Seeing Is Not Sharing: Some Vision-Language Models Overestimate Common Ground in Asymmetric Dialogue

Nan Li, Albert Gatt, Massimo Poesio
Utrecht University, Utrecht, The Netherlands
{n.li, a.gatt, m.poesio}@uu.nl

arXiv:2606.31719v1 [cs.CL] 30 Jun 2026

Abstract

In collaborative dialogue, shared perception does not guarantee shared interpretation. Mutual understanding must be established through interaction. We investigate whether vision-language models (VLMs) can distinguish what *could* be shared from what *has been* shared between dialogue participants through grounding. We formulate this as an interpretation-matching task on 13,077 annotated reference expressions from HCRC MapTask dialogues, and evaluate VLMs under systematically controlled manipulations of dialogue context and map-information access. Our results show that providing authentic map images improves overall performance but shifts models toward over-predicting alignment. Textual descriptions of the same map content reproduce this bias, while non-informative images suppress alignment predictions entirely, indicating that the bias is driven by task-relevant map content, not the visual channel. This improvement comes at the cost of degraded accuracy on non-aligned cases. Calibration analysis and reference-chain tracking further suggest that models rely on static referential cues on the maps rather than tracking how grounding unfolds through dialogue history. We observe these patterns most clearly in Qwen3-VL-8B-Instruct and, to varying degrees, in four additional models from two architecture families. In models that exhibit the bias, map content, whether presented visually or textually, is treated as evidence of mutual understanding, conflating potential with established common ground.

1 Introduction

In everyday collaborative situated dialogue, two people attending to the situation may extract different information from it or interpret the same information in different ways. Grounding, the incremental process by which dialogue participants establish mutual understanding, is what bridges this gap (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark and Brennan, 1991). A speaker's reference expression (RE) becomes part of the common ground only after the addressee recognises, accommodates, and confirms it; until then, the RE remains *potential* rather than *established* shared knowledge.

While information asymmetry is present to some degree in all dialogue, in certain collaborative tasks it is introduced by design, making asymmetry a *structural and controlled* feature of the task. In what follows, we use *asymmetric dialogue* to refer specifically to settings where participants hold different private task-relevant information by design, and where neither can directly access the other's. The HCRC MapTask (Anderson et al., 1991) is a canonical instance: two participants navigate a route using maps that differ in landmark placement, landmark names, or the number of identically-named landmarks, so that apparent agreement can mask genuine divergence in interpretation. Because the discrepancies are designed into the maps, landmarks visible on both maps create *potential* referential overlap whose resolution depends entirely on the grounding process in the dialogue. A model evaluating such dialogue from the outside, as an overhearer with access to one or both maps, faces the same asymmetry: it can observe what *could* be shared, but must infer from the dialogue what *has been* shared. Recent perspectivist annotation work (Li et al., 2026) has made it possible to evaluate, for each RE, whether two MapTask participants actually share the same grounded interpretation at a given point in the dialogue — a judgment we call *interpretation matching*.

We use this resource to investigate three research questions:

RQ1 Can large vision-language models (VLMs) capture personal interpretations of interlocutors toward the same reference expression in

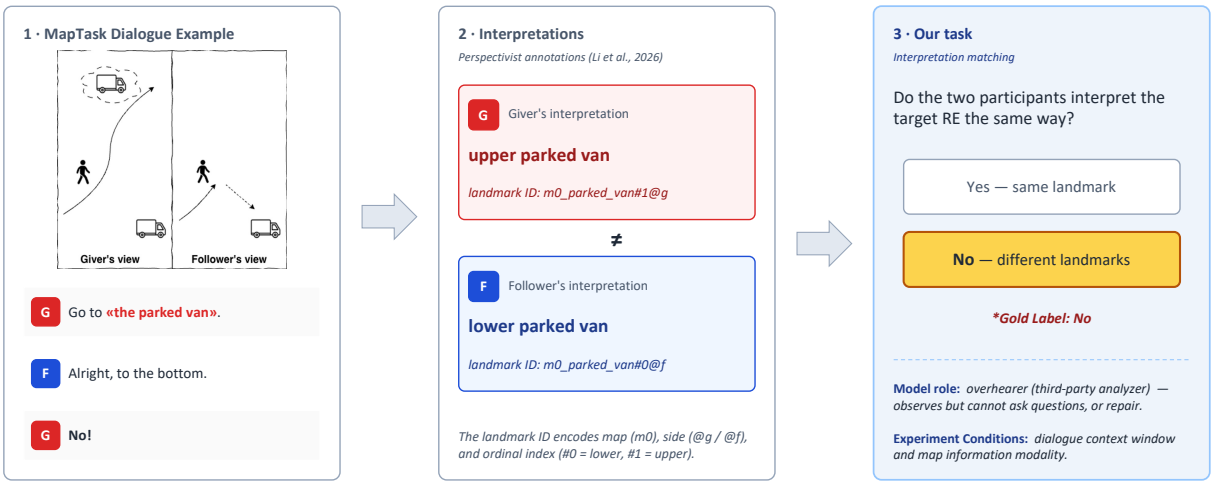

Figure 1: The interpretation matching task, illustrated on a misalignment example. **Panel 1** shows a simplified example of a MapTask dialogue with the target RE *the parked van*: the giver's map contains two parked vans while the follower's map has only one. **Panel 2** shows the perspectivist annotations (Li et al., 2026): the giver grounds *the parked van* to the upper van (`m0_parked_van#1@g`, see Appendix A for the definition about the landmark ID) and the follower to the lower one (`m0_parked_van#0@f`). **Panel 3** formulates our task: given the map(s) and dialogue, we ask the model to decide whether the two participants' interpretations of the marked RE match. The gold label here is NO.

asymmetric dialogue?

RQ2 Which modality of information contributes more to VLMs' assessment of interpretation alignment?

RQ3 Do VLMs exhibit systematically different behaviours on different types of alignment and misalignment cases?

We frame the task as a binary judgment — do the two participants' interpretations of a marked RE match? — and evaluate VLMs under systematic manipulations of two independent variables: the amount of available dialogue context and the type of map information provided. We evaluate models from two open-source VLM families, Qwen3-VL and Gemma3, at scales from 2B to 12B parameters. Qwen3-VL-8B-Instruct, the best-performing model in preliminary evaluation, serves as the primary model for detailed condition-grid analysis; the remaining models are compared on a baseline grid to assess cross-model generality (§3.3, §5.4).

Our main findings are:

- Providing authentic map images shifts the model toward over-predicting alignment. The model treats landmark co-presence as evidence of mutual understanding, conflating *what could be shared* with *what has been shared*. Textual descriptions of the same map content reproduce this bias, showing that it is driven by task-relevant map content rather than by the visual channel.
- Non-informative visual inputs (blank maps, shuffled landmarks) do not reproduce the bias; they make the model *more conservative*, not more prone to predicting alignment. The bias therefore possibly requires map *content*, whether delivered visually or textually.
- Calibration and reference-chain analysis converge on the same explanation: the model relies more on static referential cues on the maps rather than tracking how grounding unfolds through dialogue history.

