B. Ablation Heterogenous Model Sizes

In this section, we conduct an ablation study to analyze the impact of using heterogeneous model sizes for the speaker and listener roles on CUB dataset. We explore four configurations using the Qwen-2.5 VL 3B and 7B models: a symmetric setup with matched sizes (3B/3B and 7B/7B) and asymmetric setups with mismatched sizes (3B speaker/7B listener and 7B

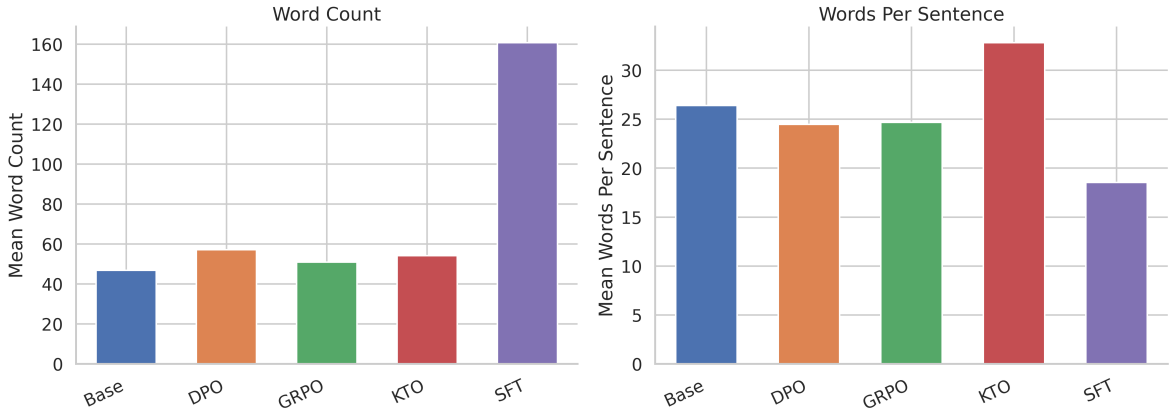


Figure 6. Word count analysis for different algorithms. We observe that SFT shows the highest word count per generation (top). However, it is lowest in word count per sentence (bottom). This shows that SFT is more verbose and highlights multiple aspects as compared to other algorithms.

speaker/3B listener). The results, presented in Tab. 8, evaluate the zero-shot (Base) performance and the maximum accuracy (Max) achieved after adaptation with GRPO and KTO.

Our first key observation relates to the zero-shot baseline performance. Counter-intuitively, the largest model configuration (7B/7B) does not yield the best baseline accuracy. In fact, it has the lowest average baseline among all setups (81.4%), primarily due to significant performance drops on the Cataract (63.0%) and Detached Retina (76.0%) distortions. The strongest baseline performance is achieved by the asymmetric configuration with a smaller speaker and a larger listener (3B/7B), which reaches an average accuracy of 88.8% for GRPO and 88.0% for KTO. This suggests that a more powerful listener model is crucial for robustly interpreting descriptions, even those generated by a smaller speaker.

Next, we analyze the gains from adaptation. We observe that both GRPO and KTO deliver consistent performance improvements across all model configurations. The magnitude of this improvement is most pronounced in the 7B/7B setup, which had the lowest baseline. For instance, GRPO and KTO improve the average accuracy by +6.2 and +6.4 points, respectively, with remarkable gains on the Cataract distortion (+19.0 for GRPO, +18.0 for KTO) where the baseline was weakest. This indicates that adaptation is highly effective at correcting the initial vulnerabilities of larger models.

From a practical standpoint, the optimal configuration for achieving the highest final accuracy is the 3B speaker and 7B listener setup. After adaptation with GRPO, this configuration reaches an average accuracy of 92.0%, the highest across all experiments. This result highlights a compelling strategy for efficient system design: pairing a smaller, computationally cheaper speaker model with a larger, more capable listener model provides a strong initial baseline that can be further enhanced through adaptation to achieve state-of-the-art performance. This asymmetric approach appears to offer a better trade-off between performance and computational resources than simply using the largest available models for both roles.

C. Performace of Top Tier Models

To contextualize the performance of our adaptation methods, we conducted a comprehensive evaluation of several leading proprietary, closed-source large multimodal models (LMMs). We benchmarked their zero-shot performance on the ImageNet and CLEVR datasets under three challenging visual distortions: Cataract, Grayscale, and Tunnel Vision. The results, presented in Table 9 and Table 10, offer critical insights into the capabilities and limitations of current state-of-the-art models and highlight the value of targeted adaptation. We use the same prompting strategy as shown in main manuscript.

On the ImageNet dataset (Table 9), the top-tier models demonstrate remarkable robustness. Models like GPT-4o, GPT-5.1, and Claude-Sonnet-4 maintain exceptionally high accuracy, often exceeding 95% even with significant visual distortions. GPT-4o, for instance, achieves a perfect 100% accuracy on grayscale images, suggesting its internal representations are largely invariant to the absence of color information for this real-world object recognition task.

In comparison, the open-source Qwen2.5-VL 7B model in a zero-shot (ZS) setting shows a noticeable performance gap, with accuracies of 75.0% on Cataract, 79.0% on Grayscale, and 85.0% on Tunnel Vision. While these are respectable scores,

