

Silence over Error: Risk-Averse Navigation Landmark Extraction via Progressive MLLM Alignment

Xin Yang Ziqi Cao Changning Shao Junhao Lin

Fanxiang Zeng* Yue Liu*

Amap, Alibaba Group

{xiyi.yx,ximeng.czq,jeffrey.scn,bianliang.ljh}@alibaba-inc.com

{fanxiang.zfx,yue.liu}@alibaba-inc.com

Abstract

Existing pedestrian navigation systems often provide direction-only instructions with limited visual confirmation at complex intersections or along long straight paths. We study coordinate-free landmark extraction from dashcam image sequences. The model returns a set of landmarks with structured attributes (e.g., category and description), or abstains when visual evidence is insufficient. Although general MLLMs can recognize street-scene objects, they do not reliably follow expert rules for filtering ambiguous candidates. Moreover, a false landmark can mislead a pedestrian, making precision more important than exhaustive coverage in this task. We therefore develop a progressive alignment framework, supported by a landmark annotation and augmentation pipeline that converts navigation experts’ selection criteria into staged supervision signals. We further propose a hard-gated bipartite matcher for box-free evaluation and a navigation-aware risk-sensitive reward that penalizes false positives more heavily than abstentions without inducing blanket silence. On the held-out benchmark, our model achieves 92.29% precision while maintaining practical coverage, outperforming strong MLLM baselines. A 30-day A/B test with 2M users on the Amap platform further shows that landmark-augmented instructions reduce departure-confirmation latency by 23.5%. Together, these results demonstrate that risk-aware alignment enables reliable, abstention-capable landmark extraction for navigation.

1 Introduction

Existing pedestrian navigation systems (Wang et al., 2022a; Yang et al., 2026) largely replicate vehicle-oriented turn-by-turn paradigms. At complex intersections and during long straight walks, such instructions provide little visual confirmation; absolute-direction commands are also counter-intuitive and difficult to perceive in real time. For

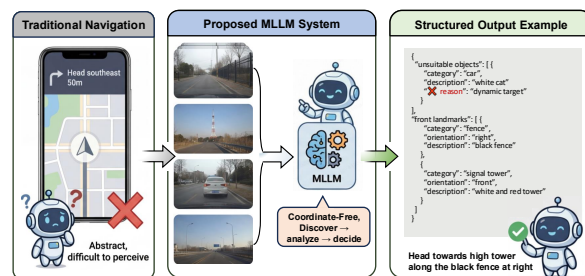


Figure 1: Traditional abstract navigation instructions vs. landmark-augmented guidance system. Offline, the model maps a dashcam image sequence to a structured landmark prediction or an empty output (silence), and associates the result with its road segment ID. Online, a key–value lookup by road ID retrieves the stored landmark, which is combined with an instruction.

pedestrians, instructions such as “walk 50 meters along the black fence on the right and then turn left at the panda statue, with a distant signal tower visible ahead” are more intuitive than abstract commands like “head southeast.” Although prior works on Navigation Instruction Generation (Chen et al., 2019; Wang et al., 2022b; Cui et al., 2026) and Vision-and-Language Navigation (Lin et al., 2024; Xu et al., 2025; Tian et al., 2024) ground navigation language in visual observations, they do not directly address safety-filtered structured extraction of a broader category of visually salient non-POI entities such as large road signs. Recent MLLMs (OpenAI et al., 2023; Gemini Team et al., 2023; Wang et al., 2025b; Bai et al., 2025) make end-to-end extraction feasible, yet off-the-shelf use remains unsafe for guidance: SceneScout (Jain et al., 2026) reports errors in roughly 28% of descriptions, including plausible hallucinations.

Building a usable MLLM landmark extractor requires addressing three coupled issues.

Navigation-grounded structured landmark extraction. Although image captioning is a strong suit of MLLMs, the core difficulty lies not in recognition, but in deciding which objects are suitable

*Corresponding authors.

for guidance. The model must learn navigation-specific extraction criteria, operationalized by human navigation experts, including: (1) landmark suitability criteria such as saliency and uniqueness; (2) precise appearance descriptions that remain comprehensible and concise for voice broadcasting in noisy outdoor environments while highlighting the core distinctive features; and (3) reliable structured attributes following unified standards (e.g., category and relative orientation). Furthermore, existing street-view images are mainly sourced from dashcams rather than pedestrian perspectives, and this viewpoint discrepancy coupled with inconsistent image quality further complicates learning such extraction criteria. This form of task-specific alignment is difficult to obtain through prompt engineering alone and requires systematic fine-tuning.

We instantiate these criteria through expert annotations organized into ten rejection categories (e.g., “dynamic target,” “invisible from pedestrian paths”), enabling the model to explicitly learn not only “what to predict” but also “why to reject.” We operationalize these criteria through a landmark annotation and augmentation pipeline that turns expert pass/reject judgments into staged alignment signals. Expert annotations and rejection categories support multi-task SFT, spatio-temporal cropping augments both positive and zero-landmark cases, contrastive pair construction supplies DPO preferences, and GRPO uses online rollouts scored by the matching-based risk-sensitive reward.

The resulting extractor is deployed through the offline-to-online pipeline shown in Figure 1. During offline batch inference, it maps each dashcam image sequence to an empty output or a structured landmark prediction which is stored with road segment IDs in a key-value index. During online serving, the system retrieves the stored landmark and combines its attributes with an instruction template. If no landmark is available, the original direction-only instruction is retained.

Coordinate-free evaluation. Object-detection systems rely on Intersection-over-Union (IoU) between bounding boxes for both training and evaluation. However, box annotation is expensive on long sequential dashcam frames, ill-defined for landmarks with ambiguous boundaries (e.g., fences, distant mountains), and ill-suited to current MLLMs whose precise spatial localization remains limited. Text-generation metrics such as BLEU/ROUGE-style n-gram overlap may reward fluent lexical similarity while missing navigation-critical attribute

bindings, such as “a red building with a green roof” vs. “a green building with a red roof.”

We therefore commit to a fully coordinate-free formulation, in which each prediction and each ground truth is a tuple of discrete and textual attributes rather than a spatial region. For this setting, we design a Hungarian-based matching scheme (Carion et al., 2020; Zhang et al., 2023) that fuses semantic similarity with multi-dimensional discrete attribute scores under a cascade of hard gates that block spatially or categorically incompatible pairs. The same matcher serves both offline evaluation and GRPO reward scoring, avoiding reward-metric mismatch.

Risk-sensitive abstention. MLLMs can hallucinate objects and visual attributes, including incorrect colors, positions, and relationships (Zhou et al., 2024; Li et al., 2023; Huang et al., 2024). In navigation, such errors—for example, a wrong color or left/right prediction—can mislead users. In other words, in this task false positives are far more harmful than false negatives. When faced with complex visual input, MLLMs used for landmark extraction should not only recognize objects but also maintain *restraint*—proactively avoiding high-risk targets. As precision is more important than recall, an intuitive solution is to rigidly restrict the model to predict only a minimal number of the most reliable landmarks. However, attempting to force the model to acquire this trait through traditional “imitation learning” paradigms would require manually covering a large number of long-tail scenarios and landmark combinations, as well as setting complex priority rules and various thresholds, which is impractical in open-world scenarios.

We therefore introduce the Navigation-Aware Risk-sensitive Reward Function (NARF) for Group Relative Policy Optimization (GRPO) (Shao et al., 2024): based upon the matching score, NARF penalizes false positives more strongly than abstention while rewarding high-confidence correct outputs. After extensive rollout exploration, a risk-averse policy emerges—the model learns to forego marginal predictions that carry high risk in cluttered scenes, confidently outputting only the most reliable landmarks, or abstaining when the evidence is insufficient. For instance, the model learns that predicting a building in densely built-up CBD areas is highly likely to incur penalties (high risk, low reward), whereas extracting large road signs with textual information on arterial roads is a stable and unambiguous choice (low risk, high reward).

Our contributions are threefold:

- We design a progressive alignment framework that converts expert landmark-selection criteria into staged supervision, preference, and reinforcement signals through landmark annotation and augmentation.
- We couple a soft-hard bipartite matcher with NARF for box-free evaluation and precision-oriented optimization without trivial silence.
- We validate the system on a real-world street-view benchmark and a 30-day A/B test with approximately 2M users, showing higher precision and lower wayfinding latency.

2 Related Work

2.1 Landmarks in Pedestrian Navigation

Landmarks provide salient spatial reference points that help pedestrians localize themselves and confirm orientation at decision points (Lynch, 1960; Millonig and Schechtner, 2007; Hile et al., 2009; Yesiltepe et al., 2021). Human-centered studies link effective landmark guidance to salience, user orientation skills, instruction timing, and the cognitive stages of outdoor wayfinding (Röser, 2015; Ohm et al., 2016; Golab et al., 2022; Alinaghi et al., 2025); these findings contextualize, rather than define, our operational annotation criteria.

Navigation Instruction Generation methods incorporate landmarks through different forms of supervision: Touchdown (Chen et al., 2019) provides human-authored street-navigation instructions, MARKY-MT5 (Wang et al., 2022b) and FCA-NIG (Cui et al., 2026) use localized landmark supervision, while Text360Nav (Nishimura et al., 2024) provides scene captions without structured extraction. Outdoor Vision-and-Language Navigation methods (Lin et al., 2024; Xu et al., 2025; Tian et al., 2024) leverage language, visual observations, and multimodal spatial reasoning, but do not target safety-filtered structured attributes. Across these lines of work, three gaps remain: (1) localized supervision often depends on costly bounding boxes ill-suited for ambiguous entities like fences or mountains; (2) limited coverage of salient non-POI objects; and (3) absence of an explicit *abstention* mechanism for false-positive control.

2.2 MLLM Preference Alignment

Preference alignment has evolved from RLHF (Ouyang et al., 2022) and offline

DPO (Rafailov et al., 2023) to critic-free policy optimization variants (Shao et al., 2024; Yu et al., 2025; Zheng et al., 2025; Liu et al., 2025; Ahmadian et al., 2024). Multimodal preference alignment and GRPO-style optimization have recently been explored for generation, reasoning, and safety (Liang, 2025; Xue et al., 2025; Rong et al., 2026; Zhang et al., 2025b; Luo et al., 2025; Yi et al., 2025), as well as navigation instruction generation (Zhao et al., 2026); these works do not target coordinate-free spatial prediction accuracy. Related work also studies risk-sensitive and multi-objective optimization (Jiang et al., 2026; Zhang et al., 2025a; Li et al., 2025), but does not define rewards over visual landmark matches or navigation-specific false-positive costs.

None of these address the asymmetric risk of navigation landmark prediction, where a false positive is far more harmful than a missed detection. A complementary NLP line teaches models to abstain rather than fabricate (Wen et al., 2025; Cao, 2024), but mainly studies text-only refusal or factual QA settings. NARF encodes silence-vs-error asymmetry directly into a hierarchical RL objective over coordinate-free bipartite matching, turning abstention into an emergent policy for visual grounding where wrong predictions carry navigation risk.