Our work contributes (1) an evaluation methodology for probing VLMs' ability to distinguish potential from established common ground in information-asymmetric dialogue; and (2) an empirical characterisation of a model-dependent failure mode, observed most clearly in Qwen3-VL-8B, in which VLMs conflate possible referential overlap with communicative alignment.

2 Related Work

Grounding and Overhearers Common ground is built incrementally through interaction: interlocutors negotiate, confirm, and repair meaning, rather than receiving it directly from shared perceptual access (Clark and Wilkes-Gibbs, 1986;

Clark and Brennan, 1991). A well-established consequence is the *overhearer illusion*: overhearers can hear every word but, lacking the ability to contribute grounding acts, reach systematically weaker interpretations than addressees (Schober and Clark, 1989). This asymmetric performance carries over to language models: dialogue systems trained and evaluated on static transcripts are structurally overhearers, which shapes what they can learn about grounding and clarification (Madureira and Schlangen, 2024). Our evaluation makes that stance explicit, placing VLMs as overhearers of asymmetric human dialogue and testing whether they mistake co-presence of potential referents on the maps, delivered either visually or as text, for established shared interpretation.

Reference in Dialogue and VLM Evaluation Reference corpora have long served as testbeds for how interlocutors build shared interpretations through repeated mention, partial information, or clarification (Anderson et al., 1991; Haber et al., 2019; Udagawa and Aizawa, 2019; Chiyah-Garcia et al., 2023), yet each RE is typically annotated with a single gold referent, implicitly assuming speaker-addressee convergence. Under information asymmetry this breaks: the perspectivist annotation of Li et al. (2026) shows that the two participants can hold distinct grounded interpretations even under apparent agreement, the judgement of which overhearers make hard. Recent VLM evaluations report systematic gaps with humans as well: VLM overhearers underperform human matchers on referential dialogue and do not improve with repeated discussion (Wang et al., 2025), and VLM participants fail to entrain, form conceptual pacts, or initiate grounding acts as humans (Zeng et al., 2026; Shaikh et al., 2025). We directly ask whether a VLM, given the same asymmetric evidence as the participants, can recognise that they have not yet reached a shared interpretation, evaluating VLMs on perspectivist MapTask annotations under a variety of controlled map-information conditions.

3 Experimental Setup

3.1 Dataset

We use a published corpus of perspectivist annotations of the HCRC MapTask corpus (Li et al., 2026) that separately records, for each RE, the speaker's intended landmark and the addressee's interpreted landmark. The dataset comprises 13,077 annotated REs from 128 HCRC MapTask dialogues. Each dialogue involves two participants (a giver and a follower) collaborating to reproduce a route on the follower's map under the giver's guidance, with slightly different maps. The discrepancies can include landmark name differences, missing landmarks, and differences in quantity, creating a rich environment for misalignment in grounding. See Appendix A for more dataset details.

3.2 Task Design

We want to evaluate VLMs' *interpretation matching* ability, and formulate this as a binary judgment task: *Given a MapTask dialogue excerpt containing one marked RE, the model must decide whether the two participants currently share the same grounded interpretation of that expression or not.* Figure 1 illustrates a misalignment case, where the giver and follower ground *the parked van* to different landmarks.

An instance is labelled YES (aligned) when the two participants' landmark IDs match; and NO (not aligned) otherwise, covering both pending states (not yet grounded) and misunderstandings (grounded to different landmarks). The ground truth class distribution of the corpus is imbalanced: 72.1% aligned (YES) and 27.9% not aligned (NO).

We manipulate two independent variables to investigate how information access shapes model judgments: *text access* (the dialogue context window) and *map access* (the map-information modality).

Text access (dialogue context window) We vary how much dialogue context the model receives via four windows of increasing size: **curL** (current transaction, up to and including the line containing the target RE), **curT** (current transaction in full), **startL** (dialogue from beginning through the line containing the target RE), and **startT** (dialogue from beginning through the end of the current transaction). A *transaction* is a human-annotated MapTask dialogue excerpt that typically corresponds to a sequence of movements from one landmark to another.

These windows allow us to test whether (1) broader dialogue history, which encodes prior grounding episodes, and (2) future interactions, which usually encode repair sequences and clarification exchanges, help the model track grounding.

Map access (map-information modality) We vary what map information the model receives. In the *baseline condition grid*, conditions include:

Text-only (no map information), **Both maps** (authentic giver and follower map images), and **Giver-only / Follower-only** (a single authentic map image).

To further investigate the impact of map information and disentangle content from input channel, the *full-grid* modality experiment adds four conditions. Two are textual: **Text-landmark-names** (a textual list of landmark names on each map) and **Text-discrepancy-detail** (textual descriptions about the map discrepancies). Two are non-informative visual controls: **Blank maps** (uniform empty images, 1024×1024, RGB 128/128/128) and **Shuffled maps** (map images with landmarks from an unrelated map pair). Appendix D shows concrete example fillings of `${map_access}` for the two textual conditions.

The blank-map and shuffled-map conditions serve as controls: if the over-alignment bias is a generic multimodal artifact, these conditions should reproduce it; if it is content-driven, they should not.

Prompt design We apply zero-shot learning. The model receives a system prompt framing it as a dialogue analysis expert overhearing a MapTask dialogue. The prompt further describes map asymmetry (landmarks may be missing, duplicated, or placed differently), and defines the interpretation-matching task. The user prompt provides the target RE, the dialogue context with the target RE wrapped in «. . . » markers, and (where applicable) map images. Full prompt templates are given in Appendix C.

3.3 Models and Inference

We select models from two open-source VLM families that represented the state of the art at the time of our experiments: Qwen3-VL (Bai et al., 2025) and Gemma3 (Gemma Team, 2025). Within each family, we evaluate models ranging from 2B to 12B parameters (Qwen3-VL-2B/4B/8B-Instruct; Gemma3-4B/12B-it), constrained by the memory of a single A100 GPU. All models are instruction-tuned, non-thinking versions. In preliminary evaluation across all five models, Qwen3-VL-8B-Instruct achieved the highest macro-F1, so we use it as the primary model for the full condition-grid analysis. The remaining four models are evaluated on the baseline condition grid to assess generality (§5.4).

We use vLLM (Kwon et al., 2023) to deploy the models and to accelerate the inference progress. We use greedy decoding (temperature 0) with constrained output via vLLM logit masking to valid YES/NO tokens, ensuring valid binary responses. We set `random_seed` to 42 in vLLM's sampling parameters to ensure reproducibility.

Map images are resized to a maximum side length of 1024 pixels and delivered as PNG. We selected 1024 pixels as the minimum resolution at which Qwen3-VL-8B-Instruct reliably identified landmark names, icons, and locations in preliminary checks.

3.4 Evaluation

We report accuracy, macro-averaged F1, per-class recall ($\text{recall}_{\text{pos}}$ for aligned, $\text{recall}_{\text{neg}}$ for not-aligned), and the model's *yes-rate* (proportion of YES predictions) as a measure of response bias. In addition, we conduct three further analyses in §5: calibration analysis using token-level logits (§5.1), a status-level breakdown that decomposes performance by grounding state (§5.2), and reference-chain tracking that examines how predictions evolve across repeated mentions of the same landmark (§5.3).