Method	Model Configuration		Type	Accuracy [%]				
	Speaker	Listener		AMD	Cataract	Detached Retina	Grayscale	Average
GRPO	Qwen-2.5 VL 3B	Qwen-2.5 VL 3B	Base	89.0	67.0	98.0	94.0	87.0
			Max	91.0 (+2.0)	67.0 (+0.0)	99.0 (+1.0)	95.0 (+1.0)	88.0 (+1.0)
	Qwen-2.5 VL 3B	Qwen-2.5 VL 7B	Base	94.0	79.0	98.0	84.0	88.8
			Max	98.0 (+4.0)	86.0 (+7.0)	100.0 (+2.0)	84.0 (+0.0)	92.0 (+3.2)
	Qwen-2.5 VL 7B	Qwen-2.5 VL 3B	Base	92.0	64.0	97.0	93.0	86.5
			Max	97.0 (+5.0)	72.0 (+8.0)	98.0 (+1.0)	98.0 (+5.0)	91.3 (+4.8)
	Qwen-2.5 VL 7B	Qwen-2.5 VL 7B	Base	90.0	63.0	76.0	88.0	81.4
			Max	93.0 (+3.0)	82.0 (+19.0)	77.0 (+1.0)	93.0 (+5.0)	87.6 (+6.2)
KTO	Qwen-2.5 VL 3B	Qwen-2.5 VL 3B	Base	88.0	67.0	96.0	94.0	86.3
			Max	90.0 (+2.0)	70.0 (+3.0)	100.0 (+4.0)	98.0 (+4.0)	89.5 (+3.2)
	Qwen-2.5 VL 3B	Qwen-2.5 VL 7B	Base	94.0	80.0	98.0	80.0	88.0
			Max	96.0 (+2.0)	85.0 (+5.0)	99.0 (+1.0)	82.0 (+2.0)	90.5 (+2.5)
	Qwen-2.5 VL 7B	Qwen-2.5 VL 3B	Base	92.0	67.0	96.0	93.0	87.0
			Max	93.0 (+1.0)	75.0 (+8.0)	97.0 (+1.0)	97.0 (+4.0)	90.8 (+3.8)
	Qwen-2.5 VL 7B	Qwen-2.5 VL 7B	Base	90.0	63.0	76.0	88.0	81.4
			Max	94.0 (+4.0)	81.0 (+18.0)	79.0 (+3.0)	93.0 (+5.0)	87.8 (+6.4)

Table 8. Detailed comparison of Zero-shot (Base) and Maximum Evaluation Accuracy across different distortions, including an overall Average on CUB dataset. Each model configuration is presented with separate rows for Base and Max accuracy, with gains over base shown in green.

they are clearly surpassed by the larger proprietary models. After fine-tuning with Direct Preference Optimization (DPO), the Qwen model's performance improves to 82.0%, 84.0%, and 92.0% respectively. This adaptation narrows the gap but does not close it entirely. The results on ImageNet suggest that for general perceptual tasks, the sheer scale and extensive pre-training of top-tier models provide a significant and durable advantage in robustness.

The results on the CLEVR dataset (Table 10) paint a dramatically different picture and underscore a critical vulnerability in large, generalist models. CLEVR requires compositional reasoning, a task that is more abstract than simple object recognition. While most top-tier models perform well under the Cataract and Tunnel Vision distortions, their performance collapses on grayscale images. For example, GPT-4o's accuracy plummets to 22.0%, and Gemini-3-Pro's drops to a mere 14.0%. The best-performing model, GPT-5.1, only reaches 36.0% accuracy, which is identical to the zero-shot performance of the much smaller Qwen2.5-VL 7B model. This indicates that the reliance on color cues for object differentiation and relational reasoning is a systemic weakness for un-adapted models on this synthetic dataset.

This is where the power of targeted adaptation becomes strikingly evident. After fine-tuning with DPO, the Qwen2.5-VL 7B model's accuracy on grayscale CLEVR images skyrockets from 36.0% to 75.0%. With this adaptation, the open-source model not only recovers its reasoning ability but vastly outperforms every single top-tier proprietary model on this specific task. Furthermore, its adapted score of 77.0% on Cataract is also superior to that of GPT-4o (74.0%) and Claude-Sonnet-4 (58.0%).

In summary, while top-tier models excel in general robustness on real-world image datasets like ImageNet, they can be surprisingly brittle when faced with specific distortions on tasks requiring complex reasoning, such as CLEVR. Our findings demonstrate that a smaller, open-source model can be fine-tuned to surpass these leading models in such challenging, niche scenarios, highlighting that targeted adaptation is a powerful and efficient strategy for achieving specialized robustness. Lastly, these results once again shed the light on the importance of task specific approaches for adaptation. Every task and distortion pair provides unique challenges. Some may be more distant from the learnt knowledge of the model and others may be closer. Lastly, we observe that GPT 5.1 shows the best trade off between the two datasets by showing average performance of 96.3% on Imagenet and taking a second place in CELVR dataset with 70% average accuracy.

D. Ablation over prompt template

Our standard prompt template , designed to elicit distinguishing features of the target image, is as follows:

Speaker	Accuracy [%]			Avg
	Cataract	Grayscale	Tunnel Vision	
GPT-5.1	**95.0**	97.0	97.0	**96.3**
GPT-4o	94.0	**100.0**	**99.0**	97.7
Claude-Sonnet-4	**95.0**	98.0	**99.0**	97.3
Gemini-3-Pro	92.0	98.0	96.0	95.3
Grok-4	90.0	97.0	98.0	95.0
Qwen2.5-VL 7B (ZS)	73.0	36.0	77.0	62.0
Qwen2.5-VL 7B (DPO)	82.0	84	92	86.0

Table 9. Performance of top-tier speaker models on the **ImageNet** dataset, compared with an adapted open-source model (Qwen2.5-VL 7B). Listener is Qwen2.5-VL 7B.

Speaker	Accuracy [%]			Avg
	Cataract	Grayscale	Tunnel Vision	
GPT-5.1	76.0	36.0	**98.0**	70.0
GPT-4o	74.0	22.0	88.0	61.3
Claude-Sonnet-4	58.0	26.0	94.0	59.3
Gemini-3-Pro	61.0	14.0	89.0	54.7
Grok-4	59.0	13.0	89.0	53.7
Qwen2.5-VL 7B (ZS)	73.0	36.0	77.0	62.0
Qwen2.5-VL 7B (DPO)	**77.0**	**75.0**	79.0	**77.0**

Table 10. Performance of top-tier speaker models on the **CLEVR** dataset, compared with an adapted open-source model. Note the dramatic improvement of the adapted model on grayscale images. Listener is Qwen2.5-VL 7B.