3 Our Approach

3.1 Task Formulation

Given a temporally ordered sequence of T dashcam images $\mathcal{I} = \{I_1, I_2, \dots, I_T\}$, the model π_θ is required to produce a structured prediction set:

$$\pi_\theta(\mathcal{I}) \rightarrow \hat{\mathcal{Y}} = \{\hat{y}_k\}_{k=1}^{\hat{N}}, \quad (1)$$

$$\hat{y}_k = (d_k, c_k, o_k, \delta_k, [s_k, e_k])$$

where $\hat{N} \geq 0$ is the prediction count ($\hat{N} = 0$ denotes abstention), with fields: description d_k , category c_k , orientation $o_k \in \{\text{left}, \text{right}, \text{front}\}$, distance $\delta_k \in \{\text{near}, \text{far}\}$, and visibility range $[s_k, e_k] \subseteq [1, T]$. No bounding boxes are involved in either annotation or prediction.

3.2 Progressive Alignment Framework

We use a multi-stage alignment framework, supported by a landmark annotation and augmentation pipeline, to convert navigation experts’ landmark-selection criteria into staged learning signals.

Annotation Protocol.

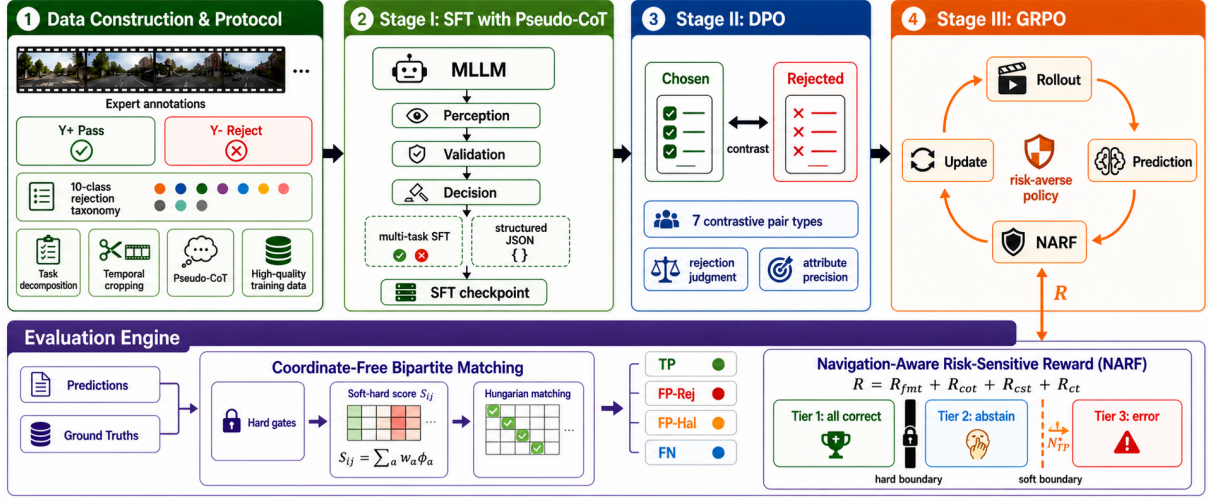


Figure 2: Overview of the proposed system. **(1) Data:** Expert annotations with rejection taxonomy, cropping augmentation and pseudo-CoT. **(2) SFT:** Multi-task training with joint pos/neg supervision. **(3) DPO:** Contrastive pairs for attribute refinement. **(4) GRPO:** Online exploration driven by NARF with asymmetric boundaries.

The dataset was annotated by eight internal navigation-domain experts with long-term experience. A shared guideline operationalized landmark suitability through saliency, uniqueness, and concise descriptions. It contained frequent positive examples, predefined rejection categories with three to five examples each. The team refined the guideline and its examples through three rounds of pilot annotation, joint review of disagreements, and discussion of difficult cases.

Initial candidate proposals are sampled at temperature 1.0 from multiple zero-shot frontier MLLMs (Qwen3-VL and Gemini-2.5-Pro), with later rounds using the latest SFT checkpoint as the proposal generator. Expert annotators are instructed to *exhaustively* identify all usable landmarks in each scene: they select, revise, and label candidates as *pass* $\mathcal{Y}^+$ or *reject* $\mathcal{Y}^-$ (with a categorized reason), merge duplicates, and manually add missing salient landmarks.

This exhaustive annotation strategy mitigates the impact of subjective selection criteria on dataset consistency by retaining *all* non-violating objects rather than cherry-picking “the best” ones. Furthermore, it simplifies evaluation: any model prediction that fails to match either a pass or reject annotation will be treated as a hallucination under this annotation scheme. At the same time, this protocol yields over-complete ground-truth (GT) sets, often containing more valid landmarks than navigation needs and thus amplifying post-finetuning over-prediction (addressed by NARF in Sec. 3.4).

Rejection reasons are consolidated into ten categories (e.g., *insufficient saliency*, *indistinct appearance*; full taxonomy in App. G), providing supervision for SFT, contrastive signal for DPO, and reward shaping for GRPO. Additionally, rejected objects participate in bipartite matching (Sec. 3.3), distinguishing predictions matching an expert-rejected object (FP-Rej, *recorded hallucination*) from those matching nothing (FP-Hal, *novel hallucination*)—both harmful, but the former indicates that the model has learned spurious patterns from the annotation data itself.

Stage I: Multi-Task SFT. To avoid the model merely memorizing input-output patterns, we decompose the end-to-end extraction into four auxiliary tasks—*pass/reject classification*, *saliency ranking*, *range prediction*, and *description generation*—trained jointly with the primary end-to-end task under a shared prompt (full task definitions in App. H). Temporal cropping further augments data and teaches abstention: each sample with $T \geq 3$ frames uses a random offset $\tau \sim \text{Uniform}(1, T-1)$ to form a suffix subsequence, with object ranges adjusted accordingly (Eq. 14 in App. H); objects no longer visible after cropping are discarded.

Zero-landmark samples come from two sources: (i) naturally empty scenes (e.g., highways, blurred images) annotated as $\mathcal{Y}^+ = \emptyset$, and (ii) cropping-induced empties where all landmark-containing frames are removed. These provide explicit supervision for “output nothing” behavior.

For the primary task, we prepend a synthetic CoT to each target, following a fixed *discover*→*analyze*→*decide* template constructed from annotations, forcing the model to execute a complete verification process before output.

Stage II: Multi-Dimensional DPO. We construct seven types of contrastive preference pairs targeting recall completeness, rejection judgment (false acceptance), range/orientation precision, rejection thoroughness, false rejection, and output conciseness (duplicate predictions); each combines with cropping augmentation. Full pair construction details are in App. I.

Stage III: Risk-Sensitive GRPO. The final stage applies GRPO with our NARF reward (Sec. 3.4) to encode the asymmetric cost structure, which SFT and DPO cannot directly express.

3.3 Coordinate-Free Bipartite Matching

In object detection, predicted and GT boxes are matched via IoU. Our task produces no spatial coordinates; instead, each prediction and each ground truth is a tuple of discrete and textual attributes. We design a *soft-hard fused* matching scheme on the Hungarian algorithm (Carion et al., 2020; Zhang et al., 2023), serving as both (i) the offline evaluation metric and (ii) the GRPO reward scoring.

Hard-Gated Cost Matrix. For each pair $(\hat{y}_i, y_j)$, a cascade of five hard gates blocks the pair ($S_{ij} = -1$) if any fires: (1) *type mismatch* ($c_i \neq c_j$); (2) *lateral conflict* ($\{o_i, o_j\} = \{L, R\}$); (3) *range disjoint*; (4) *color conflict*, when both sides specify conflicting single colors (e.g., “red shop” vs. “blue shop”); (5) *semantic floor*, $f_{\text{sem}}(d_i, d_j) < \tau_1$ where f_{sem} is the reranker (Wang et al., 2025a) similarity and τ_1 is a floor threshold below which the pair is deemed unrelated. The full formal definition is in App. E.1 (Eq. 7).

Soft Similarity Scoring. For pairs that pass all hard gates, we compute a weighted combination of attribute-level scores:

$$S(\hat{y}_i, y_j) = \sum_{a \in \mathcal{A}} w_a \phi_a(\hat{y}_i, y_j) \quad (2)$$

where $\mathcal{A} = \{\text{sem}, \text{typ}, \text{dst}, \text{rng}, \text{dir}, \text{clr}\}$, $\phi_{\text{sem}} = f_{\text{sem}}(d_i, d_j)$ carries the largest weight ($w_{\text{sem}}=0.4$) as it subsumes color, shape, and context; $\phi_{\text{typ}} = 1$ (verified by hard gate); $\phi_{\text{dst}} = \mathbb{1}[\delta_i = \delta_j]$.

Range overlap ϕ_{rng} adopts a tiered design rewarding *conservative* predictions (range subsumed by GT) over merely overlapping ones: exact match $\rightarrow 1.0$; pred subsumed by GT $\rightarrow 0.3 + 0.7\rho$; partial

overlap $\rightarrow 0.1 + 0.2\rho$; disjoint $\rightarrow 0$, where ρ is the temporal IoU (Eq. 8 in App. E). A prediction in frames 2–3 vs. GT in 1–4 is *safe*, but the reverse may reference a different object.

Orientation matching ϕ_{dir} encodes asymmetric error severity: exact match $\rightarrow 1.0$; left↔right confusion $\rightarrow -0.2$ (most dangerous, opposite side of road); front-side confusion $\rightarrow 0.5$ (partial credit).

The *color soft score* ϕ_{clr} provides a mild bonus ($w_{\text{clr}}=0.1$) when both sides have exactly one single-category, single-color mention and the colors agree; otherwise it defaults to zero.

Optimal Assignment and Classification. We assemble the $\hat{N} \times M$ similarity matrix $\mathbf{S}$, set $S_{ij} = -1$ for blocked pairs, and solve a maximum-weight partial assignment via the Hungarian algorithm:

$$\sigma^* = \arg \max_{\sigma \in \Pi_{\emptyset}} \sum_{i: \sigma(i) \neq \emptyset} S_{i, \sigma(i)} \quad (3)$$

where $\Pi_{\emptyset}$ allows one-to-one prediction–candidate matches and $\sigma(i) = \emptyset$ for unmatched predictions. Each prediction $\hat{y}_i$ is then classified:

$$\hat{y}_i \mapsto \begin{cases} \text{TP}, & S_{i, \sigma^*(i)} \geq 0 \wedge y_{\sigma^*(i)} \in \mathcal{Y}^+ \\ \text{FP-REJ}, & S_{i, \sigma^*(i)} \geq 0 \wedge y_{\sigma^*(i)} \in \mathcal{Y}^- \\ \text{FP-HAL}, & \sigma^*(i) = \emptyset \text{ or } S_{i, \sigma^*(i)} < 0 \end{cases} \quad (4)$$

Unmatched pass ground truths are labeled as FN.

3.4 Navigation-Aware Risk-Sensitive Reward

To further align the model with navigation-specific risk preferences via GRPO, a straightforward approach would be to adopt a standard additive reward, which linearly sums TP bonuses and FP penalties. Such rewards encourage over-prediction: high-quality TPs can offset FPs, which is unsafe under the post-SFT over-prediction tendency (Sec. 3.2). NARF instead partitions outputs into three quality tiers with *asymmetric boundaries*: FP-free non-empty outputs are unconditionally separated from silence and erroneous outputs, while the silence-vs-error boundary is deliberately *soft* to prevent reward hacking via blanket abstention.

