Beyond Local Surprise: Grounded Dialogue as Selective Belief Revision under Referential Uncertainty

Anonymous ACL submission

Abstract

When a speaker refers to a scene that the listener cannot directly see, the listener must decide whether to preserve its current understanding or revise it as new utterances arrive. Many language systems treat local mismatch as a cue for updating: divergence from the current understanding encourages adjustment. Yet conversational understanding may be more conservative, interpreting mismatch input relative to prior understanding rather than immediately revising it. We introduce a controlled, data-driven framework for turn-by-turn preserve/revise decisions in dialogue, where competing revision policies are learned under otherwise identical conditions. We compare four theory-driven revision strategies, each reflecting a different assumption about when listeners should preserve or revise. Two findings stand out. First, a mismatch-driven policy that updates solely based on local divergence reacts strongly to mismatch but destabilizes grounding and degrades retrieval. Second, an uncertainty-sensitive policy extends mismatch-based updating with accumulated evidence, preserving coherent understanding while maintaining strong retrieval performance. Surprisingly, coherent understanding emerges from a counterintuitive pattern: local mismatch promotes preservation, whereas accumulated uncertainty promotes revision, suggesting that listeners maintain prior understanding despite local mismatch and revise only when uncertainty sufficiently accumulates. This pattern is consistent with conceptual pact theory.

1 Introduction

Human communication succeeds because listeners must continually decide whether to preserve an established interpretation or revise it as new conversational evidence arrives. For example, when a speaker mentions *my new pet*, a listener may initially picture a dog; when a later turn references *her tank*, that interpretation may shift toward a fish or reptile. This process, known as conversational grounding (Clark and Schaefer, 1989; Clark and Brennan, 1991), is inherently incremental: each conversational turn is interpreted relative to what has already been established between speakers (Shaikh et al., 2024). Grounding therefore depends not only on updating understanding, but on deciding when prior interpretations should be preserved and when they should be revised. By constructing systems that solve this grounding problem turn by turn, we can use their behavior to study how interpretations are preserved and revised over time.

Many computational approaches to grounded language understanding model interpretation as an incremental update process, where interpretations are continuously revised as new evidence arrives. In incremental reference resolution, for example, models update a probability distribution over candidate referents as each new word narrows or shifts the interpretation (Kennington and Schlangen, 2017). Consider the referring expression *red book*: hearing *red* initially activates a broad set of possible objects, while *book* subsequently narrows the listener's belief toward a more specific referent (Yule, 2013). Implicit in many such approaches is a simple computational intuition: larger mismatch should warrant stronger revision of current understanding. Yet whether mismatch alone should determine when and how much listeners revise an established interpretation remains unclear. Incremental models demonstrate that understanding evolves over time, but leave open the question of how revision should be driven.

Conceptual pact theory suggests that conversational partners preserve shared interpretations rather than renegotiating meaning at every local mismatch (Brennan and Clark, 1996). This raises a different computational possibility: mismatch alone may not be sufficient to trigger revision. Instead, effective grounding may depend not on isolated mismatch, but on whether repeated mismatch accumulates into uncertainty about current under-

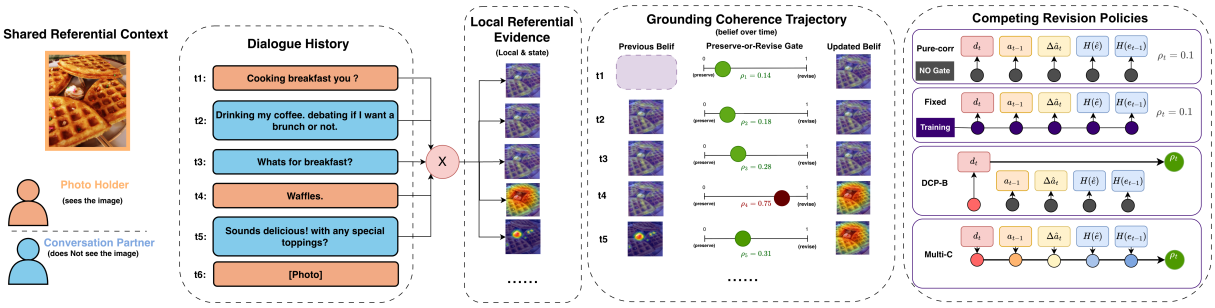

Figure 1: Overview of the preserve-or-revise framework for grounded dialogue. Grounding evolves over dialogue turns by integrating local referential evidence with prior belief through a preserve-or-revise mechanism. The underlying architecture is fixed while revision policies vary: PURE-CORR, FIXED, DCP-B, and MULTI-C.

standing. To investigate this idea, we introduce a simulation framework that compares competing revision policies in grounded visual dialogue (Figure 1) (Zang et al., 2021). Across revision policies, a clear pattern emerges: reacting strongly to local mismatch destabilizes grounding. Although local divergence reliably detects mismatch, using it as the primary trigger for revision repeatedly disrupts accumulated understanding. In contrast, stable grounding emerges when revision is guided by uncertainty over accumulated evidence. Unexpectedly, uncertainty promotes revision while surprise suppresses it, suggesting that listeners preserve interpretations despite local mismatch and revise only when uncertainty sufficiently accumulates. The code will be released upon acceptance.