Table 11. Comparison of prompted vs. normal (Zero-Shot) accuracy for Grayscale and Cataract distortions on the CUB and CLEVR datasets. The "Prompted Acc." reflects performance with the distortion-aware instructions. We report results with Qwen2.5-VL 7B as listener and speaker both.

Accuracy Type	Grayscale	Cataract
Prompted Acc. [%]	74.0	**80.0**
Normal Acc. [%]	**76.0**	63.0

(a) Performance on the **CUB** dataset.

Accuracy Type	Grayscale	Cataract
Prompted Acc. [%]	**50.0**	56.0
Normal Acc. [%]	36.0	**73.0**

(b) Performance on the **CLEVR** dataset.

```
You see two images. Image 1: [IMG_1], Image 2: [IMG_2]. Focus on Image
1. Your goal is to describe Image 1 in a way that clearly distinguishes
it from Image 2. Do not mention Image 2 in your description at all and
just talk about Image 1. Highlight unique features of Image 1 as a whole
that differentiate it from Image 2.
```

To make the model "distortion-aware," we introduce a listener-centered perspective by inserting an additional line of text just before the final sentence. This extra instruction explicitly describes the nature of the perceptual impairment. For the distortions analyzed here, the specific additions are:

- **For Grayscale:** `Your partner is sufffering from color blindness and will see a grayscale version of your image.`
- **For Cataract:** `Your partner is suffering from cataracts and will see a blurred version of your image along with some black artifacts.`

This modification directly tasks the model with generating descriptions that are robust and interpretable under the specified visual degradation. We report results with Qwen2.5-VL 7B as listener and speaker both.

The effectiveness of this distortion-aware prompting strategy varies significantly depending on the dataset and the nature of the distortion, as shown in Table 11.

On the **CUB dataset**, which features fine-grained classification of birds, the results are mixed. For the **Cataract**, distortion aware-prompting yields a substantial performance gain, increasing accuracy from 63.0% to 80.0%. This suggests that informing the model about the blur and artifacts successfully guides it to focus on more robust features. Conversely, for the **Grayscale** distortion, the prompted accuracy (74.0%) is slightly lower than the normal accuracy (76.0%). In the context of CUB, where bird species are often distinguished by subtle shapes, crests, and beak forms, color may be a less critical feature. The baseline model likely already focuses on these non-color features, and the explicit instruction about color blindness might act as a minor, unhelpful distraction, causing a negligible drop in performance.

On the **CLEVR dataset**, which consists of simple 3D shapes where color is a primary attribute, the impact of distortion aware-prompting is more dramatic and polarized. For the **Grayscale** distortion, the performance without the special prompt

is extremely low (36.0%), as the loss of color information is catastrophic for a task that heavily relies on it. By explicitly telling the model that its partner is colorblind, the prompted accuracy jumps to 50.0%. This significant improvement indicates that the prompt successfully forces the model to pivot its descriptive strategy away from the color information and towards alternative distinguishing features. In a striking reversal, for the **Cataract** distortion, the prompted accuracy plummets to 56.0% from a high normal accuracy of 73.0%. The baseline model, despite the blur. However, when prompted about cataracts, the model's performance collapses. This suggests a form of "over-correction," where the model, anticipating severe information loss, may generate overly simplistic or vague descriptions that fail to distinguish the target, or it may struggle to identify any feature it deems robust enough to mention.

We further examine the generations ot understand why cataract distortion aware prompt fails. We share some sample generations in Table. 12

To better understand the drop in accuracy for the Cataract condition on CLEVR, we qualitatively compare generations produced with and without cataract-aware prompting (Table 12). Three consistent patterns emerge.

(1) Decoding instability. Cataract-aware descriptions frequently include leaked internal tokens or partial loops (e.g., `addCriterion`), often appearing mid-sentence and disrupting otherwise coherent text. These artifacts occur far less frequently in the normal setting, where generations remain mostly intact even when imperfect.

(2) Loss of discriminative detail. Many cataract-aware outputs collapse into vague summaries (e.g., "a simple arrangement of geometric shapes") instead of enumerating object types, counts, or spatial cues. CLEVR depends heavily on such combinatorial details, and their omission removes the very signals needed for disambiguation.

(3) Hallucinated content. Cataract-aware prompting increases scene-independent errors, such as duplicated colors ("blue blue"), extra spheres, or incorrect materials. These hallucinations are especially damaging in CLEVR, where identity depends on simple, atomic properties like color, shape, and size.

Together, these effects indicate that cataract-aware prompting pushes the speaker into an over-cautious, unstable generative mode: it avoids strong commitments, produces underspecified or contradictory descriptions, and becomes more prone to decoding failures. In contrast, the normal prompt elicits structured, object-level descriptions that remain robust even under simulated blur. This disparity explains why distortion-aware prompting *reduces* accuracy for Cataract on CLEVR, despite improving performance in other settings such as CUB.

Cataract-aware (distorted prompt)	**Original (no distortion-aware prompt)**
sample 0	
The image showcases a variety of geometric shapes without mentioning layout or relative complexity.	Image 1 features a diverse collection of shapes with clear differences in layout, density, and object types (gold sphere, gray sphere, green cube), providing strong distinguishing cues.
sample 1	
A collection of shapes with distinct colors and materials. A addCriterion (artifact) The image contains a green cylinder, purple cube, teal cube, teal cylinder, and golden sphere.	Image 1 features a coherent set of objects (green cylinder, purple cube, teal cube, teal cylinder, golden sphere) arranged cleanly on a gray background.
sample 2	
A scene with a yellow sphere, gray sphere, and golden cylinder. addCriterion . . .	A clean description of three distinct objects with correct materials and spacing.
sample 3	
A variety of geometric shapes. Red spheres stand out out their their addCriterion . . .	A detailed structured description listing cubes, spheres, cylinders in specific colors (red, green, yellow, gray, teal) with correct spatial distinctions.
sample 4	
A simple arrangement of geometric shapes on a plain background.	Explicit listing of objects: green sphere, green cylinder, red cylinder, with explanation of spacing and composition.
sample 6	
A variety of shapes including a gray cylinder and red cylinder; "shiny blue blue" addCriterion . . .	A precise object set: large gray cylinder, smaller red cylinder, shiny blue cube, two teal spheres, with a correct comparison of structure and density.