3.4.1 Three-Tier Content Reward

Let $M^+ = |\mathcal{Y}^+|$ denote the number of accepted GTs and $\hat{N}_{\text{FP}}$ the total false positives (FP-REJ + FP-HAL) from Sec. 3.3. We partition outputs into three mutually exclusive tiers:

$$R_{\text{ct}} = \begin{cases} r_1 + \sum_{k \in \text{TP}} Q_k & \hat{N} > 0, \hat{N}_{\text{FP}} = 0 \\ \quad + r_{\text{all}} \mathbb{1}[\hat{N} = M^+] & \\ r_1 \mathbb{1}[M^+ = 0] + r_2 \mathbb{1}[M^+ > 0], & \hat{N} = 0 \\ \min(r_3 + P_{\text{err}} + \alpha \sum_{k \in \text{TP}} Q_k, & \hat{N}_{\text{FP}} > 0 \\ \quad r_{\text{ceil}}) & \end{cases} \quad (5)$$

Tier 1 (FP-free predictions) sums a positive base r_1 , per-TP quality scores Q_k (defined below), and a bonus r_{all} for predicting *all* pass objects. **Tier 2** (silence) penalizes *empty predictions* with a mild $r_2 < 0$, but yields r_1 when no valid landmark exists, rewarding correct abstention. **Tier 3** (any FP) starts from a punitive base $r_3 < r_2 < 0$, discounts any TP quality by $\alpha < 1$, and caps the total reward at r_{ceil} . The strict ordering $r_3 < r_2 < r_1$ together with the α discount and ceiling jointly enforce the asymmetric tier separation analyzed in Sec. 3.4.3.

3.4.2 Per-Prediction Quality Score

The per-TP quality Q_k combines (i) a base TP reward, (ii) a reranker-similarity bonus with two-segment piecewise scaling that rewards predictions exceeding a high-quality threshold more generously, and (iii) four attribute penalties over color, direction, distance, and range. The penalty structure encodes navigation-specific severity: a left $\leftrightarrow$ right confusion is penalized more heavily than a front $\leftrightarrow$ side one, reflecting that lateral errors are more disorienting for pedestrian wayfinding. The Tier-3 error term decomposes as $P_{\text{err}} = \hat{N}_{\text{FP-REJ}} p_{\text{rej}} + \hat{N}_{\text{FP-HAL}} p_{\text{hal}}$ with $|p_{\text{rej}}| > |p_{\text{hal}}|$, encoding the navigation principle that violating an explicit rejection criterion is more harmful than a generic hallucination. Full forms are in App. F.

3.4.3 Asymmetric Tier Boundary Design

Hard Boundary: Tier 1 vs. Tiers 2–3. We enforce $\min R_{\text{ct}}^{(1)} > \max R_{\text{ct}}^{(2,3)}$, so no silent or erroneous output can outscore a clean non-empty output; GRPO’s group-relative advantage then yields an unambiguous gradient toward FP-free outputs.

Soft Boundary: Tier 2 vs. Tier 3. In contrast, we set $r_{\text{ceil}} > r_2$, allowing a Tier-3 output to exceed silence only when its TP count exceeds a *critical count* obtained by equating the two tier rewards:

$$N_{\text{TP}}^* = \left\lceil \frac{r_2 - r_3 - P_{\text{err}}}{\alpha \cdot \bar{Q}} \right\rceil \quad (6)$$

with $\bar{Q}$ the average per-TP quality estimated once from DPO-checkpoint rollouts before GRPO. This

preserves strong risk aversion when errors dominate while permitting confident predictions in unambiguous scenes. A strict hard boundary ($r_{\text{ceil}} < r_2$) would instead invite blanket abstention; we analyze this *silence collapse* phenomenon and its resolution in Sec. 4.3 (Fig. 3).

3.4.4 Format and Consistency Rewards

The final GRPO reward sums the content term with three structural terms: $R = R_{\text{fmt}} + R_{\text{cot}} + R_{\text{cst}} + R_{\text{ct}}$, where R_{ct} is the tiered content reward of Eq. (5). R_{fmt} validates JSON parseability and enum legality, returning a dominating negative $p_{\text{parse}} \ll 0$ on parse failure; R_{cot} grants capped bonuses for the presence of reasoning keywords (*discover*, *analyze*, *decide*) in the CoT prefix; R_{cst} penalizes mismatches between the number of “accept” decisions in the CoT, the JSON object count, and the declared proper_obj_num field.

4 Experiments

4.1 Experimental Setup

Datasets and Implementation Details We curate over 2,500 dashcam image sequences across diverse driving scenarios. Expert annotators provide *pass* and *reject* annotations, with rejects grouped into ten categories; 240 held-out sequences form the evaluation benchmark. We evaluate Qwen3-VL-32B/8B-Instruct (Bai et al., 2025), using LoRA (Hu et al., 2022) (rank = 16, $\alpha = 32$) on all linear layers while freezing the vision encoder and aligner. SFT stage uses $\sim 12.5\text{k}$ multi-task samples; DPO uses $\sim 6.8\text{k}$ contrastive pairs; GRPO performs online rollout on $\sim 3.5\text{k}$ prompts with $G=16$ rollouts each. Detailed hyperparameters are in App. A.

Metrics Predictions are evaluated coordinate-free via bipartite matching (Sec. 3.3). **Recall** and **Precision** measure accepted-GT recovery and prediction correctness; **FP-Rej** / **FP-Hal** are the per-prediction rates of rejection-criterion violations and unmatched hallucinations classified by Eq. (4); **Dir. Acc** / **Dist. Acc** report the fraction of TPs with correct orientation and distance labels. Because pass annotations deliberately retain all acceptable landmarks, Recall measures redundant candidate coverage rather than segment usability. One reliable landmark can be sufficient for the origin instruction, whereas one false landmark can mislead a user. We therefore treat Precision and FP rates as primary offline risk proxies and Recall as a secondary coverage metric.

4.2 Main Results

Table 1 summarizes the baselines and three progressive alignment stages. All trainable rows use Qwen3-VL-32B-Instruct; the 8B replication in App. C follows the same trend. Closed-source baselines retain FP-Hal rates above 30%, while Stage-I SFT raises Precision to 83.45% and reduces FP-Hal to 11.82%. DPO further raises Precision to 87.58%. GRPO with NARF reaches 92.29% Precision with FP-Rej at 2.48% and FP-Hal at 5.23%, while Recall decreases to 45.52% as the model suppresses marginal and redundant candidates.

In addition, we estimated uncertainty using 20,000 sequence-level paired-bootstrap replicates over the 240 held-out sequences. Final SFT obtains 64.29% Recall (95% CI: 61.09–67.63) and 83.45% Precision (80.24–86.58); GRPO-NARF obtains 45.52% Recall (42.14–48.91) and 92.29% Precision (88.99–95.30).

The original matcher and GRPO reward share a frozen Jina reranker. We therefore reevaluated fixed model predictions with GTE-rerank-v2, which was not used in training, while leaving the hard gates and structured scores unchanged. Without threshold retuning ($T_1/T_2 = 0.2/0.4$), GRPO-NARF retains the highest Precision among the models in this sensitivity study (87.62% Precision, 50.56% Recall). Complete results and the separately labeled post-hoc calibration are reported in App. B.

4.3 Ablation Studies

We ablate the progressive stages in Table 1 to validate key design choices. Furthermore, to analyze the behavioral shifts during GRPO, we track custom rollout metrics—specifically the ratio of rollouts falling into different Tier classifications and the average number of predicted objects—to reveal the mechanics behind reward hacking and our risk-averse regularization.

Stage-I SFT: Data Enrichment Ablation We progressively enrich the SFT data to isolate each design choice. Starting from basic pass-only annotations (Base SFT: Prec. 79%, FP-Rej 5.2%), adding rejection labels and pseudo-CoT temporarily *degrades* Precision to 75% and raises FP-Rej to 8.5%, indicating short-term over-prediction. Multi-task scaffolding recovers Precision to 82% and reduces FP-Hal to 10.5%, while final augmentation further suppresses FP-Rej to 4.7%. This non-monotonic trajectory reveals that rejection awareness must be paired with multi-task regularization.

Table 1: Main results on Qwen3-VL-32B. External APIs/larger models are baselines; all values are percentages. Full metrics and 8B replication are in App. C.

Model / Stage	Recall	Prec.	FP-Rej	FP-Hal
Commercial APIs & Base Models				
GPT-5.2	34.17	57.28	6.10	36.62
GPT-5.4	27.45	56.16	4.58	39.26
Gemini 3 Flash	38.24	57.23	4.82	37.95
Gemini 3.1 Pro	35.29	61.61	3.67	34.72
Qwen3.5-Plus	31.79	63.41	5.87	30.73
Qwen3-VL-235B-A22B	30.81	62.50	5.11	32.39
Qwen3-VL-32B (Base)	31.09	49.55	7.59	42.86
Stage-I: SFT Ablation				
Base SFT (No Rej/CoT)	57.56	79.04	5.19	15.77
+ Reject & Pseudo-CoT	62.04	75.21	8.49	16.30
+ Multi-task Learning	65.41	81.93	7.54	10.53
Final SFT (+ Crop Aug)	64.29	83.45	4.73	11.82
Stage-II: DPO Refinement				
DPO w/o Aug	60.15	85.12	4.05	10.83
Final DPO (+ Crop Aug)	58.26	87.58	3.37	9.05
Stage-III: GRPO Alignment				
GRPO (Linear)	54.30	86.20	4.10	9.70
GRPO (Hard Bnd.)	12.45	96.80	1.10	2.10
GRPO (NARF)	45.52	92.29	2.48	5.23

Divergent Error Dynamics: Recorded vs. Novel Hallucinations An interesting pattern emerges from the SFT ablation: while FP-Hal drops sharply from 42.86% to 15.77% after basic SFT, FP-Rej exhibits a *non-monotonic* trajectory—initially decreasing to 5.19%, then rebounding to 8.49% when rejection labels and pseudo-CoT are introduced.

We hypothesize that SFT’s maximum-likelihood objective cannot explicitly push the model *away* from rejection patterns—it teaches what to output, not what to avoid. Consequently, the model may inadvertently overfit to the rejection descriptions themselves, learning spurious associations (e.g., “this scene contains a coffee shop”) without internalizing the critical distinction (“green-signed is wrong”). This limitation motivates the contrastive and reinforcement stages: DPO explicitly contrasts chosen vs. rejected outputs, while GRPO’s reward function penalizes any false positive, teaching the model to abstain when uncertain.