4 Results

Table 1 presents the main results across the baseline conditions. All results in this section use Qwen3-VL-8B-Instruct; cross-model generality is assessed in §5.4. Here are the key findings:

F1: Map access improves detection but shifts the model toward over-predicting alignment. Comparing within the same model (Qwen3-VL-8B), at startT, adding both maps improves F1_{macro} from .591 (text-only) to .671 (both maps), a gain of .080. However, this improvement is driven by a dramatic shift in recall profile: $\text{recall}_{\text{pos}}$ rises from .590 to .822, while $\text{recall}_{\text{neg}}$ drops from .677 to .518. The yes-rate shifts from .515 to .727, exceeding the gold base rate of .721. The evidence for over-prediction is not the yes-rate alone (which is close to the base rate) but the *direction of the recall trade-off*: map access systematically pushes $\text{recall}_{\text{neg}}$ down while inflating $\text{recall}_{\text{pos}}$, meaning the model sacrifices its ability to detect non-aligned cases in favour of aligned ones. Calibration analysis in §5.1 and the status-level breakdown in §5.2 further confirm this: map conditions produce confident errors specifically on gold-NO instances. We analyse the status-level consequences of this shift

Map access	Text	Accuracy	$F1_{macro}$	$Recall_{pos}$	$Recall_{neg}$	Yes-rate
Text-only	curL	.442	.439	.261	.910	.213
	curT	.639	.604	.647	.616	.574
	startL	.533	.532	.409	.855	.336
	startT	.614	.591	.590	.677	.515
Both maps	curL	.627	.593	.633	.609	.566
	curT	.654	.618	.665	.625	.584
	startL	.694	.630	.769	.501	.694
	startT	.737	.671	.822	.518	.727
Giver-only	curL	.626	.586	.650	.565	.590
	curT	.688	.632	.747	.536	.668
	startL	.723	.644	.827	.454	.749
	startT	.756	.669	.879	.436	.791
Follower-only	curL	.619	.569	.663	.504	.617
	curT	.653	.594	.716	.490	.658
	startL	.718	.632	.832	.422	.761
	startT	.743	.650	.872	.408	.794

Table 1: Baseline results across text-access and map-access conditions for the model Qwen3-VL-8B-Instruct. Best $F1_{macro}$ per map-access level is **startT** in all map conditions; for text-only, curT slightly outperforms startT.

in §5.2, including its differential impact on aligned, pending, and misunderstood instances.

Single-map conditions amplify the bias. The giver-only condition at startT achieves similar $F1_{macro}$ (.669) to both maps (.671), but with an even higher yes-rate (.791) and lower $recall_{neg}$ (.436). The follower-only condition shows a comparable pattern (yes-rate .794, $recall_{neg}$ .408). Access to either map is sufficient to trigger the over-alignment bias, and providing both maps actually moderates it slightly by introducing cross-map discrepancy evidence.

Map info	$F1_{macro}$	R_{pos}	R_{neg}	Yes-rate
Baseline				
Text-only	.591	.590	.677	.515
Both maps	.671	.822	.518	.727
Textual map information				
Landmark names	.636	.756	.533	.675
Discrepancy Desc.	.668	.810	.528	.716
Fake visual controls				
Blank maps	.407	.220	.910	.184
Shuffled maps	.402	.222	.884	.193

Table 2: Map-information modality comparison at startT (both-maps access level). All conditions use Qwen3-VL-8B-Instruct. R_{pos} = recall on aligned; R_{neg} = recall on not-aligned.

F2: Textual map descriptions reproduce the over-alignment bias; only content-free visual controls avoid it. To determine whether the over-alignment bias is driven by map *content* or by the visual input channel, we compare authentic maps against textual map descriptions and non-informative visual controls. All conditions use the same model (Qwen3-VL-8B-Instruct) at the startT text window, ensuring a clean within-architecture comparison. Table 2 shows the results.

Blank maps (yes-rate .184) and shuffled maps (yes-rate .193) produce *lower* yes-rates than the text-only baseline (.515), let alone authentic maps (.727). The model becomes more conservative when it receives images from which it cannot extract task-relevant content. This rules out the hypothesis that the over-alignment bias is caused by the mere presence of images rather than by the information they carry.

Both textual conditions produce yes-rates (.675–.716) close to real maps (.727) and well above the text-only baseline (.515). Macro $F1_{macro}$ for the textual conditions (.636–.668) sits just below real maps (.671) and well above text-only (.591). The recall profiles echo real maps: elevated $recall_{pos}$ (.756–.810) and reduced $recall_{neg}$ (.528–.533), the same directional shift as real maps (.822 / .518), only slightly less extreme.

The same landmark information — which landmarks appear on which maps and where they differ — triggers over-alignment whether it is presented visually or textually. The over-alignment bias is therefore about *what the model learns about the scene*, not about the visual channel per se. The visual presentation does contribute a small additional shift (a few points on yes-rate and on $recall_{pos}$) relative to the textual presentation of the same content, consistent with spatial co-presence in images being

a slightly stronger perceptual cue, but this residual effect is small compared to the content effect itself.

The conditions thus split into two groups by yes-rate: inputs with task-relevant map content (real maps .727, textual descriptions .675–.716) over-predict alignment; inputs without map content (text-only .515, blank maps .184, shuffled maps .193) do not. Map content thus seems to drive the bias, while the visual presentation channel has only a secondary amplifying effect.

F3: Broader dialogue context helps, but this is mitigated by map access. The text-window effect (curL → startT) produces a .152 $F1_{macro}$ gain in the VL text-only condition, but only .078 under both maps. This means when the model has no access to maps, giving it more dialogue context helps a lot more. When it already has both maps, extra dialogue context still helps, but much less. This smaller gain suggests that, under map access, the model relies more heavily on static referential cues from the maps and benefits less from additional dialogue evidence about how grounding unfolds over time.

Future dialogue interactions also help. Across all map conditions, both curT → startT and curL → startL produce consistent gains in $F1_{macro}$. This means that subsequent turns often contain useful repair, confirmation, or clarification evidence that makes the grounding outcome more legible.

5 Further Analysis and Discussion

The preceding results show that map access improves overall performance while amplifying the over-alignment bias, and that this bias is driven by task-relevant map content rather than by the visual channel itself. We now probe the mechanism behind this pattern through calibration analysis (§5.1), a status-level trade-off analysis (§5.2), reference-chain tracking (§5.3), and cross-model comparison (§5.4).

5.1 Over-Prediction and Calibration

Since we use vLLM's constrained decoding to force a single YES/NO token, the cumulative log-probability of the generated token directly gives the model's *confidence*: conf := exp(logprob). We also compute Expected Calibration Error (ECE) (Pakdaman Naeini et al., 2015; Guo et al., 2017) to measure how calibration quality varies depending on whether the model's default response happens to be correct.

Condition	ECE	ECE_{yes}	ECE_{no}	Conf.
Text-only	.249	.263	.235	.863
Both maps	.174	.094	.403	.912
Giver-only	.172	.057	.484	.927
Follower-only	.185	.061	.524	.929

Table 3: Calibration by gold label at startT (n = 13,077). All conditions use Qwen3-VL-8B-Instruct. ECE_{yes}/ECE_{no} = ECE on gold-YES/gold-NO instances. Conf. = mean prediction confidence.