Table 12. Representative CLEVR descriptions with and without cataract-aware prompting. Red = hallucinations, Blue = missing specificity, Purple = decoding artifacts. These error types explain why cataract-aware prompting degrades performance.

(a) Original

(b) Cataract

(c) Age-related Macular Degeneration

(d) Tunnel Vision

(e) Detached Retina

(f) Grayscale

Figure 7. Examples of visual distortions used in our experiments. Each row shows two instances of the same impairment type, with the caption below.

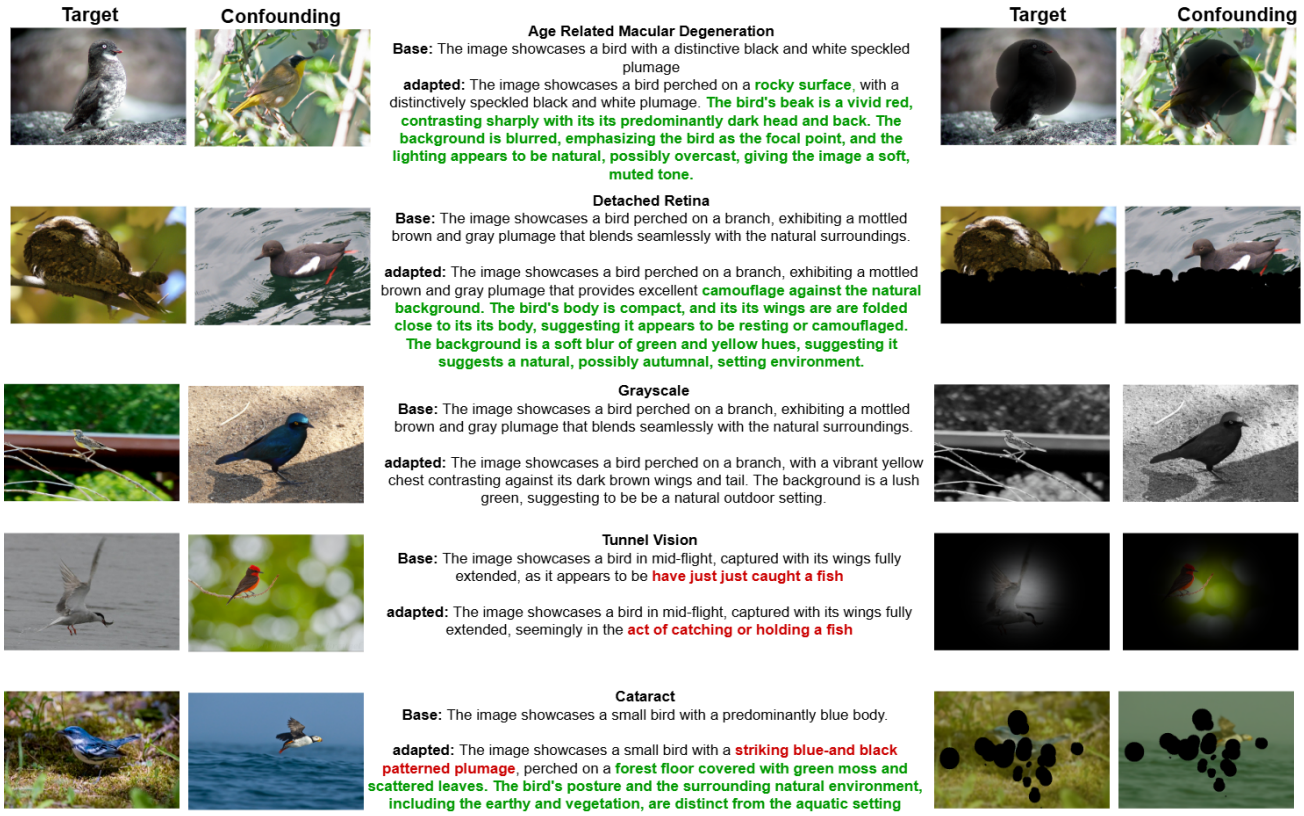


Figure 8. **Qualitative results For GRPO.** We share the qualitative results for GRPO (Shao et al., 2024) where left image is being described in each description. On left we share original images and on right we share distorted images. **Red text** represent change of information from original description while **bold text** represents new information.

E. Perceptual Distortions

F. Hyperparameters

Unless otherwise noted, all models are adapted from `Qwen2.5-VL-7B`. Qwen2.5-VL-7B is used as a listener and a speaker. Furthermore, we also report results using Intervl3.5 in section 5.2. For preference data generation we use Qwen2.5-VL 32B with context length of 2048 tokens.

Dataset-level defaults.

- **CUB:** batch size $= 3$, learning rate $= 5 \times 10^{-6}$, sampling temperature $= 1.0$.
- **CLEVR:** batch size $= 3$, learning rate $= 5 \times 10^{-6}$, sampling temperature $= 1.1$.

Method-specific settings.

- **GRPO (online):** number of generations per instance $G = 4$; micro-batch size $= 1$; sampling temperature follows the dataset default (Imagnet: 1.0, CUB: 1.0, CLEVR: 1.1).
- **DPO (offline):** sampling temperature $= 1.0$; context length $= 2048$.
- **SFT (offline):** sampling temperature $= 1.0$; context length $= 2048$.