Preference Augmentation in DPO Removing cropped augmentation (“DPO w/o Aug”) hurts Precision (88%→85%) and increases FP-Hal, suggesting that synthesizing plausible negative responses for explicit contrast sharpens discriminative bound-

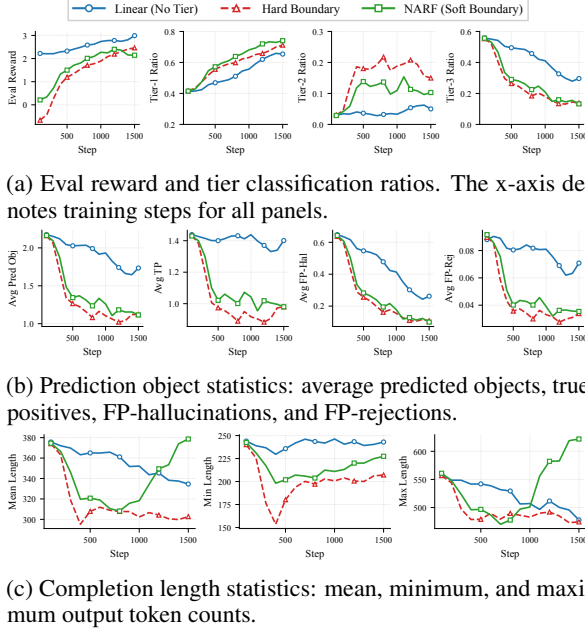


Figure 3: Training dynamics under three reward formulations (8B model; 32B exhibits consistent trends).

aries without additional labeling.

GRPO Reward Design: From Over-Prediction to Calibrated Risk Aversion SFT+DPO checkpoints exhibit an *over-prediction bias*: near-zero Tier-2 ratios, high predicted object counts (~ 2.2), and elevated FP rates (Avg FP-Hal ≈ 0.65). SFT and DPO teach the model *what to predict* but fail to teach *when not to predict*, motivating GRPO reward shaping. We compare three reward formulations.

Linear GRPO adds TP bonuses and FP penalties, enabling a “gambling” policy. As shown in Fig. 3(a-b), avg_pred_obj remains persistently high (~ 2.0 throughout training) and Tier-3 ratio declines slowly. The effective per-prediction accuracy is only $\sim 70\%$ (Avg TP ≈ 1.4 out of 2.0 predictions), yet this low hit rate is never penalized. Thus, it achieves high Recall (54%) but underperforms on Precision (86%) and fails to suppress FP-Rej (4.1%).

Hard tier boundaries ($r_{\text{ceil}} < r_2$) make any FP worse than silence, but trigger *silence collapse*: since the model cannot know *a priori* whether a prediction is TP or FP, predicting nothing becomes the safest strategy. As shown in Fig. 3, avg_pred_obj drops from ~ 2.2 to ~ 1.0 within 500 steps and Tier-2 ratio climbs steadily. Completion length gives the same signal: the minimum drops from ~ 245 to ~ 155 tokens as outputs become terse to minimize FP risk, and Recall collapses to 12%.

NARF’s soft boundary ($r_{\text{ceil}} > r_2$) lets high-TP-ratio outputs exceed silence. Like the hard variant, NARF reduces avg_pred_obj to ~ 1.0 and drives TP-to-prediction ratio toward near-100%, but induces a *calibrated* risk-averse policy rather than absolutist silence. A revealing signal is the completion length trajectory: NARF’s mean length initially drops (suppressing over-prediction), then *recovers* to near-baseline levels—a characteristic U-shaped curve. The increasing maximum length further confirms richer descriptions for high-confidence scenes. In contrast, linear’s length remains flat and hard’s length monotonically decreases (progressive retreat into silence). This recovery suggests that the model is no longer relying solely on silence to avoid penalties. NARF achieves 92% Precision while maintaining practical Recall (45%), with Tier-2 acting as a dynamic regularizer rather than an absolute behavioral sink.

4.4 Real-World Navigation Deployment

Setup and outcome definition. For the initial evaluation, we focused on the *departure scene*, where extracted landmarks are formatted into origin instructions: “Do you see the [description]? When you start, it should be on your [front/left/right], [nearby/in the distance].” The primary metric is the **median time from instruction playback completion to stable walking onset**, which measures how quickly and clearly users confirm their departure direction. A valid navigation event required at least 20 trajectory points sampled at one-second intervals, and no speed above 4 m/s. Starting after instruction playback completion, stable walking onset was the first observation in the earliest five-point window in which all speeds were at least 0.8 m/s and the maximum circular heading difference was at most 120 degrees.

Serving policy and coverage. We evaluated the system on our Amap pedestrian-navigation platform in Beijing, Shanghai, Hangzhou, and Shenzhen. Offline inference processed 500K image-covered road segments that satisfied the road-length, curvature, and data-quality filters. The 8B model ran first and the 32B model was invoked when 8B abstained; 411K segments (82.2%) yielded at least one valid landmark. During the experiment, approximately 40% of treatment-arm origin events played a landmark-augmented instruction. All other treatment events used the same direction-only fallback as control.

City	Users	Control	Landmark	Impr.
Beijing	~678K	7.46	5.73	23.2%
Shanghai	~553K	7.77	6.17	20.6%
Hangzhou	~315K	7.47	5.77	22.8%
Shenzhen	~471K	7.44	5.37	27.8%
All four	~2.02M	7.54	5.77	23.5%

Table 2: 30-day production A/B test. Control/Landmark are pooled event-level median times (s) from origin-instruction playback completion to stable walking onset. Trt. users are treatment-assigned unique users who initiated at least one navigation event. Impr. is the relative reduction from control.

Deployment protocol. Users were randomly assigned at the user level for the 30-day experiment (Apr. 20–May 20, 2026), remained in one arm, and did not cross over. At least 20% of users were assigned to treatment. Assignment did not target city, time, route length, or other business variables. The approximately 2.02M users reported below are treatment-assigned users who initiated at least one navigation event in the four cities.

Results. The pooled control median was 7.54 s (95% CI: 7.43–7.61) and the treatment median was 5.77 s (5.66–5.83), an absolute reduction of 1.77 s (1.64–1.87) and a relative reduction of 23.5% (22.7–24.3%). Confidence intervals use a user-cluster percentile bootstrap: users were resampled with replacement within each arm, all valid events associated with a sampled user were retained, and repeatedly sampled users contributed their event sets repeatedly. City-level relative reductions ranged from 20.6% to 27.8% (Table 2). These results measure departure-confirmation latency under assignment to a landmark-enabled policy.

5 Conclusion

We present a coordinate-free MLLM framework for navigation landmark extraction. A progressive alignment pipeline—multi-task SFT, contrastive DPO, and NARF-driven GRPO—addresses hallucination and unsuitable landmark selection without bounding-box supervision. By coupling soft-hard bipartite matching with a three-tier risk-sensitive reward, our method induces calibrated abstention, sharply reduces false positives, and improves both offline precision and real-world wayfinding latency.

Limitations

Dependence on the semantic scorer. Both the bipartite matcher and the NARF reward use a frozen reranker (Wang et al., 2025a) for semantic similarity. The reported offline metrics may therefore inherit its errors or biases and should not be interpreted as scorer-independent ground truth.

Hyperparameter portability of NARF. NARF’s three-tier reward involves multiple hyperparameters ($r_1, r_2, r_3, r_{\text{ceil}}, \alpha$, tier boundaries, and attribute-level penalties), tuned manually to encode pedestrian-navigation risk asymmetry. Porting NARF to a different safety-critical domain (e.g., medical triage or autonomous-driving perception) would require re-deriving these constants from domain-specific cost models, which we leave to future work.

Outcome scope. The A/B test measures departure-confirmation latency at the first navigation step, rather than complete-route task success or safety outcomes. It therefore supports the practical value of landmark-augmented origin guidance, not route-level effectiveness.

Ethical Considerations

Data provenance, privacy, and retention. The dashcam sequences were legally purchased from licensed third-party providers, and their acquisition and use passed internal legal and security review. Before delivery, detected faces were blurred, detected license plates were mosaiced, and textual overlays were masked. We received only de-identified images and task-relevant metadata, without personal identity fields. Both are access-controlled and retained for at most one year; the procurement agreement permits model training but prohibits public redistribution of the raw imagery.

User data and governance. The production A/B test was conducted as a product evaluation under the existing User Agreement and Privacy Policy with internal approval. Stable assignment used pseudonymized identifiers; identity and profile attributes were excluded from analysis, only aggregate anonymized metrics are reported, and analysis records were retained for one month. During the experiment, users could raise objections through customer service, which could change their experiment assignment.

Potential harm and deployment safeguards.

An incorrect landmark or left–right attribute could mislead pedestrians. Deployment therefore combines precision-oriented filtering with the existing direction-only fallback, incorrect-instruction reporting, complaint and safety-event monitoring, and immediate removal of reported inaccurate landmarks. Before each city-batch release, experts audit 100–500 cases; release requires a case-level pass rate above 95%, where a case passes only if it contains no incorrect landmark prediction. Since experiment launch, at most 10 of more than 10,000 user-feedback entries have concerned landmark accuracy, and no verified safety incident attributable to landmark instructions has been observed.

Acknowledgments

We thank Qijing Guo, Xianlong Pan, and Xiang Zhang from the engineering team for their support in developing the online serving pipeline, including road-ID-based key–value retrieval and template-based instruction assembly. We also thank navigation experts Gebin Mo and Xin Wen from the product team for formalizing the landmark-selection criteria, iteratively refining the annotations, and designing the instruction templates.

References

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267. Association for Computational Linguistics.
- Negar Alinaghi, Ioannis Giannopoulos, Markus Kattenbeck, and Martin Raubal. 2025. [Decoding wayfinding: Analyzing wayfinding processes in the outdoor environment](#). *International Journal of Geographical Information Science*, 39(8):1770–1800.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.
- Lang Cao. 2024. [Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3628–3646.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. [End-to-end object detection with transformers](#). In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.
- Howard Chen, Alane Suhr, Dipendra Misra, Noah Snaveley, and Yoav Artzi. 2019. [TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12538–12547.
- Yibo Cui, Liang Xie, Yu Zhao, Jiawei Sun, and Erwei Yin. 2026. [Generating vision-language navigation instructions incorporated fine-grained alignment annotations](#). *Information Fusion*, 130:104107.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. 2023. [Gemini: A family of highly capable multimodal models](#). *arXiv preprint arXiv:2312.11805*.
- Antonia Golab, Markus Kattenbeck, Georgios Sarlas, and Ioannis Giannopoulos. 2022. [It’s also about timing! when do pedestrians want to receive navigation instructions](#). *Spatial Cognition & Computation*, 22(1–2):74–106.
- Harlan Hile, Radek Grzeszczuk, Alan Liu, Ramakrishna Vedantham, Jana Košćeka, and Gaetano Borriello. 2009. [Landmark-based pedestrian navigation with enhanced spatial reasoning](#). In *Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 59–76. Springer.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. [OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Gaurav Jain, Leah Findlater, and Cole Gleason. 2026. [SceneScout: Towards AI-driven access to street level imagery for blind users](#). In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–22. ACM.
- Yuhua Jiang, Jiawei Huang, Yufeng Yuan, Xin Mao, Yu Yue, Qianchuan Zhao, and Lin Yan. 2026. [Risk-sensitive reinforcement learning for alleviating exploration dilemmas in large language models](#). In *International Conference on Learning Representations*.