Table 3 shows the results with Qwen3-VL-8B-Instruct at startT across all 13,077 instances. Here are the key findings:

Map conditions are better calibrated on aligned instances but miscalibrated on non-aligned ones. Under both-maps, the model is biased toward YES (yes-rate .727) and is well-calibrated on gold-YES instances (ECE_{yes} = .094) but badly miscalibrated on gold-NO instances (ECE_{no} = .403). The text-only condition, which is not strongly biased toward either class (yes-rate .515), is moderately calibrated on both classes (ECE_{yes} = .263, ECE_{no} = .235). The asymmetry is sharpest under single-map conditions: follower-only achieves ECE_{yes} = .061 but ECE_{no} = .524.

Maps make the model more confident, and more confidently wrong on non-aligned instances. Mean confidence rises modestly with map access (.863 → .912 / .927 / .929), but the distribution of that confidence shifts: the ECE gap between gold-YES and gold-NO instances widens from .028 (text-only) to .463 (follower-only). On gold-YES instances the model is confident and right; on gold-NO instances it is equally confident but wrong. In other words, when the model gets non-aligned instances wrong, it is confidently wrong. Map evidence drives the model to predict YES with high certainty even when alignment does not hold. This class-conditioned miscalibration is the strongest evidence for over-prediction: on gold-NO instances, the model is not merely wrong but *confidently* wrong, indicating that map content drives spurious certainty rather than merely shifting a threshold.

5.2 The Status-Level Trade-Off

The overall improvement from map access conceals an asymmetric trade-off. Table 4 decomposes accuracy by the gold grounding status of each RE: *aligned* (gold YES; n = 9,435), *pending* (not yet grounded; gold NO; n = 3,403), or *misunderstood*

(grounded to different landmarks; gold NO; n = 239).

Condition	Aligned	Pending	Misund.
Baseline			
Text-only	.590	.691	.473
Both maps	.822	.523	.456
Giver-only	.879	.441	.372
Follower-only	.872	.419	.255
Textual map information			
Landmark names	.756	.544	.372
Disc. detail	.810	.540	.368
Fake visual controls			
Blank maps	.220	.913	.866
Shuffled maps	.222	.886	.858

Table 4: Accuracy by gold grounding status at startT (both-maps access level). Each cell shows the fraction of correct predictions within that status group. All conditions use Qwen3-VL-8B.

Maps boost aligned accuracy at the cost of pending and misunderstood accuracy. Maps boost aligned accuracy by 23–29 percentage points (.590 → .822 / .872 / .879) while simultaneously dropping pending accuracy by 17–27 points (.691 → .419 / .441 / .523; McNemar $p < 10^{-6}$ for all map conditions). On misunderstood REs, the effect is condition-dependent: the both-maps drop (.473 → .456) is not significant ($p = 0.724$, $n = 239$), but single-map conditions show large significant declines. Follower-only drops to .255 ($p = 3 \times 10^{-9}$) and giver-only to .372 ($p = 8 \times 10^{-3}$).

Both maps moderate the misunderstood collapse. Having both maps is actually the *least extreme* map condition on misunderstood (0.456 vs. 0.372 giver-only vs. 0.255 follower-only), likely because cross-map discrepancies provide corrective evidence that moderates the bias relative to single-map conditions. The accuracy drop on pending cases, by contrast, is broad and significant for *all* map conditions.

Textual descriptions follow the same trade-off as authentic maps. Textual map descriptions show the same directional profile as authentic maps relative to the text-only baseline (.590 / .691 / .473 for aligned / pending / misunderstood): aligned accuracy rises sharply while pending and misunderstood accuracy drop. Discrepancy-detail reaches .810 / .540 / .368 and landmark names .756 / .544 / .372 — close to both-maps (.822 / .523 / .456) on aligned and pending, but notably *below* both-maps on misunderstood. Their $F1_{macro}$ (.636 / .668) sits just under both-maps (.671).

Fake visual controls, by contrast, push the profile in the opposite direction: blank and shuffled maps produce near-identical hyper-conservative patterns (.220 / .222 aligned, .913 / .886 pending, .866 / .858 misunderstood), collapsing into a near-constant NO response (yes-rates .184 / .193).

The conditions therefore form two groups: content-rich inputs (real maps and textual descriptions) over-predict alignment, while content-free visual inputs (blank, shuffled) amplify caution to the opposite extreme. Telling the model what is on the maps or showing it produces similar behaviour; it is the presence of task-relevant content about potential common ground that drives the trade-off, not the modality.

The model confuses potential with established common ground. Map content tells the model what *could* be shared (landmarks appearing on both maps); dialogue history tells it what *has been* shared (interpretations established through interaction). The model over-weights the former. This is the computational analogue of the *overhearer's illusion*: overhearers systematically overestimate their understanding of a conversation because they have access to referential context but lack the interactive grounding process that establishes mutual understanding between participants. The model is structurally an overhearer, which means it observes what both participants could share but cannot well assess what they have confirmed through dialogue.

5.3 Reference-Chain Analysis

We use the *reference chains* defined in Li et al. (2026), sequences of mentions of the same landmark within a dialogue, to test whether repeated mention helps the model converge on the correct judgment. We group 1,665 chains into six buckets by *chain length*, the number of mentions of that landmark in the dialogue (1, 2, 3, 4–5, 6–8, 9+), and report accuracy together with yes-rate, because chain-length effects are partly driven by response bias that macro F1 alone obscures. See Figure 2.

Over-prediction of alignment grows with repeated mention. Map conditions maintain high accuracy across chain lengths (both-maps: .681 at length 1 → .791 at 9+), while text-only degrades (.719 → .599; Figure 2a). However, this comes with steadily rising yes-rates under map conditions (both-maps: .549 → .789; single-map reaches .840–.845; Figure 2b). Since aligned REs dominate at

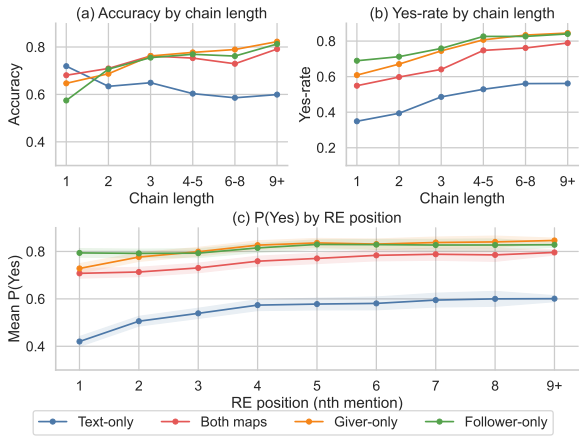

Figure 2: (a) Accuracy and (b) yes-rate by reference-chain length; (c) mean $P(\text{YES})$ by RE position (nth mention within a chain), with 95% CI shading. Map conditions maintain high accuracy but with steadily increasing yes-rate, and $P(\text{YES})$ rises with RE position in most situations.

longer chain lengths (post-grounding mentions accumulate), this inflation lets maps appear more accurate than they are. The same pattern emerges within chains: $P(\text{YES})$ rises steadily with RE position in most conditions, especially at the early mentions (Figure 2c) — text-only from .420 at the first mention to .601 at 9+, and map conditions (.708–.794) by +.034 to +.118.