G. Adapting Algorithms

To investigate adaptation under divergent visual spaces, we study both *offline* and *online* post-training algorithms. Offline methods allow for cost-effective *ante-hoc* adaptation when it is known in advance that a user population or set of agents shares a specific perceptual limitation. In contrast, online methods adapt through direct interaction with divergent agents, capturing the need for *in situ* personalization. Below we describe each algorithm in relation to the preference dataset $\{(\mathbf{x}, (m^+, m^-))\}$ introduced in Section 2.2, where $\mathbf{x} = (x_{\text{tgt}}, x_{\text{conf}})$ is the undistorted target–confound pair, m^+ denotes the distortion-consistent response, and m^- the distortion-inconsistent response. In all cases, adaptation is applied via LoRA modules for efficiency.

Supervised Fine-Tuning (SFT). SFT provides a baseline for offline adaptation by directly fine-tuning the speaker on $(\mathbf{x}, m^+)$ pairs, ignoring the negative samples:

$$\mathcal{L}_{\text{SFT}}(\pi_\theta) = -\mathbb{E}_{(\mathbf{x},\, m^+)\sim\mathcal{D}} \left[\sum_{t=1}^{|m^+|} \log \pi_\theta(m_t^+ \mid \mathbf{x}, m_{<t}^+) \right]. \tag{1}$$

Direct Preference Optimization (DPO). DPO (Rafailov et al., 2023) directly optimizes the policy on preference pairs without a reward model. Given a reference model π_{ref}, the loss is

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(\mathbf{x},m^+,m^-)} \left[\log \sigma\Big(\beta \Big(\log \tfrac{\pi_\theta(m^+|\mathbf{x})}{\pi_{\text{ref}}(m^+|\mathbf{x})} - \log \tfrac{\pi_\theta(m^-|\mathbf{x})}{\pi_{\text{ref}}(m^-|\mathbf{x})} \Big) \Big) \right], \tag{2}$$

where $\beta > 0$ controls the KL regularization.

Kahneman–Tversky Optimization (KTO). KTO (Ethayarajh et al., 2024) optimizes directly from preference pairs by modeling desirable responses (m^+) and undesirable responses (m^-) with asymmetric sensitivity. The implicit reward is

$$r_\theta(\mathbf{x}, m) = \log \frac{\pi_\theta(m \mid \mathbf{x})}{\pi_{\text{ref}}(m \mid \mathbf{x})}, \tag{3}$$

and the reference point is

$$z_0 = D_{\text{KL}}(\pi_\theta(\cdot \mid \mathbf{x}) \,\|\, \pi_{\text{ref}}(\cdot \mid \mathbf{x})). \tag{4}$$

However, z_0 is intractable. To this end, KTO approximates z_0 as follows:

$$\hat{z}_0 = \max\left(0,\ \frac{1}{B}\sum_{i=1}^{B} \log \frac{\pi_\theta(y_{(i+1) \bmod B} \mid x_i)}{\pi_{\text{ref}}(y_{(i+1) \bmod B} \mid x_i)}\right) \tag{4}$$

The value function is then

$$v(\mathbf{x}, m) = \begin{cases} \lambda_+ \, \sigma\big(\beta(r_\theta(\mathbf{x}, m) - \hat{z}_0)\big), & m = m^+, \\ \lambda_- \, \sigma\big(\beta(\hat{z}_0 - r_\theta(\mathbf{x}, m))\big), & m = m^-, \end{cases} \tag{5}$$

where $\lambda_+, \lambda_- > 0$ control sensitivity to positive and negative samples, $\beta > 0$ adjusts curvature, and σ is the sigmoid. The training objective is

$$\mathcal{L}_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) \ = \ \mathbb{E}_{(\mathbf{x},m^+,m^-)\sim D}\Big[\lambda_m - v(\mathbf{x}, m)\Big]. \tag{6}$$

This formulation captures prospect-theoretic asymmetry by treating distortion-consistent responses (m^+) and inconsistent responses (m^-) differently.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) is an online reinforcement learning algorithm that normalizes rewards across groups. For $\mathbf{x}$, we sample $\{m_i\}_{i=1}^G \sim \pi_\theta(\cdot \mid \mathbf{x})$ with listener rewards $\{r_i\}$. The group-relative advantage is

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \tag{7}$$

The training objective is

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{\mathbf{x},\{m_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Bigg(\frac{\pi_\theta(m_i \mid \mathbf{x})}{\pi_{\theta_{\text{old}}}(m_i \mid \mathbf{x})}A_i, \\ \text{clip}\Big(\frac{\pi_\theta(m_i \mid \mathbf{x})}{\pi_{\theta_{\text{old}}}(m_i \mid \mathbf{x})}, 1-\epsilon, 1+\epsilon\Big)A_i\Bigg)\Bigg] - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}). \tag{8}$$

By leveraging relative comparisons within a group, GRPO stabilizes training while avoiding critic complexity.

H. Extended Qualitative Results

Figure 9. **Qualitative results For KTO (Ethayarajh et al., 2024)** with Qwen2.5-VL 7B as speaker and listener. The left image is being described in each description. On left we share original images and on right we share distorted images. In each description text with color blue represents the method name. **Red text** represent change of information from original description while **green text** represents new information.

To better understand how adaptation manifests in practice, we provide qualitative results qualitative results for two best performing methods in online (KTO (Ethayarajh et al., 2024)) and offline methods (DPO (Rafailov et al., 2023)) in Fig. 9 and Fig. 9 respectively. In Fig. 9 and Fig. 10 the leftmost columns shows the original (target and confounding) images, while the rightmost columns shows the corresponding distorted versions. For more qualitative results, we refer the reader to the appendix H.