- Sai Srinivas Kancheti, Aditya Kanade, Rohit Sinha, Vineeth N. Balasubramanian, and Tanuja Ganu. 2026. Faithful GRPO: Improving visual spatial reasoning in multimodal language models via constrained policy optimization. *arXiv preprint arXiv:2604.08476*.
- Xuying Li, Zhuo Li, Yuji Kosuga, and Victor Bian. 2025. Optimizing safe and aligned language generation: A multi-objective GRPO approach. *arXiv preprint arXiv:2503.21819*.
- Yangfu Li, Yuning Gong, Hongjian Zhan, Teng Li, Yuanhuiyi Lyu, Tianyi Chen, Qi Liu, Ziyuan Huang, Zhihang Zhong, Dandan Zheng, and Yue Lu. 2026. Perceptual flow network for visually grounded reasoning. *arXiv preprint arXiv:2605.02730*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics.
- Xu Liang. 2025. Group relative policy optimization for image captioning. *arXiv preprint arXiv:2503.01333*.
- Bingqian Lin, Yunshuang Nie, Ziming Wei, Yi Zhu, Hang Xu, Shikui Ma, Jianzhuang Liu, and Xiaodan Liang. 2024. [Correctable landmark discovery via large models for vision-language navigation](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8534–8548.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. [Understanding R1-Zero-like training: A critical perspective](#). In *Proceedings of the Second Conference on Language Modeling*.
- Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng, and Yujiu Yang. 2025. [Unlocking multimodal mathematical reasoning via process reward model](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 49851–49899. Curran Associates, Inc.
- Kevin Lynch. 1960. *The Image of the City*. The MIT Press.
- Alexandra Millonig and Katja Schechtner. 2007. [Developing landmark-based pedestrian-navigation systems](#). *IEEE Transactions on Intelligent Transportation Systems*, 8(1):43–49.
- Chieko Nishimura, Shuhei Kurita, and Yohei Seki. 2024. [Text360Nav: 360-degree image captioning dataset for urban pedestrians navigation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15783–15788. ELRA and ICCL.
- Christina Ohm, Stefan Bienk, Markus Kattenbeck, Bernd Ludwig, and Manuel Müller. 2016. [Towards interfaces of mobile pedestrian navigation systems adapted to the user’s orientation skills](#). *Pervasive and Mobile Computing*, 26:121–134.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Xuankun Rong, Wenke Huang, Tingfeng Wang, Daiguo Zhou, Bo Du, and Mang Ye. 2026. [SafeGRPO: Self-rewarded multimodal safety alignment via rule-governed policy optimization](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7901–7911.
- Florian Röser. 2015. *The Cognitive Observer-Based Landmark-Preference Model: What Is the Ideal Landmark Position at an Intersection?* Ph.D. thesis, Justus-Liebig-Universität Gießen.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Huilin Tian, Jingke Meng, Wei-Shi Zheng, Yuan-Ming Li, Junkai Yan, and Yunong Zhang. 2024. [Loc4Plan: Locating before planning for outdoor vision and language navigation](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4073–4081. ACM.
- Feng Wang, Yuqing Li, and Han Xiao. 2025a. jina-reranker-v3: Last but not late interaction for listwise document reranking. *arXiv preprint arXiv:2509.25085*.
- Qu Wang, Haiyong Luo, Jianquan Wang, Lei Sun, Zhangchao Ma, Chaoyi Zhang, Meixia Fu, and Fang Zhao. 2022a. [Recent advances in pedestrian navigation activity recognition: A review](#). *IEEE Sensors Journal*, 22(8):7499–7518.

- Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldrige, and Peter Anderson. 2022b. [Less is more: Generating grounded navigation instructions from landmarks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15428–15438.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025b. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. [Know your limits: A survey of abstention in large language models](#). *Transactions of the Association for Computational Linguistics*, 13:529–556.
- Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. 2025. [FLAME: learning to navigate with multimodal LLM in urban environments](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 9005–9013. AAAI Press.
- Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. 2025. [Dance-GRPO: Unleashing GRPO on visual generation](#). *arXiv preprint arXiv:2505.07818*.
- Yiming Yang, Hao Fu, Fanxiang Zeng, Xikai Yang, Yue Liu, and Ning Guo. 2026. [Revolutionizing turn-by-turn navigation with cloud-edge deep learning](#). *IEEE Transactions on Intelligent Transportation Systems*, 27(7):7882–7892.
- Demet Yesiltepe, Ruth Conroy Dalton, and Ayse Ozbil Torun. 2021. [Landmarks in wayfinding: A review of the existing literature](#). *Cognitive Processing*, 22(3):369–410.
- Huahui Yi, Kun Wang, Qiankun Li, Miao Yu, Liang Lin, Gongli Xi, Hao Wu, Xuming Hu, Kang Li, and Yang Liu. 2025. [Safer-vlm: Toward safety-aware fine-grained reasoning in multimodal models](#). *arXiv preprint arXiv:2510.06871*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [DAPO: An open-source LLM reinforcement learning system at scale](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244. Curran Associates, Inc.
- Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2023. [DINO: DETR with improved denoising anchor boxes for end-to-end object detection](#). In *International Conference on Learning Representations (ICLR)*.
- Lijun Zhang, Lin Li, Yajie Qi, Huizhong Song, Yaodong Yang, Jun Wang, and Wei Wei. 2025a. [Risk-aware direct preference optimization under nested risk measure](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 73591–73625. Curran Associates, Inc.
- Yifan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Tingting Gao, Zhang Zhang, Fan Yang, Di Zhang, Liang Wang, and Rong Jin. 2025b. [MM-RLHF: The next step forward in multimodal LLM alignment](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76625–76654. PMLR.
- Yi Zhao, Siqi Wang, and Jing Li. 2026. [Laf-grpo: In-situ navigation instruction generation for the visually impaired via GRPO with llm-as-follower reward](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 34994–35002. AAAI Press.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. 2025. [Group sequence policy optimization](#). *arXiv preprint arXiv:2507.18071*.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. [Analyzing and mitigating object hallucination in large vision-language models](#). In *International Conference on Learning Representations*.

A Implementation Details

Training Hyperparameters All three stages apply LoRA (Hu et al., 2022) (rank = 16, $\alpha = 32$) to all linear layers, keeping the vision encoder and aligner frozen.

- **Stage I (SFT):** $\sim 12.5k$ samples (five task types, each with original and cropped variants), 2 epochs, learning rate 8×10^{-5} .
- **Stage II (DPO):** $\sim 6.8k$ contrastive preference pairs (seven perturbation types $\times$ two variants), 2 epochs, learning rate 2×10^{-6} , $\beta = 0.8$, $rpo_alpha = 1.0$, and sigmoid DPO loss.
- **Stage III (GRPO):** $\sim 3.5k$ prompts, 1.5k steps, learning rate 2.5×10^{-6} , $G = 16$ rollouts per prompt, global batch size 64.

The 8B model follows the same hyperparameters; its stage-wise trends are consistent with the 32B findings (Table 6).

Evaluation Inference Settings All evaluations use `max_tokens=4000`. Except for GRPO checkpoints, all systems use temperature 0.0 with the default repetition penalty. GRPO checkpoints use temperature 0.4 and repetition penalty 1.05.

GRPO Rollout and Reward Parameters During GRPO training, each prompt is sampled with $G=16$ rollouts, temperature 1.0, repetition penalty 1.1, maximum completion length 1000, $\beta=0.0$, one optimization iteration per generation group, and reward scaling disabled. The implementation registers a single weighted reward, NARF, together with zero-weight monitor rewards that log Tier-1/2/3 ratios, average prediction count, JSON failure rate, average TP, FP-Rej, and FP-Hal. Table 3 summarizes the concrete reward and matching constants used by the main NARF experiments.

B Semantic-Scorer Sensitivity Analysis

The original evaluator and GRPO reward use a frozen Jina reranker for semantic similarity. To test whether the reported precision-first ordering depends on this scorer, we reevaluated fixed model predictions with GTE-rerank-v2, which was not used in training. Hard gates and structured attribute scores were unchanged.

Table 4 applies the original Jina thresholds ($T_1/T_2 = 0.2/0.4$) without retuning. GRPO-NARF reaches 87.62% Precision and 50.56% Recall and retains the highest Precision among the

Table 3: Key evaluation and GRPO reward constants used in the implementation.

Component	Setting
Matching thresholds	Match floor $\tau_1 = 0.2$; high-quality threshold $\tau_2 = 0.4$.
Matching weights	type / range / dir. / dist. / semantic / color-soft = 0.2/0.2/0.1/0.1/0.4/0.1.
Format and consistency	JSON parse failure -10.0 ; parseable JSON base $+0.2$; missing field, invalid enum, or invalid range -0.1 each; CoT keyword reward $+0.1$ capped at $+0.2$; count consistency reward $+0.2$.
NARF tier bases	$r_1 = 2.0$, $r_2 = -2.0$, $r_3 = -4.0$, $r_{tp} = 1.0$, $r_{all} = 3.0$, $\alpha = 0.5$, $r_{ceil} = 0.0$.
Quality scaling	Semantic bonus multiplier $\lambda_h = 1.0$ above τ_2 and $\lambda_l = 0.5$ below τ_2 .
Attribute/error penalties	color conflict -1.0 ; left/right confusion -0.8 ; front/side confusion -0.5 ; distance mismatch -0.2 ; range penalty scale -0.5 ; FP-Rej -2.0 ; FP-Hal -1.0 ; FN 0.0.

models included in this study. Absolute values change because the two rerankers have different score distributions.

For completeness, Table 5 reports a separately identified post-hoc GTE calibration ($T_1/T_2 = 0.175/0.3$). This setting was selected after observing GTE score behavior and is not a preregistered or training-time configuration.

Offline uncertainty. Confidence intervals for the 240-sequence benchmark use 20,000 sequence-level paired-bootstrap replicates. Each replicate samples 240 sequences with replacement and applies the same sampled indices to all compared models and threshold settings. Final SFT reaches 64.29% Recall (95% CI: 61.09–67.63) and 83.45% Precision (80.24–86.58); GRPO-NARF reaches 45.52% Recall (42.14–48.91) and 92.29% Precision (88.99–95.30). These are marginal metric intervals. We do not claim paired significance for stage-to-stage differences because paired-difference intervals or tests are not reported.

C Cross-Scale Replication and Extended Ablation Analysis

This section reports the full cross-scale results for all stages and ablations (Table 6) and provides extended analysis of the GRPO training dynamics that the main text (Sec. 4.3) summarizes.