5.4 Cross-Model Comparison

To assess whether the above findings and model behavioural patterns generalise beyond a single model, we evaluate four additional VLMs from two VL model families on the baseline condition grid at startT. Table 5 shows the results.

	Text-only		Both maps			
Model	$F1_m$	YR	$F1_m$	YR	ECE	Conf
Qwen3-VL-2B	.561	.483	.566	.707	.144	.789
Qwen3-VL-4B	.518	.368	.411	.225	.494	.907
Qwen3-VL-8B	.591	.515	.671	.727	.174	.912
Gemma-3-4B	.445	.279	.510	.400	.457	.973
Gemma-3-12B	.322	.107	.416	.231	.531	.948

Table 5: Cross-model comparison at startT on the full dataset (n = 13,077). $F1_m$ = $F1_{macro}$; YR = yes-rate.

Models respond to map information in qualitatively different ways. Adding maps increases the yes-rate for Qwen3-VL-8B (.515 → .727), Qwen3-VL-2B (.483 → .707), and Gemma-3-4B (.279 → .400), the same over-alignment direction observed in the primary experiments. However, Qwen3-VL-4B responds in the *opposite* direction: maps make it more conservative (yes-rate .368 → .225, $F1_{macro}$ .518 → .411). Gemma-3-12B is extremely conservative across all conditions (yes-rate .107 text-only, .231 both-maps), barely engaging with the task. These divergent responses suggest that the over-alignment bias is not a universal property of vision-language architectures but depends on model-specific factors.

One likely contributor to Gemma3's weaker performance is its limited ability to parse the information-dense hand-drawn MapTask map images. The two families differ substantially in how they encode visual input. Qwen3-VL uses a native dynamic-resolution vision encoder (Bai et al., 2025): images are tiled into patches at their input resolution, producing a variable number of visual tokens that scales with image area. For our 791×1024 map images, this yields hundreds of tokens with 2D-RoPE-based spatial positional encoding, preserving fine-grained detail such as small landmark labels and icons. Gemma3 uses a SigLIP-based vision encoder (Gemma Team, 2025) that resizes images to a fixed 896×896 resolution and average-pools the patch representations down to a budget of 256 tokens per image, regardless of image size or content complexity. This aggressive compression likely discards spatial detail that is critical for reading the MapTask maps. Prior to the main experiments, we ran a sanity check in which all five models were prompted to list landmark names and describe their spatial positions on each of the 32 map images (Appendix E). Qwen3-VL models achieved 88–90% F1 on landmark identification, compared to 81–82% for Gemma3 models. Gemma3 models also introduced character-level naming errors and spatial mislocations that Qwen3-VL models did not produce. If a model cannot reliably extract map content, map access cannot produce the content-driven over-alignment bias we observe in Qwen3-VL.

Figures 7, 8, and 9 in Appendix F show the full condition-level breakdown. The status-level analysis (Figure 9) reveals that the aligned vs. non-aligned case trade-off observed for Qwen3-VL-8B generalises across models: map access consistently improves aligned-case F1 while degrading performance on pending and misunderstood REs, with the exception of Qwen3-VL-4B, where the overall conservative shift suppresses performance across all status categories.

Model size does not predict task performance. Within both families, the scaling relationship is non-monotonic: Qwen3-VL-2B outperforms 4B on both-maps (.566 vs. .411), and Gemma-3-4B outperforms 12B (.510 vs. .416). Calibration data reveals why: Qwen3-VL-2B achieves the best calibration (ECE = .144) because it is the least confident (mean confidence .789), while Gemma-3-12B achieves the worst (ECE = .531) at the second highest confidence (.948). Larger models become more confident without becoming more discerning. The bottleneck may be the tendency to commit to reference-related evidence-driven grounding establishment with excessive confidence.

6 Conclusion

We set out to test whether VLMs can judge *interpretation matching* in information-asymmetric collaborative dialogue, using an evaluation methodology based on systematically controlled map-information and dialogue-context conditions applied to the HCRC MapTask. The main result is not simply that maps help. Authentic maps improve overall performance, but they do so by pushing models toward YES: landmark co-presence is treated as evidence of mutual understanding. Textual descriptions of the same map content reproduce the same bias, while non-informative visual inputs (blank, shuffled) reverse it into hyper-caution. The problem is therefore not multimodality in general, nor a specific visual-perceptual cue, but a tendency to read task-relevant information about *potential* common ground as evidence that it has been *established* through grounding.

Further analyses show where this bias is structured. Under map conditions, models become confidently biased toward aligned judgments, gain on already aligned cases while losing accuracy on pending and misunderstood ones, and grow more likely to predict alignment across repeated mentions. Taken together, these patterns suggest that current VLMs are better at assessing potential referential overlap from map content than at tracking grounding as an interactional and incremental process. In MapTask terms, they can infer what the interlocutors *could* be talking about, but they do not reliably distinguish this from what the interlocutors have actually established through grounding.

Our experiments are limited to an overhearer setting in a single domain, so an important next step is to test interactive settings where models can ask for clarification, express uncertainty, or revise their judgments over time. More broadly, extending the evaluation beyond MapTask and relating the observed heterogeneity to model properties via mechanistic interpretability analysis (e.g., attention flow analysis, Zhang et al., 2025) may clarify whether the observed over-alignment is a general behavioural tendency in VLMs for collaborative dialogue, or a narrower failure mode tied to this task, setting, and/or model family.

Limitations

Generalization Our primary evaluation relies on one single model (Qwen3-VL-8B-Instruct) for the detailed condition grid, with additional models tested only on baseline conditions.

The dataset derives from HCRC MapTask dialogues, a single corpus with specific properties that may shape the observed effects: a 72.1%/27.9% class imbalance toward aligned cases, MapTask-specific discrepancy types (missing, duplicated, or renamed landmarks), and a rare misunderstood category (239 of 13,077 instances). Suitable datasets for replication should provide perspectivist annotations recording both interlocutors' referent interpretations separately, not just task success or dialogue-level grounding labels, which limits the pool of available corpora. Generalisation to other information-asymmetric settings remains to be established.

Task Design The evaluation places the model in an overhearer position, characterising a judgment failure rather than an interaction failure.

We use greedy decoding to elicit a binary judgment, which may not reflect the model's full distributional beliefs about alignment. A softer evaluation protocol (e.g., allowing multi-choice selection or free-form responses) might reveal a more nuanced picture.

Textual Map Reconstruction Our textual conditions (*Text-landmark-names* and *Text-discrepancy-detail*) convey landmark name lists and inter-map discrepancies, but they do not fully reconstruct the spatial layout of the maps. The behavioral gap between textual and visual conditions may therefore partly reflect this incomplete reconstruction rather than a genuine visual-channel effect.

Acknowledgments

We appreciate the helpful comments and suggestions from the anonymous reviewers. This work is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant NGF.1607.22.002, *Dealing with Meaning Variation in NLP*.

Code and Data Availability

We release our code and prompt templates to facilitate reproducibility[1].

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi, and Helen Hastie. 2023. 'What are you referring to?' Evaluating the ability of multi-modal dialogue models to process clarificational exchanges. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 175–182, Prague, Czechia. Association for Computational Linguistics.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition.*, pages 127–149. American Psychological Association.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. ACM.