We show qualitative results for KTO (Ethayarajh et al., 2024) in Fig. (Ethayarajh et al., 2024). Each description is generated for the target image. Text in **red** indicates changes relative to the base description, and **green text** marks newly introduced information. Across distortions, we observe that speaker systematically adapts by enriching descriptions with context most relevant to the listener's perceptual limitations. For instance, under Age-Related Macular Degeneration (AMD), the adapted model supplements the base description with environmental cues such as the bird standing in shallow water,

thereby leveraging contextual features when fine-grained details are obscured. Similarly, under Detached Retina, adaptation emphasizes the blurred green foliage in the background, aligning the description with the listener's altered visual field. In case of Grayscale, the adapted model highlights structural contrasts (e.g., "thin, bare branch"), compensating for the absence of color cues. In the case of Tunnel Vision, adaptation directs attention to sharp details of the bird's body and posture and emphasizes its presence in flight, providing information inside the narrow visual field of the listener. Finally, for Cataract, the adapted description incorporates references to the blurred green and brown background, again matching the listener's impaired visual input.

We show qualitative results for DPO (Rafailov et al., 2023) (Rafailov et al., 2023). Each description is generated for the target image. Text in **red** indicates changes relative to the base description, and **green text** marks newly introduced information. Across distortions, we observe that DPO (Rafailov et al., 2023) enriches descriptions by emphasizing fine-grained attributes and structural cues that align with the listener's distorted perception. In case of Age-Related Macular Degeneration (AMD), for example, the adapted model introduces new details such as the "bird's beak being slightly open" and the broader "earthy-toned environment", compensating for the loss of central visual detail. For Detached Retina, adaptation emphasizes strong color contrasts (e.g., black head, orange-brown feathers) and background cues, allowing the description to remain informative despite missing peripheral vision. Additionally, emphasize on the upper part of the body of bird (e.g., black head) is also relevant to the listener because the body of the bird is occluded. In case of Grayscale, the adapted model highlights sharper contrasts in structure and position (e.g., "sharp focus on the bird against a softly blurred background" instead of blurred "green" background), directly addressing the absence of color cues. In the case of Tunnel Vision, descriptions adapt by focusing on body posture and motion, noting that the bird "dives towards the water" and has a "streamlined body", thereby ensuring that critical motion cues remain visible within the narrow field of vision. Additionally, we also notice that speaker changes 'bird just caught a fish' to bird diving towards water since the fish is not clearly visible in the narrow field of vision. Finally, under "Cataract", the adapted outputs include both fine visual traits (e.g., "short, pointed beak") and environmental grounding "natural, outdoor setting"), effectively compensating for the hazy appearance caused by blurred input. These qualitative results show that DPO adaptation emphasizes discriminative features and contextual cues in ways that directly address distortion-specific limitations, complementing the quantitative findings where DPO often achieved the highest peak improvements. We also observe the level of details provided is more than KTO9 as the speaker is able to mention fine-grained details such as "bird's beak being slightly open". This further validates that in case of DPO10, speaker is benefittings strongly by distilling from a much stronger model with access to listener specific distortion.

Overall, these qualitative examples confirm that speaker can adapt not just by reformatting but by prioritizing perceptual attributes most relevant to listener in both online and offline settings.

H.1. Failure Cases

In order to further validate our findings, We share failure cases in this section. We observe that in failure cases the adapted model produces descriptions that are misaligned with the listener's view (Fig. 11). In case of grayscale, the adapted model continues to emphasize color attributes under, such as describing "vibrant yellow plumage" or "subtle hints of green on the wings," while the listener cannot access color information. Another observation is the loss of contextual grounding, where the adapted description becomes overly concise and omits background cues or environmental context that could support disambiguation, especially under cataract or detached retina distortions. These errors reduce the interpretability and usefulness of adapted outputs, as the speaker either relies on inaccessible features or provides insufficient detail.

I. Datasets

To evaluate adaptation under divergent visual perceptions, we conduct experiments on two established multimodal benchmarks: **CLEVR** (Johnson et al., 2017) and **CUB** (Wah et al., 2011). These datasets offer complementary characteristics, enabling us to test both compositional reasoning and fine-grained visual recognition in our reference game setup.

CLEVR. (Johnson et al., 2017) provides synthetic scenes with varied shapes, colors, materials. CLEVR contains 100k images, each depicting 3–10 objects rendered from a factorial combination of shapes, sizes, materials, and colors, with splits of 70k/15k/15k for training, validation, and testing.

CUB. (Wah et al., 2011) provides natural, fine-grained imagery across 200 bird species. We form target–distractor pairs from visually similar birds to increase identification difficulty. The dataset includes 11,788 images, with roughly half