C.1 Cross-Scale Consistency

The 8B rows in Table 6 show that the same progressive trends hold at a smaller model scale. Base SFT achieves 72.83% Precision with 21.89% FP-Hal; adding rejection labels and pseudo-CoT again triggers the characteristic Precision dip (68.07%) and FP-Rej spike (6.59%), mirroring the 32B pat-

Model	Reported Jina R / P	GTE R / P	$\Delta R / \Delta P$
GPT-5.2	34.17 / 57.28	32.49 / 54.46	-1.68 / -2.82
GPT-5.4	27.45 / 56.16	26.47 / 54.15	-0.98 / -2.01
Gemini 3 Flash	38.24 / 57.23	35.57 / 53.25	-2.67 / -3.98
Gemini 3.1 Pro	35.29 / 61.61	33.33 / 58.19	-1.96 / -3.42
Qwen3.5-Plus	31.79 / 63.41	30.39 / 60.61	-1.40 / -2.80
Qwen3-VL-235B-A22B	30.81 / 62.50	28.85 / 58.52	-1.96 / -3.98
Qwen3-VL Base 8B	14.01 / 45.66	13.73 / 44.75	-0.28 / -0.91
Qwen3-VL Base 32B	31.09 / 49.55	29.13 / 46.43	-1.96 / -3.12
GRPO-NARF 32B	45.52 / 92.29	50.56 / 87.62	+5.04 / -4.67

Table 4: GTE-rerank-v2 sensitivity analysis using the original Jina thresholds ($T_1/T_2 = 0.2/0.4$) without retuning. Values are Recall / Precision (%); deltas are percentage-point changes from the reported Jina-based results.

Model	Reported Jina R / P	GTE R / P	$\Delta R / \Delta P$
GPT-5.2	34.17 / 57.28	37.11 / 62.21	+2.94 / +4.93
GPT-5.4	27.45 / 56.16	30.39 / 62.18	+2.94 / +6.02
Gemini 3 Flash	38.24 / 57.23	42.02 / 62.89	+3.78 / +5.66
Gemini 3.1 Pro	35.29 / 61.61	37.39 / 65.28	+2.10 / +3.67
Qwen3.5-Plus	31.79 / 63.41	33.33 / 66.48	+1.54 / +3.07
Qwen3-VL-235B-A22B	30.81 / 62.50	31.51 / 63.92	+0.70 / +1.42
Qwen3-VL Base 8B	14.01 / 45.66	15.13 / 49.32	+1.12 / +3.66
Qwen3-VL Base 32B	31.09 / 49.55	34.03 / 54.24	+2.94 / +4.69
GRPO-NARF 32B	45.52 / 92.29	52.38 / 90.78	+6.86 / -1.51

Table 5: GTE-rerank-v2 sensitivity analysis with separately selected post-hoc thresholds ($T_1/T_2 = 0.175/0.3$). Values and deltas follow Table 4.

tern. The full SFT recovers Precision to 76.13% and suppresses FP-Rej to 3.98%. DPO further lifts Precision to 79.38% and reduces FP-Rej to 3.70%, consistent with the 32B trajectory. NARF then pushes 8B Precision to 92.75% while suppressing FP-Rej to 1.09%, closely mirroring the 32B results.

Notably, the 8B GRPO ablations reproduce the same phenomena observed in 32B: the linear reward baseline achieves higher Recall (43.70%) but lower Precision (88.89%), while the hard-boundary variant triggers silence collapse (Recall drops to 27.87%). Despite the smaller model capacity, the relative improvement patterns, the non-monotonic FP-Rej behavior, and the GRPO reward dynamics remain consistent, suggesting that our data and progressive pipeline are robust across model scales.

C.2 Scene-Complexity Adaptive Behavior

Beyond the aggregate training dynamics analyzed in the main text (Sec. 4.3), per-batch rollout statistics reveal that the NARF-trained model dynamically modulates its prediction count according to scene complexity. In high-ambiguity scenes (e.g., dense urban areas with many similar buildings), where the expected per-prediction accuracy is low, the critical count (Eq. 6) is unlikely to be met, so the model reduces predictions or abstains. In low-ambiguity scenes (e.g., a distinctive sculpture on

a suburban road), the model predicts confidently and produces longer descriptions. This adaptive behavior is induced by the reward landscape and appears in held-out scenarios, in contrast to heuristic post-processing rules that require manual threshold tuning per scene type.

C.3 Cascaded Deployment Implication

The cross-scale consistency observed above carries a direct deployment consequence: the 8B NARF model achieves 92.75% precision (Table 6), albeit with lower recall than 32B. This supports a practical cascaded strategy: run 8B first and invoke 32B only when 8B abstains, reducing computational cost while maintaining high prediction reliability, as in our 411K-segment deployment (Sec. 4.4).

D Case-Based Walkthrough of the Proposed Method

This section uses three held-out validation case panels to make the method concrete without adding material to the main paper. The panels show (i) coordinate-free bipartite matching, (ii) recomputed NARF content rewards on real validation outputs, and (iii) prediction changes after progressive fine-tuning. All case panels are rendered from stored validation predictions and evaluator outputs.

Table 6: Full quantitative results across baselines, progressive alignment stages, and ablation variants. Each experiment reports 8B and 32B results in adjacent sub-rows when applicable, extending the main-text table with JSON validity, direction accuracy, and distance accuracy.

Model / Stage	Scale	Recall	Prec.	FP-Rej	FP-Hal	JSON	Dir.	Dist.
Commercial APIs & Base Models								
GPT-5.2	–	34.17	57.28	6.10	36.62	99.17	80.74	73.77
GPT-5.4	–	27.45	56.16	4.58	39.26	99.58	88.27	77.04
Gemini 3 Flash	–	38.24	57.23	4.82	37.95	96.25	80.59	81.32
Gemini 3.1 Pro	–	35.29	61.61	3.67	34.72	99.58	84.13	81.35
Qwen3.5-Plus	–	31.79	63.41	5.87	30.73	100.0	83.70	75.33
Qwen3-VL-235B-A22B	–	30.81	62.50	5.11	32.39	100.0	79.09	75.00
Qwen3-VL (Base)	8B	14.01	45.66	6.85	47.49	44.17	77.00	65.00
	32B	31.09	49.55	7.59	42.86	99.17	78.48	74.89
Stage-I: SFT Ablation								
Base SFT (No Rej/CoT)	8B	54.06	72.83	5.28	21.89	100.0	95.85	93.01
	32B	57.56	79.04	5.19	15.77	100.0	93.43	93.67
+ Reject & Pseudo-CoT	8B	56.44	68.07	6.59	25.34	97.92	94.04	92.31
	32B	62.04	75.21	8.49	16.30	99.58	93.00	92.78
+ Multi-task Learning	8B	58.82	75.13	5.90	18.96	98.33	93.81	94.05
	32B	65.41	81.93	7.54	10.53	99.17	94.00	93.36
Final SFT (+ Crop Aug)	8B	58.96	76.13	3.98	19.89	97.08	92.87	95.01
	32B	64.29	83.45	4.73	11.82	99.17	93.68	93.25
Stage-II: DPO Refinement								
DPO w/o Aug	8B	57.42	77.95	4.37	17.68	95.42	94.39	94.39
	32B	60.15	85.12	4.05	10.83	99.58	93.85	92.50
Final DPO (+ Crop Aug)	8B	57.14	79.38	3.70	16.93	94.17	94.61	94.36
	32B	58.26	87.58	3.37	9.05	100.0	94.23	92.31
Stage-III: GRPO Alignment								
GRPO (Linear)	8B	43.70	88.89	3.13	7.98	92.50	95.51	94.55
	32B	54.30	86.20	4.10	9.70	99.17	93.50	92.80
GRPO (Hard Bnd.)	8B	27.87	93.87	0.47	5.66	88.75	95.48	95.48
	32B	12.45	96.80	1.10	2.10	100.0	94.50	94.10
GRPO (NARF)	8B	35.85	92.75	1.09	6.16	93.33	96.48	95.70
	32B	45.52	92.29	2.48	5.23	100.0	94.03	93.43

D.1 Reward Audit on a Real Case

Figure 4 replays the reward calculation for a base-model output on sample_00033 from the validation record. The panel shows the parsed JSON predictions, response-level reward components, the full GT-by-prediction score matrix with hard-gate reasons and attribute scores, and the Hungarian assignment that yields TP, FP-Rej, and FN labels.

D.2 NARF Reward Calculation Across Tiers

Figure 5 and Table 7 show selected validation model outputs from real cases with rewards re-computed from the reward formula used in our experiments. Because rewards attach to a particular model output rather than to a sample alone, Table 7 also lists the source model for each case.

D.3 Behavioral Changes Across Alignment Stages

Figure 6 compares the base, SFT, and final DPO+GRPO models on four real cases. Each row displays the original frame and the stored structured outputs with evaluator labels.

D.4 Aggregate Case Statistics

The case-level phenomena are also visible across all 240 validation samples for the systems with stored reranker-matrix dumps. Figure 7 shows that the final model has the largest Tier-1 mass and the smallest Tier-3 mass among these compared systems. The reward CDF in Figure 8 shows the same safety pattern from another view: the final model has the smallest left tail, meaning fewer severe low-reward outputs.

Single-case reward audit from validation record

sample_00033 | Base prediction | response parsing, hard gates, Hungarian assignment, and NARF content reward

sample_00033 | Base



Source: stored response_text + reranker matrix dump

Parsed JSON predictions

P0 蓝色车道路牌 / blue road sign

type=路牌 dir=front dist=far range=f1

P1 黄色岗亭 / yellow booth

type=亭子 dir=right dist=near range=f1

Reward decomposition

+0.20

Got

+0.20

JSON

+0.20

Consistency

-5.22

Content

-4.62

Full response reward

Tier 3 = $-4.00 + 1 * (-2.0) + 0 * (-1.0) + 0.5 * 1.56 = -5.22$

Counts: TP=1, FP-Rej=1, FP-Hal=0, FN=1 | FN is reported but P_FN=0 in the content reward

Pairwise score matrix and assignment path

Score = $0.2 * \text{type} + 0.2 * \text{range} + 0.1 * \text{dir} + 0.1 * \text{dist} + 0.1 * \text{color} + 0.4 * \text{semantic}$; hard gates set blocked cells to -1.

GT candidates	Pred 0 蓝色车道路牌 / blue road sign type=路牌 dir=front dist=far range=f1	Pred 1 黄色岗亭 / yellow booth type=亭子 dir=right dist=near range=f1
GT 0 PASS 蓝色车道路牌 / blue road sign type=路牌 dir=front dist=far range=f1	S=0.923 TP assignment type=1.0 range=1.0 dir=1.0 dist=1.0 color=1.0 semantic=0.559	blocked blocked: type gate GT type=路牌 pred type=亭子 type=0.0 range=0.0 dir=0.0 dist=0.0 color=0.0 semantic=-0.095
GT 1 PASS 弧形玻璃幕墙大楼 / curved glass building type=建筑 dir=right dist=near range=f1	blocked blocked: type gate GT type=建筑 pred type=路牌 type=0.0 range=0.0 dir=0.0 dist=0.0 color=0.0 semantic=-0.084	blocked blocked: semantic $-0.05 < 0.20$ type=1.0 range=1.0 dir=1.0 dist=1.0 color=0.0 semantic=-0.054
GT 2 REJECT 黄色垃圾分类亭 / yellow trash kiosk type=亭子 dir=right dist=near range=f1	blocked blocked: type gate GT type=亭子 pred type=路牌 type=0.0 range=0.0 dir=0.0 dist=0.0 color=0.0 semantic=-0.093	S=0.833 FP-Rej assignment type=1.0 range=1.0 dir=0.5 dist=1.0 color=1.0 semantic=0.457

Assignment result: Pred0 -> GT0 = TP; Pred1 -> GT2 = FP-Rej; GT1 remains unmatched = FN.