Nan Li, Albert Gatt, and Massimo Poesio. 2026. Grounded misunderstandings in asymmetric dialogue: A perspectivist annotation scheme for MapTask. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4988–5001, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Brielen Madureira and David Schlangen. 2024. It couldn't help but overhear: On the limits of modelling meta-communicative grounding acts with supervised learning. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 149–158, Kyoto, Japan. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29. Association for the Advancement of Artificial Intelligence.

Michael F Schober and Herbert H Clark. 1989. Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127. Association for the Advancement of Artificial Intelligence.

[1] https://github.com/chnln/seeing-is-not-sharing

Zhengxiang Wang, Weiling Li, Panagiotis Kaliosis, Owen Rambow, and Susan Brennan. 2025. LVLMs are bad at overhearing human referential communication. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16758–16782, Suzhou, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peter Zeng, Weiling Li, Amie J. Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory Zelinsky, Susan E. Brennan, and Owen Rambow. 2026. LVLMs and humans ground differently in referential communication. *Preprint*, arXiv:2601.19792.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19781–19791. IEEE.

A Dataset and Annotation

Our experiments use the perspectivist annotation release of the HCRC MapTask corpus (Li et al., 2026), derived from the original corpus of Anderson et al. (1991). The release contains 13,077 reference expressions (REs) annotated across 128 dialogues, paired with 16 distinct map pairs (`m0`–`m15`, each used in 8 dialogues). For each RE the annotation records the giver's and follower's interpretations separately, which is what makes the YES/NO interpretation-matching task studied here well-defined.

Landmark discrepancy types The 16 map pairs contain four kinds of landmark variation: **identical** landmarks appear at the same position under the same name on both maps; **lexical variants** appear at the same position but under different names (e.g., *white water* on the giver's map vs. *rapids* on the follower's; 10 such pairs are documented); **existence discrepancies** (landmarks appearing on only one of the two maps); and **multiplicity discrepancies** (landmarks appearing *twice* on one map but only once on the other; 16 such landmarks are documented, one per map pair). Existence and multiplicity discrepancies are the structural source of most misalignment in this corpus; lexical variants are resolved by name unification during annotation so they do not on their own produce *misunderstood* REs.

Landmark ID format Because the original MapTask corpus uses the same landmark ID for all instances of same-named landmarks, the annotation release introduces a unified ID scheme that disambiguates multiplicity landmarks and tracks map-side provenance. Each landmark ID has the format

`<map_id>_<concept>#<ordinal>@<side>`,

where `<map_id>` is `m0`–`m15`; `<concept>` is the original landmark name (e.g., `diamond_mine`); `#<ordinal>` is present only for *multiplicity* landmarks that appear twice on one map, with `0` denoting the lower/bottom instance and `1` the upper/top; and `@<side>` is `g` for the giver's map or `f` for the follower's. Examples: `m9_stony_desert@g`, `m9_site_of_plane_crash#0@g`, `m2_stone_creek#1@f`.

Annotation cascade Each RE is annotated along a five-step cascade that models incremental resolution. Each step is evaluated only when the preceding conditions are met; when the cascade terminates early, all downstream attributes are set to `null` in the released data.

Step	Attribute	Perspective	Req.
1	`is_quantificational`	Speaker	`false`
2	`is_specified`	Addressee	`true`
3	`is_accommodated`	Addressee	`true`
4	`is_grounded`	Addressee	`true`
5	`is_imagined`	Addressee	—

Table 6: The five-step annotation cascade used to guide the LLM and improve the annotation quality (Li et al., 2026). The **Req.** column shows the value an attribute must take for the cascade to proceed; step 5 is terminal.

Understanding states Each RE is assigned one of three understanding states, derived post-hoc from the cascade attributes and the match between the giver's and the follower's interpretation fields (after lexical-variant unification). In the released dataset: **aligned** (both participants ground the RE to the same or equivalent landmark) accounts for 9,435 REs (72.1%); **pending** (the RE is quantificational, unspecified, unaccommodated, or otherwise ungrounded) accounts for 3,403 REs (26.0%); and **misunderstood** (both participants believe they

agree but ground to different landmarks) accounts for 239 REs (1.8%). The main-text analyses in §5.2 decompose model performance along these three states.

B Experimental Setup and Reproducibility

All inference runs use vLLM (Kwon et al., 2023) as the backend with greedy decoding (temperature = 0.0, `random_seed` = 42), `max_new_tokens` = 16, and constrained output via vLLM's logit masking to the YES / NO token choice. For the calibration analyses reported in §5.1, we additionally record the top-20 logprobs per generation step. All experiments run on a single NVIDIA A100 80 GB GPU. For every Qwen3-VL model we explicitly disable the built-in thinking mode to ensure a fair comparison with other models. All the models are accessed via the Hugging Face Transformers library (Wolf et al., 2020).

C Prompt Template

Figure 3 shows the full prompt template used for all conditions.

D Textual Map-Information Examples

This section shows the text that fills the `${map_access}` slot of the prompt template (Figure 3) for the two textual map-information conditions introduced in §3.2. Both examples are drawn from the same instance — dialogue `q1ec1` (map pair `m12`), target RE *a caravan park* — so the two variants can be compared directly.

`text-landmark-names`. Under this condition, the model is given only the list of landmark names on each participant's map. The system prompt's map-information line reads: *"You are given the list of landmark names on each participant's map (see below in the dialogue context)."* The `${map_access}` slot of the user prompt is filled with the block shown in Figure 4.

`text-discrepancy-detail`. In addition to the per-map landmark lists, this condition supplies an explicit textual summary of how the two maps differ (per-side exclusives, multiplicity landmarks, and shared landmarks). The system prompt's map-information line becomes: *"You are given the landmark names on each participant's map and a description of how the two maps differ (see below in the dialogue context)."* The `${map_access}` slot is filled with the block shown in Figure 5.

E Map Reading Sanity Check

Prior to the main experiments, we assessed whether the five VLMs can reliably read the hand-drawn MapTask map images by prompting each model with two tasks on all 32 maps: (1) list all landmark names visible on the map, and (2) describe each landmark's spatial position. Both tasks use greedy decoding with free-form text output (no constrained decoding).

Prompts. For task 1: *"List all landmark names you can see on this map. Output only the landmark names, one per line. Do not add numbering, descriptions, or any other text."* For task 2: *"For each landmark you can see on this map, describe its position on the map (e.g., top-left, upper-center, center, bottom-right). Format each line as: landmark name – position. Do not add any other text."*

Landmark listing. Table 7 reports recall, precision, and F1 for landmark-name identification across all 32 maps, evaluated against the gold landmark lists from the corpus metadata. Matching uses exact string comparison after lowercasing and whitespace normalisation.

Model	Recall	Precision	F1
Qwen3-VL-2B	.845	.910	.876
Qwen3-VL-4B	.861	.937	.897
Qwen3-VL-8B	.812	.956	.878
Gemma-3-4B	.766	.866	.813
Gemma-3-12B	.749	.907	.820

Table 7: Landmark-name identification on all 32 MapTask map images. Recall = fraction of gold landmarks listed; Precision = fraction of model outputs that match a gold landmark.