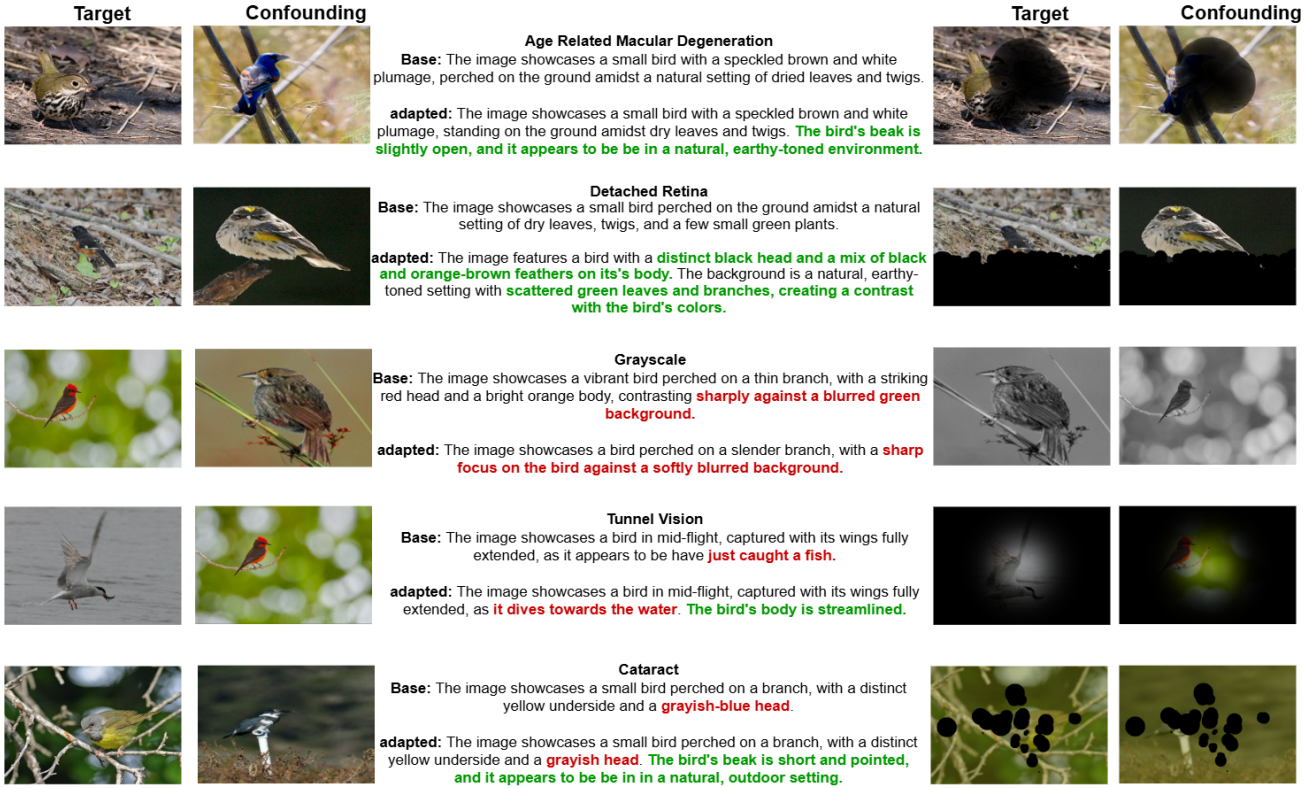


Figure 10. **Qualitative results For DPO** with Qwen2.5-VL 7B as speaker. The left image is being described in each description. On left we share original images and on right we share distorted images. In each description text with color blue represents the method name. **Red text** represent change of information from original description while **green text** represents new information.

used for training and half for testing, each annotated with bounding boxes, part keypoints, and over 300 binary attributes describing appearance and morphology.

Imagenet Imagenet is a large-scale benchmark for image classification containing over 1.2 million training images and 50,000 validation images across 1,000 object categories. The dataset is organized according to the WordNet hierarchy, where each class corresponds to a semantic synset. ImageNet has become a standard benchmark for evaluating representation learning and visual recognition models due to its scale and diversity of object categories.

J. Linguistic Analysis

To better understand how adaptation methods alter the speaker's communication strategy, we conduct a linguistic analysis of the generated descriptions. This analysis reveals the specific ways in which adapted models shift their language to be more robust to the listener's perceptual limitations, providing quantitative evidence that complements our accuracy-based results.

J.1. Quantitative Linguistic Shifts Across All Distortions

To complement our lexical analysis, we quantify how adaptation alters the readability and structural complexity of generated descriptions across distortions. We report two metrics—average word count and Flesch Reading Ease (FRE)—as shown in Fig. 12 (KTO) and Fig. 13 (GRPO). FRE is a standard measure of linguistic complexity defined as:

$$\text{FRE} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right), \tag{9}$$

where higher scores indicate simpler text, and lower scores reflect denser or more complex phrasing.

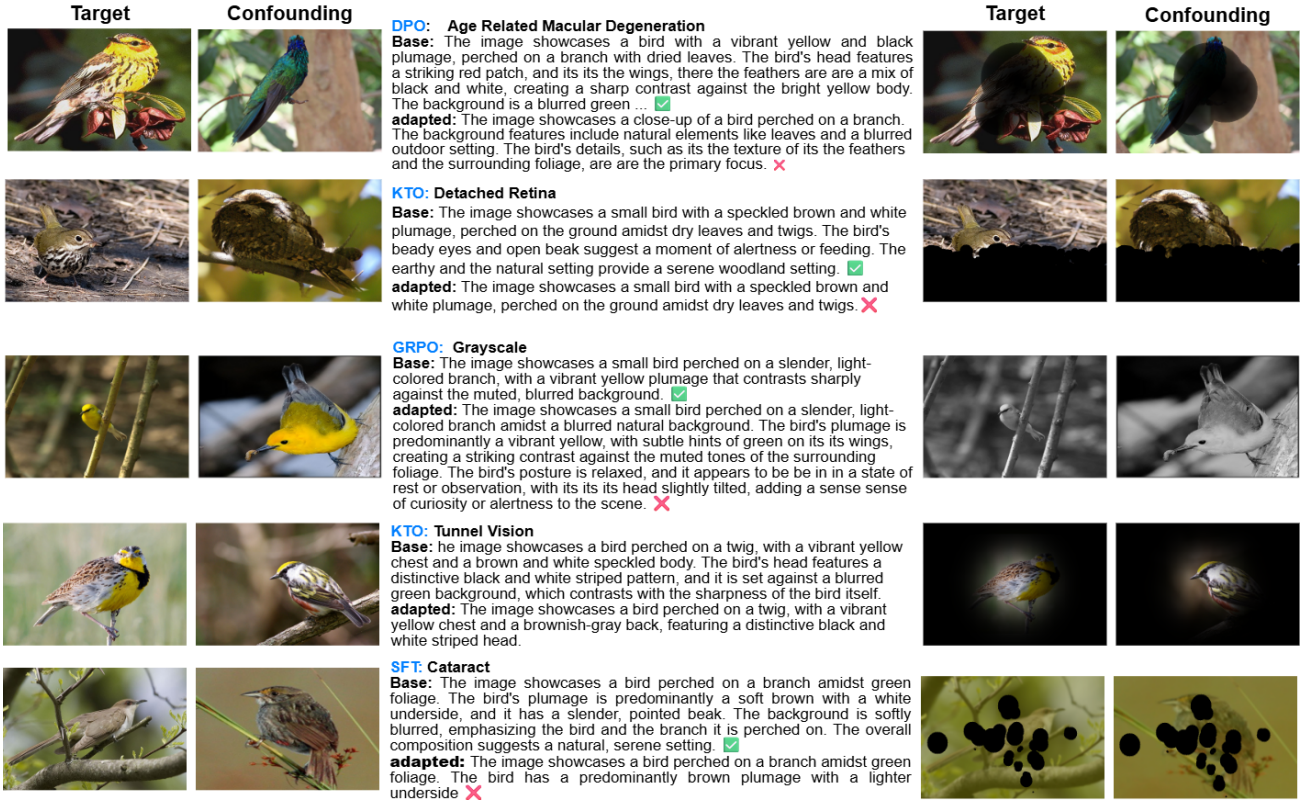
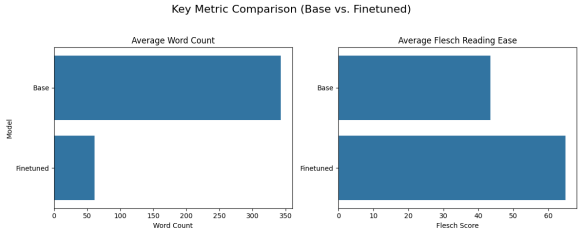