Figure 4: Single-case reward audit from the validation record. The matrix exposes the hard gates, per-attribute scores, semantic reranker scores, and final Hungarian assignment used before applying the tiered NARF content reward formula.

Real validation cases under the NARF content reward

Rewards recomputed from the reward formula used in our experiments.

sample_00150 | Base



Tier 1 partial clean

Reward +6.62

TP=3, FP-Rej=0, FP-Hal=0, FN=1

TP 蓝色车道路牌 / blue road sign

type=路牌 dir=front dist=near range=f1-f2

TP 蓝色围挡 / blue barrier

type=围挡 dir=left dist=near range=f1-f2

TP 红色围墙 / landmark

type=围墙 dir=right dist=near range=f1-f2

sample_00171 | Base



Tier 1 complete clean

Reward +8.56

TP=3, FP-Rej=0, FP-Hal=0, FN=0

TP 蓝色悬挂车道路牌 / landmark

type=路牌 dir=front dist=far range=f1-f3

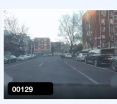
TP 白色带拱形门洞的建筑 / landmark

type=建筑 dir=right dist=near range=f1-f3

TP 蓝色路牌花朝街 / landmark

type=路牌 dir=right dist=near range=f2-f3

sample_00129 | GRPO



Tier 2 abstention

Reward -2.00

TP=0, FP-Rej=0, FP-Hal=0, FN=1

output: []

sample_00033 | Base



Tier 3 rule violation

Reward -5.22

TP=1, FP-Rej=1, FP-Hal=0, FN=1

TP 蓝色车道路牌 / blue road sign

type=路牌 dir=front dist=far range=f1

FP-R 黄色岗亭 / yellow booth

type=亭子 dir=right dist=near range=f1

sample_00001 | Base



Tier 3 hallucination

Reward -5.00

TP=0, FP-Rej=0, FP-Hal=1, FN=2

FP-H 红色中国结装饰的白色路灯杆 / landmark

type=路灯杆 dir=left dist=near range=f2-f4

sample_00009 | Base



Tier 3 rejected building

Reward -6.00

TP=0, FP-Rej=1, FP-Hal=0, FN=2

FP-R 白色外墙的多层居民楼 / landmark

type=建筑 dir=right dist=far range=f1-f2

Figure 5: Real-case NARF content rewards and structured outputs from selected validation predictions, recomputed from the reward formula used in our experiments.

Role	Source	Case	Outcome Pattern	Reward
Tier-1 partial clean	Base	sample_00150	3 TP, 0 FP, 1 FN	+6.62
Tier-1 complete clean	Base	sample_00171	3 TP, 0 FP, 0 FN	+8.56
Tier-2 abstention	DPO+GRPO	sample_00129	0 TP, 0 FP, 1 FN	-2.00
Tier-3 rule violation	Base	sample_00033	1 TP, 1 FP-Rej, 1 FN	-5.22
Tier-3 pure hallucination	Base	sample_00001	0 TP, 1 FP-Hal, 2 FN	-5.00
Tier-3 pure rule violation	Base	sample_00009	0 TP, 1 FP-Rej, 2 FN	-6.00

Table 7: NARF content reward ladder across selected validation predictions. FP-Rej denotes a prediction matched to an expert-rejected object; FP-Hal denotes an unmatched hallucination.

Stage-wise prediction changes on real validation cases

Stored Base / SFT / DPO+GRPO outputs with evaluator labels.

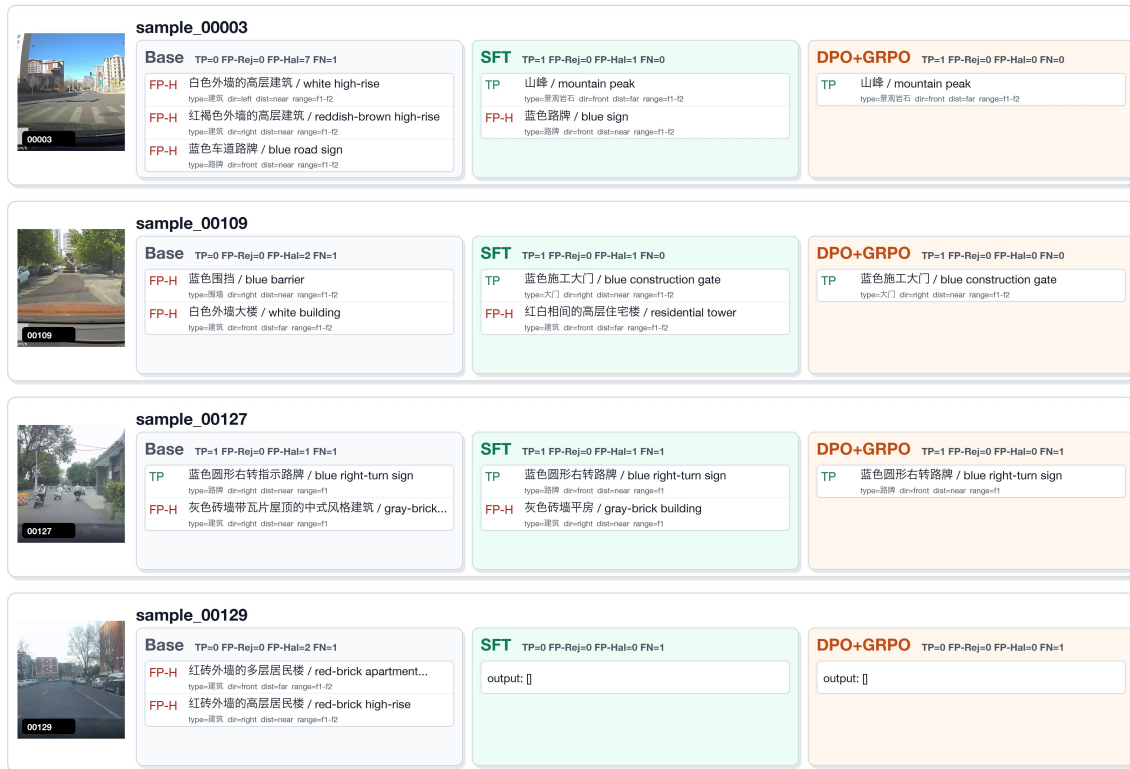


Figure 6: Stage-wise prediction changes on real validation cases.

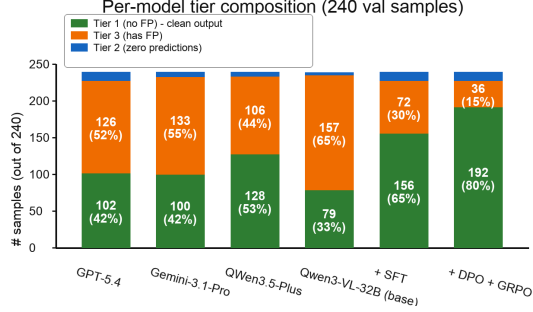


Figure 7: Tier composition over 240 validation samples for the systems with stored reranker-matrix dumps. The final DPO+GRPO model substantially increases no-FP cases relative to the base model and reduces Tier-3 cases relative to SFT.

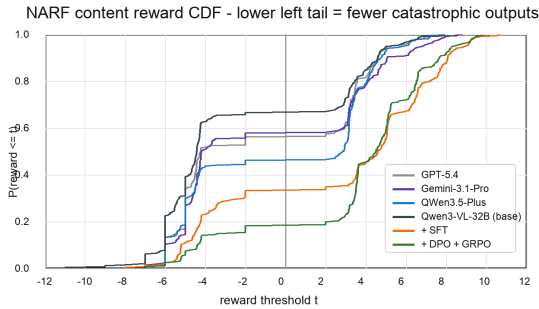


Figure 8: CDF of per-sample rewards. Lower mass in the negative-reward region indicates fewer severe low-reward outputs; the final model has the smallest left tail.

E Coordinate-Free Bipartite Matching: Full Equations

This section gives the full formal definitions of the hard-gated cost matrix and the soft attribute scoring functions that the main text (Sec. 3.3) summarizes in prose.

E.1 Hard-Gated Cost Matrix

For each prediction–GT pair $(\hat{y}_i, y_j)$, we evaluate a cascade of hard gates. If any gate fires, the pair is unconditionally blocked ($S_{ij} = -1$) regardless of other attributes:

$$G(\hat{y}_i, y_j) = \begin{cases} \perp, & c_i \neq c_j \\ \perp, & \{o_i, o_j\} = \{\text{left}, \text{right}\} \\ \perp, & [s_i, e_i] \cap [s_j, e_j] = \emptyset \\ \perp, & |\mathcal{K}_i| = |\mathcal{K}_j| = 1 \wedge \mathcal{K}_i \neq \mathcal{K}_j \\ \perp, & f_{\text{sem}}(d_i, d_j) < \tau_1 \\ \top, & \text{otherwise} \end{cases} \quad (7)$$

The five blocking rows correspond to type mismatch, lateral conflict, range disjointness, color conflict, and semantic floor. Here, $\mathcal{K}(\cdot)$ extracts mentioned colors from a description through regex matching, and f_{sem} is the semantic similarity between two descriptions computed by a reranker model (Wang et al., 2025a); the color-conflict gate fires only when both sides specify a single, conflicting color (e.g. “red-signed shop” vs. “blue-signed shop”). The threshold τ_1 acts as a floor: pairs below it are deemed entirely unrelated.

E.2 Range and Direction Soft Scores

Range Overlap. ϕ_{rng} adopts a tiered design that rewards conservative predictions (whose range is subsumed by the ground truth) more than merely overlapping ones:

$$\phi_{\text{rng}} = \begin{cases} 1.0, & [s_i, e_i] = [s_j, e_j] \\ 0.3 + 0.7 \rho, & [s_j, e_j] \supseteq [s_i, e_i] \\ 0.1 + 0.2 \rho, & [s_i, e_i] \cap [s_j, e_j] \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where $\rho = \text{IoU}([s_i, e_i], [s_j, e_j])$. The intuition is that a prediction visible in frames 2–3 when the GT spans frames 1–4 still provides *conservative* guidance, whereas a prediction in frames 1–4 when the GT is only in frames 2–3 may refer to a different object.

Orientation Matching. ϕ_{dir} encodes the asymmetric severity of directional errors in pedestrian navigation:

$$\phi_{\text{dir}} = \begin{cases} 1.0, & o_i = o_j \\ -0.2, & \{o_i, o_j\} = \{\text{left}, \text{right}\} \\ 0.5, & \text{front} \in \{o_i, o_j\} \end{cases} \quad (9)$$

A left $\leftrightarrow$ right confusion may direct the user to the opposite side of the road—the most severe directional error—hence the negative score; front $\leftrightarrow$ side confusions are less severe and receive a partial score.

F NARF: Full Functional Forms

This section provides the complete functional forms of the per-prediction quality score, the attribute penalty set, the Tier-3 error decomposition, and the total reward composition, which the main text (Sec. 3.4) summarizes in prose. Concurrent work on visually grounded reasoning, PFlowNet (Li et al., 2026), similarly combines multi-dimensional rewards with vicinal geometric shaping to promote reasoning-oriented perception while preserving visual reliability.