Qwen3-VL models achieve higher F1 (.876–.897) than Gemma3 models (.813–.820). Qwen3-VL-8B has the highest precision (.956), producing almost no spurious landmark names. Gemma3 models introduce character-level naming errors that Qwen3-VL avoids, such as "picker fence" instead of "picket fence" (both Gemma3 models) and "pits of forest fire" instead of "site of forest fire" (Gemma3-4B).

Spatial description. We also prompted each model to describe landmark positions on all 32 maps. Figure 6 shows the giver's map for map

Landmark	Model	Position
picket fence	Qwen3-VL-2B	top-left
	Qwen3-VL-4B	top-left
	Qwen3-VL-8B	top-left
	Gemma-3-4B	upper-right†
	Gemma-3-12B	top-left
east lake	Qwen3-VL-2B	top-right
	Qwen3-VL-4B	top-right
	Qwen3-VL-8B	top-right
	Gemma-3-4B	bottom-right†
	Gemma-3-12B	right-center†
FINISH	Qwen3-VL-2B	top-right
	Qwen3-VL-4B	top-right
	Qwen3-VL-8B	top-right
	Gemma-3-4B	center†
	Gemma-3-12B	right-center†
START	Qwen3-VL-2B	—
	Qwen3-VL-4B	bottom-left
	Qwen3-VL-8B	bottom-left
	Gemma-3-4B	center-left†
	Gemma-3-12B	—
camera shop	Qwen3-VL-2B	bottom-left
	Qwen3-VL-4B	bottom-left
	Qwen3-VL-8B	bottom-left
	Gemma-3-4B	center-left†
	Gemma-3-12B	top-left†
parked van (lower)	Qwen3-VL-2B	bottom-left
	Qwen3-VL-4B	bottom-left
	Qwen3-VL-8B	bottom-left
	Gemma-3-4B	bottom-right†
	Gemma-3-12B	bottom-left

Table 8: Spatial descriptions on map0g (Figure 6) for six landmarks where Gemma3 models make clear errors. Qwen3-VL models produce correct or near-correct positions on all 14 landmarks; only the six with Gemma3 errors are shown. "—" = not listed. † = position inconsistent with the map.

pair 0 (map0g), and Table 8 compares the outputs on this map for six landmarks where at least one Gemma3 model makes a clear spatial error. Qwen3-VL models produce correct or near-correct positions on all 14 landmarks; the six shown are selected to illustrate Gemma3's failure pattern. For example, Gemma-3-4B places east lake at bottom-right (it is at top-right), START and camera shop at center-left (both are at bottom-left), and picket fence at upper-right (it is at top-left). Gemma-3-12B similarly mislocates east lake and FINISH. Both Gemma3 models also output "picker fence" instead of "picket fence".

F Cross-Model Detailed Results

Figures 7, 8 and 9 provide detailed breakdowns of the cross-model comparison in §5.4.

System Prompt

```
You are a dialogue analysis expert. You are overhearing two participants doing a MapTask-style route navigation task.

MapTask Background:
- Each participant has their own map and they cannot see each other's map.
- One participant (the Giver) describes a route using named landmarks on their map.
- The other participant (the Follower) tries to follow the instructions on their own map.
- The two maps may differ (some landmarks may be missing, duplicated, or placed differently), so the two participants can end up with different personal interpretations even if the dialogue sounds smooth.

Task:
You will be given a dialogue context in which ONE target reference expression is marked with «...». Decide whether the two participants interpret that reference expression as pointing to the SAME specific landmark (interpretations match).

Guidelines:
- Use only the provided dialogue context and (if available) the provided map image(s).
- A match can be supported by clear evidence of successful grounding (e.g., consistent descriptions, confirmations, coherent subsequent navigation).
- If the expression indicates quantificational asking, or unspecified/unresolved grounding, treat it as NOT a match.

Output:
Answer with exactly one word: Yes or No.
- Yes = interpretations match.
- No = interpretations do not match.
Do not output anything else.

Information you can access in this instance:
- Dialogue text: ${text_access}
- Map information: ${map_access}
```

User Prompt

```
Below is the specific information for this judgement.

Target reference expression:
${target_ref}

Dialogue context (the target RE is wrapped in « »):
${context}

Maps: if map images are provided, they appear below.
```

Figure 3: Prompt template. Template variables (`${...}`) are filled per instance based on the text-access and map-access conditions. For example, under the **startT** text-access window and **both-maps** access level, `${text_access}` is filled with "*You can read the dialogue from the beginning of the conversation through the end of the transaction that contains the target reference expression (i.e., including the subsequent lines in that transaction after the target line).*" and `${map_access}` is filled with "*You are shown both the Giver's and the Follower's map images.*"; the two map images are appended after the user prompt.

```
Map landmark information:
- Giver's map landmarks: start, caravan park, old mill, abandoned cottage, fenced meadow, fenced meadow, west lake, trig point, monument, nuclear test site, east lake, farmed land, finish
- Follower's map landmarks: start, caravan park, picket fence, mill wheel, forest, abandoned cottage, fenced meadow, west lake, monument, golf course, east lake, farmed land
```

Figure 4: Example filling of ${map_access} under the text-landmark-names condition for dialogue q1ec1.

```
Map landmark information:
- Giver's map landmarks: start, caravan park, old mill, abandoned cottage, fenced meadow, fenced meadow, west lake, trig point, monument, nuclear test site, east lake, farmed land, finish
- Follower's map landmarks: start, caravan park, picket fence, mill wheel, forest, abandoned cottage, fenced meadow, west lake, monument, golf course, east lake, farmed land
Discrepancies between maps:
- Landmarks on Giver's map ONLY (not on Follower's): finish, nuclear test site, old mill, trig point
- Landmarks on Follower's map ONLY (not on Giver's): forest, golf course, mill wheel, picket fence
- Landmarks appearing multiple times: fenced meadow appears 2 times on Giver's map
- Shared landmarks (on both maps): abandoned cottage, caravan park, east lake, farmed land, fenced meadow, monument, start, west lake
```

Figure 5: Example filling of ${map_access} under the text-discrepancy-detail condition for dialogue q1ec1.