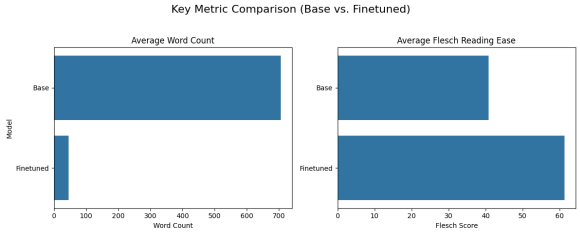
Figure 11. **Failure Cases** with Qwen2.5-VL 7B as speaker and listener. The left image is being described in each description. On left we share original images and on right we share distorted images. In each description tick represents correct and cross represents incorrect description. These are all the cases where adapted model fails and base model is successful.

Across distortions, we observe consistent and distortion-dependent shifts following adaptation:

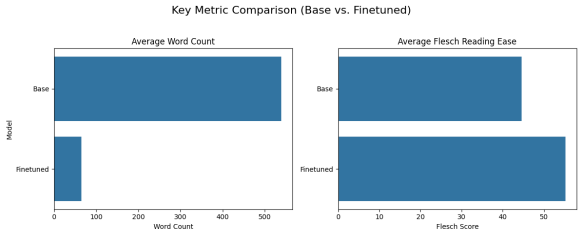
Overall Linguistic Shifts. Across all distortion types, we observe a consistent pattern in both KTO and GRPO: the Flesch Reading Ease (FRE) scores increase while the average word count decreases after adaptation. This indicates that adapted speakers produce shorter and more readable descriptions, reflecting a systematic move toward clearer and more efficient communication. Furthermore, we observe higher flesch scores for KTO achieving around 60 score in 3/4 distortions. On the other hand, we observe slightly lower flesch scores in GRPO except in Grayscale. In all cases, we observe word count decreases after adaptation. This shows that KTO aligns the model more robustly for communication with listener by making it more readable as compared to GRPO. Importantly, this shift suggests that adaptation encourages the model to rely on simpler, more direct phrasing that remains robust to perceptual impairments, rather than attempting to compensate with longer or more complex sentences.



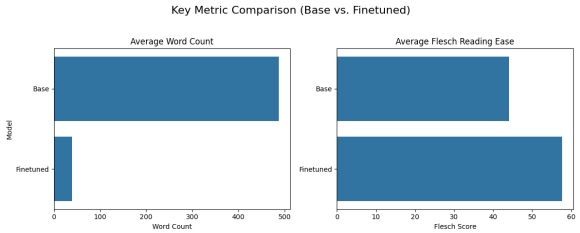
(a) Catract



(b) Tunnel Vision

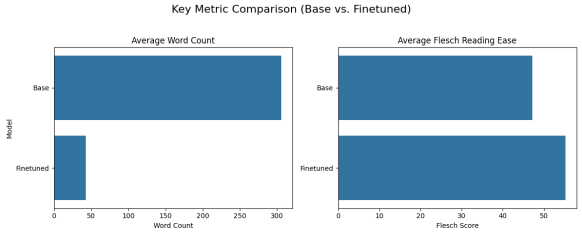


(c) Detached Retina

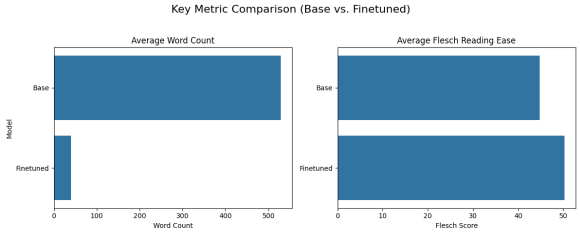


(d) Grayscale

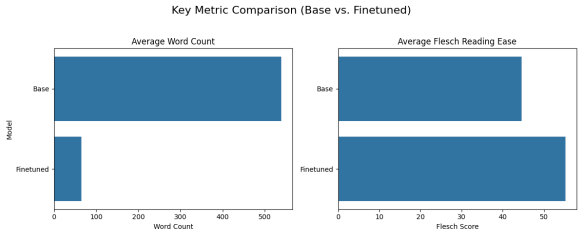
Figure 12. Shows how the text is changing before and after finetuning for KTO.



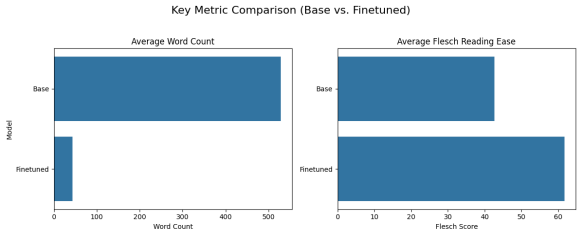
(a) Catract



(b) Tunnel Vision



(c) Detached Retina



(d) Grayscale

Figure 13. Shows how the text is changing before and after finetuning for GRPO.