F.1 Per-Prediction Quality Score

For each TP pair $(\hat{y}_k, y_{\sigma^*(k)})$, the quality score Q_k captures both semantic fidelity and attribute correctness:

$$Q_k = r_{\text{tp}} + f_{\text{sem}}^k \cdot \begin{cases} \lambda_h, & f_{\text{sem}}^k \geq \tau_2 \\ \lambda_l, & f_{\text{sem}}^k < \tau_2 \end{cases} + \sum_{a \in \mathcal{P}} p_a^k \quad (10)$$

where r_{tp} is the base TP reward, f_{sem}^k is the reranker similarity score from the matching stage, $\tau_2 > \tau_1$ is a high-quality threshold, and $\lambda_h > \lambda_l > 0$ provide differentiated scaling: predictions with $f_{\text{sem}} \geq \tau_2$ are rewarded more generously, incentivizing precise descriptions.

F.2 Attribute Penalty Set

The attribute penalty set $\mathcal{P} = \{p_{\text{clr}}, p_{\text{dir}}, p_{\text{dst}}, p_{\text{rng}}\}$ applies fine-grained deductions:

$$\begin{aligned} p_{\text{dir}}^k &= \begin{cases} 0, & o_k = o_k^* \\ p_{\text{side}}, & \{o_k, o_k^*\} = \{\text{left}, \text{right}\} \\ p_{\text{mild}}, & \text{otherwise} \end{cases} \\ p_{\text{rng}}^k &= (1 - \phi_{\text{rng}}^k) \cdot p_{\text{r_scale}} \\ p_{\text{clr}}^k &= \begin{cases} p_{\text{c_severe}}, & \text{color conflict} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (11)$$

where $p_{\text{side}} < p_{\text{mild}} < 0$ reflects the navigation-specific severity hierarchy: a left $\leftrightarrow$ right error (p_{side}) is penalized more heavily than a front $\leftrightarrow$ side error (p_{mild}). The distance penalty p_{dst}^k assigns a small constant deduction when $\delta_k \neq \delta_k^*$.

F.3 Tier-3 Error Penalty Decomposition

The error penalty in Tier 3 decomposes as:

$$P_{\text{err}} = \hat{N}_{\text{FP-REJ}} \cdot p_{\text{rej}} + \hat{N}_{\text{FP-HAL}} \cdot p_{\text{hal}} \quad (12)$$

with $|p_{\text{rej}}| > |p_{\text{hal}}|$, encoding the navigation principle that violating an explicit filtering rule is more harmful than a generic hallucination. This asymmetry makes recorded rule violations more costly than novel hallucinations during optimization.

F.4 Total Reward Composition

The complete GRPO reward sums the content term with three structural terms:

$$R = \underbrace{R_{\text{fmt}}}_{\text{JSON validity}} + \underbrace{R_{\text{cot}}}_{\text{CoT quality}} + \underbrace{R_{\text{cst}}}_{\text{consistency}} + \underbrace{R_{\text{ct}}}_{\text{Eq. (5)}} \quad (13)$$

R_{fmt} validates that the output is parseable JSON with all required fields and legal enum values; a complete parse failure yields a large negative score $p_{\text{parse}} \ll 0$ that dominates all other terms. R_{cot} grants small bonuses for the presence of structural keywords (*discover*, *analyze*, *decide*) in the reasoning prefix, capped to prevent reward hacking via verbosity. R_{cst} checks that the number of “accept” decisions in the CoT equals the number of objects in the JSON output and the declared proper_obj_num field; mismatches indicate internal logical inconsistency and are penalized proportionally. Recent work (Kancheti et al., 2026) observes that GRPO-trained VLMs can produce CoT traces inconsistent with final outputs; R_{cst} penalizes this inconsistency explicitly.

G Rejection Taxonomy

We systematically categorize the rejection reasons from expert annotations into ten categories, enabling the model to learn explicit rejection reasoning:

1. **Insufficient Saliency:** The landmark is not visually prominent in the scene and cannot be spotted at a glance.
2. **Indistinct Appearance:** Nearby objects of the same category share similar appearances (e.g., multiple road signs, symmetrically distributed fences, rows of trees), making it impossible to distinguish based on the description alone.
3. **Inadequate Description:** The object’s appearance has unique features compared to similar objects, but the description fails to capture these distinguishing characteristics.
4. **Inappropriate Position:** (1) The object’s direction does not align with the walking direction; (2) at intersections, the landmark is located on the road after turning left/right rather than the straight path; (3) the object is on the left side of the vehicle lane, too far from the pedestrian sidewalk on the right.
5. **Suboptimal Choice:** Two or more objects in the scene are more salient and better suited for navigation guidance (e.g., small blue road name signs mounted on traffic signals).
6. **Misidentification:** Category recognition errors, such as identifying pedestrians or trees as sculptures, gates as archways, reflections on car windows as sculptures, or elevated highways as pedestrian overpasses.
7. **Incorrect Description:** The category is correctly identified, but the description contains errors, including appearance errors (e.g., blue lane signs described as green, blue facades as gray, white-signage shops as red-signage), position errors, distance errors, or range errors.
8. **Outdated Description:** The description contains time-sensitive content such as “0 parking spots remaining” or “20th anniversary celebration,” particularly for signage, billboards, and gates.

9. **Dynamic Object:** Non-static objects such as pedestrians, bicycles, or vehicles.

10. **Other:** Reasons not covered by the above nine categories.

H SFT Multi-Task Decomposition

To maximize the utility of limited expert annotations, we decompose the landmark extraction task into four auxiliary tasks plus the primary end-to-end extraction task for SFT training:

1. **Task 1 - Pass/Reject Classification:** Given a list of candidate objects with all attribute fields (category, description, orientation, distance, visibility range), the model must classify each as *pass* (with a saliency score 1–5) or *reject* (with a categorized reason). This task teaches explicit acceptance/rejection reasoning.
2. **Task 2 - Saliency Ranking:** Given all passed objects, rank them into three tiers based on visual saliency:
 - *Tier 1:* Extremely prominent—unique shape, very large size, or vivid colors; the dominant visual element.
 - *Tier 2:* Clearly visible—distinctive features that are easily noticed by pedestrians.
 - *Tier 3:* Visible but requires careful inspection due to surrounding distractors.
3. **Task 3 - Visibility Range Prediction:** Given passed objects with description, category, orientation, and distance, predict the visibility range $[i_{\text{start}}, i_{\text{end}}]$ indicating which image frames contain the object.
4. **Task 4 - Description Generation:** Given passed objects with category, orientation, distance, and visibility range, generate appropriate natural language descriptions that are distinctive yet concise.
5. **Task 5 - Full Prediction (Core Task):** Given only the input images, output all passed objects with complete attributes and generate a pseudo chain-of-thought reasoning trace that verbalizes the spatial analysis process.

Each task type is generated with both original and spatio-temporally cropped variants, yielding ~12.5k SFT samples total. The multi-task design

provides regularization that helps reduce overfitting to the core task while teaching complementary skills.

Temporal Cropping Equation For each sample with $T \geq 3$ frames, a random offset $\tau \sim \text{Uniform}(1, T-1)$ produces a suffix subsequence $\mathcal{I}' = \{I_{\tau+1}, \dots, I_T\}$, with object visibility ranges adjusted accordingly:

$$[s'_k, e'_k] = [\max(s_k - \tau, 1), e_k - \tau] \quad (14)$$

Objects with $e'_k < 1$ fall entirely outside the cropped window and are discarded. Together with naturally empty scenes (e.g., highway segments, severely blurred images), cropping-induced empty samples provide explicit supervision for the “output nothing” behavior that NARF later reinforces.

Pseudo Chain-of-Thought Format The pseudo-CoT prefixes the final JSON output with a structured *discover*→*analyze*→*decide* trace that explicitly verbalizes the spatial analysis for each candidate object:

```
1. discover: candidate object [white
   stone sculpture]
   analyze: visible in frames 2-4, a
   sculpture located
   on the right side of the pedestrian
   path. Distinctive
   appearance, easily noticed by
   pedestrians.
   decide: PASS

2. discover: candidate object [gray high
   -rise residential building]
   analyze: possibly located on the left
   side of the
   pedestrian path in frames 1-2,
   however, indistinct
   appearance among similar buildings.
   decide: REJECT

```json
{"proper_obj_num": 1, "results":
 [{...}]
```
```

For *passed* objects, the *analyze* step includes: (1) visibility range, (2) category, (3) relative position to pedestrian path, and (4) saliency justification. For *rejected* objects, the *analyze* step includes: (1) hypothesized visibility range and position, and (2) the specific rejection reason from the taxonomy (Appendix G). This format teaches the model to explicitly discover candidates, analyze both acceptance and rejection criteria, and decide before producing the final structured output.

I DPO Contrastive Pair Construction

We design seven types of contrastive pairs for DPO training, each targeting a specific failure mode:

1. Type 1 - Missing High-Saliency Object:

- *Condition*: $|\mathcal{Y}^+| \geq 2$
- *Chosen*: All passed objects
- *Rejected*: Randomly drop one object from the top-2 highest-saliency objects
- *Purpose*: Penalizes omitting important landmarks

2. Type 2 - False Acceptance:

- *Condition*: $|\mathcal{Y}^-| \geq 1$
- *Chosen*: Correct pass/reject assignment
- *Rejected*: Promote 1–2 rejected objects to pass (removing rejection reasons, assigning fake saliency scores)
- *Purpose*: Penalizes accepting unsuitable objects

3. Type 3 - Visibility Range Error:

- *Condition*: $|\mathcal{Y}^+| \geq 1$ and $|\mathcal{I}| \geq 2$
- *Chosen*: Correct visibility ranges
- *Rejected*: Perturb $[i_{\text{start}}, i_{\text{end}}]$ by shifting boundaries ± 1 –2 frames
- *Purpose*: Penalizes temporal localization errors

4. Type 4 - Orientation Error:

- *Condition*: $|\mathcal{Y}^+| \geq 1$
- *Chosen*: Correct orientations
- *Rejected*: Randomly change orientation (front/left/right) for 1–2 objects
- *Purpose*: Penalizes spatial reasoning errors

5. Type 5 - Incomplete Rejection Analysis:

- *Condition*: $|\mathcal{Y}^-| \geq 2$
- *Chosen*: CoT analyzes all rejected objects
- *Rejected*: CoT only analyzes a random subset of rejected objects
- *Purpose*: Encourages exhaustive rejection reasoning

6. Type 6 - False Rejection:

- *Condition*: $|\mathcal{Y}^+| \geq 2$

- *Chosen*: Correct pass/reject assignment
- *Rejected*: Demote one top-2 passed object to reject (assigning a random rejection reason)
- *Purpose*: Penalizes wrongly rejecting valid landmarks

7. **Type 7 - Redundant Prediction (Anti-Overfitting):**

- *Condition*: Any sample
- *Chosen*: Original output
- *Rejected*: Add redundant rejected objects by copying and perturbing existing objects
- *Purpose*: Penalizes verbose or repetitive outputs

Each type is generated with original and cropped variants, yielding ~6.8k DPO pairs total.