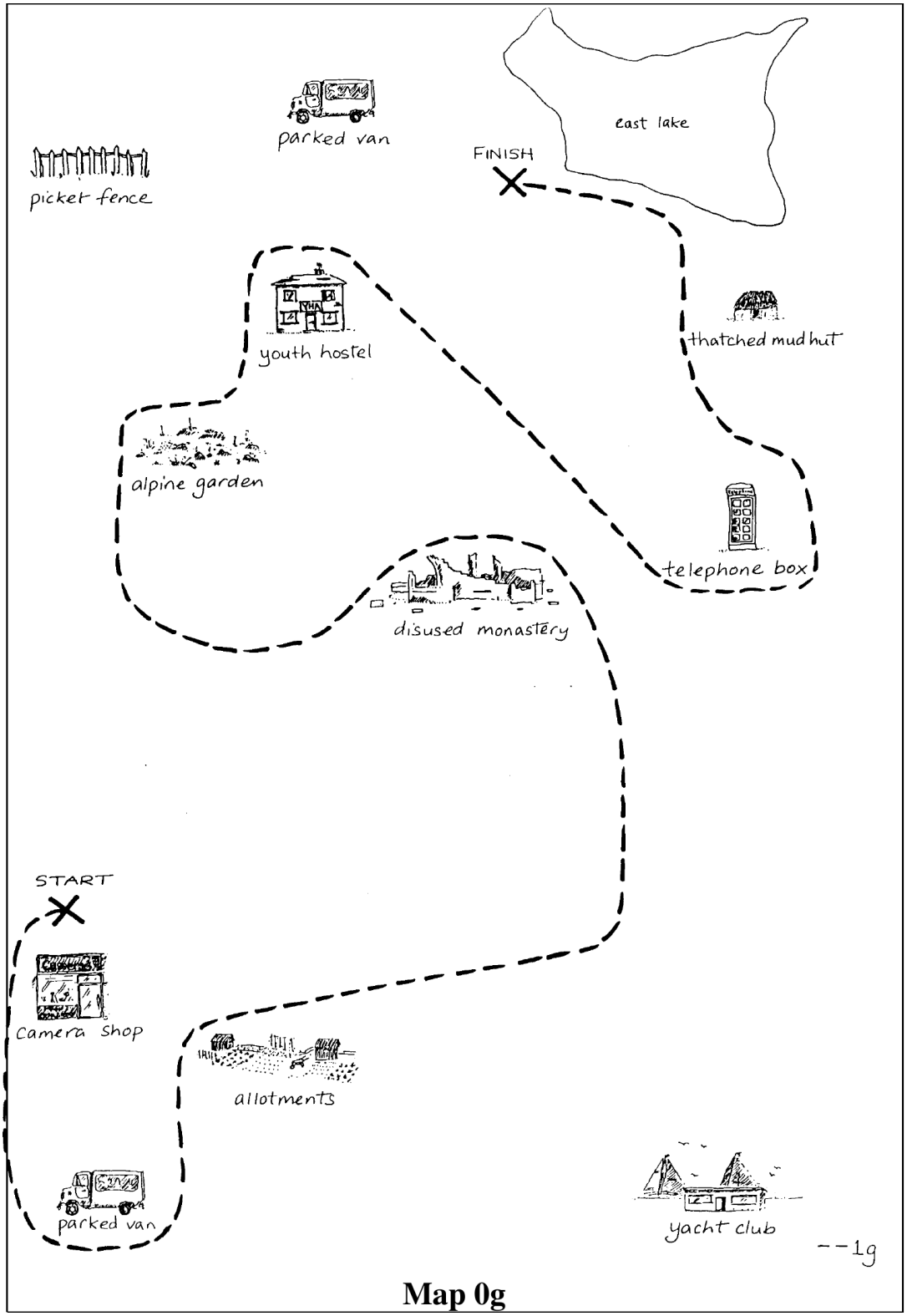

Figure 6: Giver's map for map pair 0 (map0g), used for the spatial-description comparison in Table 8.

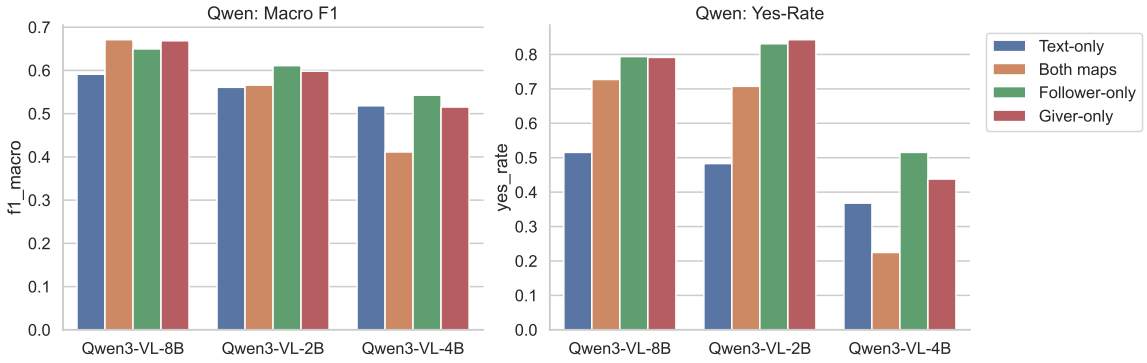

Figure 7: Macro F1 and yes-rate across map-access conditions for Qwen3-VL models (8B, 2B, 4B) at startT.

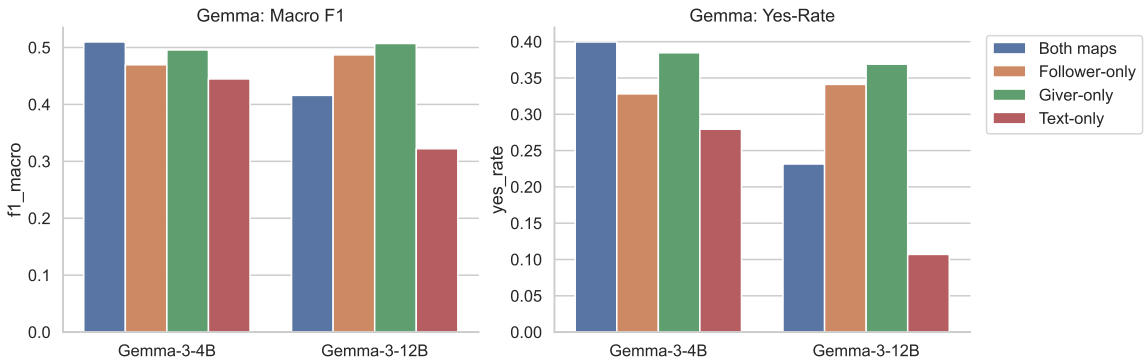

Figure 8: Macro F1 and yes-rate across map-access conditions for Gemma-3 models (4B, 12B) at startT.

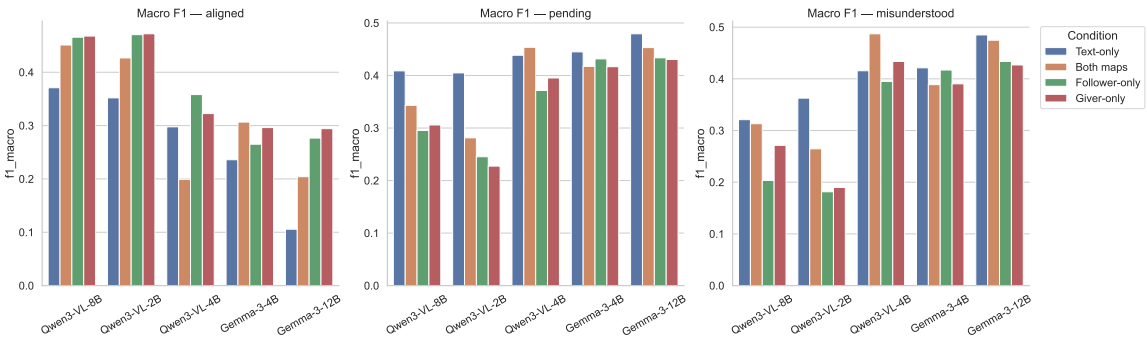

Figure 9: Macro F1 by grounding status (aligned, pending, misunderstood) across all models and map-access conditions at startT.