2 Related Work

Conversational grounding is commonly understood as an incremental process through which interlocutors establish shared understanding over time (Clark and Schaefer, 1989; Clark and Brennan, 1991). When misunderstandings arise, prior work often frames grounding in terms of conversational repair, emphasizing clarification and recovery following confusion or communication breakdown (Schegloff et al., 1977; Traum, 1994). Dialogue systems similarly treat revision as a discrete decision about whether clarification is required under uncertainty (Shi et al., 2022; Testoni and Fernández, 2024). Our work differs in focusing not on whether repair occurs, but on the computational assumptions governing when grounding should be preserved or revised.

A complementary line of work studies how listeners revise interpretations incrementally as linguistic evidence accumulates (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Kennington and Schlangen, 2017). In situated reference resolution and grounded dialogue, models update interpretations incrementally as new evidence arrives, gradually narrowing or shifting toward likely referents (Kennington and Schlangen, 2017; Schlangen and Skantze, 2011; Li and Boyer, 2016). A common intuition in these approaches is simple: larger mismatch should lead to larger updates (Meister et al., 2024). However, prior work typically assumes a revision mechanism rather than treating revision itself as the object of study. While incremental grounding and repair have been extensively modeled, preserve-or-revise behavior is rarely formulated as an explicit computational question. We instead directly compare competing assumptions about how revision should be determined under mismatch and uncertainty.

Grounded multimodal dialogue provides a natural setting for studying conversational grounding because interlocutors must align language against a shared visual environment over time (Das et al., 2017; Kim et al., 2019; Zang et al., 2021). Prior work typically evaluates grounding through downstream task success such as retrieval or collaborative performance (Zang et al., 2021; Willemsen et al., 2023; Lee et al., 2024; Fried et al., 2023). Rather than treating grounded dialogue only as a task, we leverage it as a controlled setting for demonstrating how different revision assumptions shape grounding under identical interactions. This allows grounding behavior to be compared independently of changes in perceptual representations or interaction context.

3 Method

Figure 1 provides an overview of the preserve-or-revise framework. Grounding evolves through turn-level evidence accumulation over a shared visual

scene, while competing revision policies differ only in how preserve-or-revise decisions are determined.

Grounding as an Evolving Evidence State. We model conversational grounding as an evolving evidence state over image regions, allowing preserve-or-revise behavior to be analyzed turn by turn. The shared image provides a stable referential anchor, creating a controlled setting in which different revision assumptions can be examined under identical interactions.

Given a dialogue–image pair (D, I), where $D = u_{t=1}^{T}$ denotes a sequence of utterances and I a shared visual scene, the goal is to track how grounding evolves over time rather than simply determine whether dialogue and image match. At each turn t, the model maintains an evidence state e_t over image regions, representing its current estimate of what has been grounded.

At each turn, the model determines how strongly new linguistic evidence should update the accumulated grounding state. We model this through a scalar revision signal $\rho_t \in (0, 1)$ that balances prior grounding against newly observed evidence: low ρ_t preserves accumulated grounding, whereas high ρ_t favors revision toward new evidence. By varying how ρ_t is determined while holding dialogue and visual evidence fixed, the framework isolates competing assumptions about when revision should occur.

Perception as a Fixed Prior. To isolate grounding from perceptual learning, we keep a frozen CLIP encoder (ViT-B/32; (Radford et al., 2021)) fixed throughout training. This means the model's visual–linguistic knowledge remains unchanged, while only grounding is allowed to evolve over dialogue. Within conversation, the image serves as a stable referential anchor that new language is grounded against turn by turn. As new utterances arrive, grounding updates over image regions while visual representations remain fixed, allowing different revision policies to be examined under identical perceptual conditions. Utterances are encoded using CLIP text features and images using global and patch-level visual features. A shared trainable projection W_h maps language into the visual space for grounding attention, coherence estimation, and contrastive alignment, while visual representations are reused from frozen CLIP.

Turn-local evidence and accumulated grounding state. Each utterance provides new evidence about which parts of the shared scene are currently referenced. We compute this turn-local signal as a softmax attention distribution over patch-level image features:

$$\hat{e}_t = \mathrm{softmax}(q_t^\top V) \tag{1}$$

where $\hat{e}_t \in \Delta^{49}$ captures where the current utterance points within the visual scene. This signal is intentionally local: it reflects only what the current utterance demands, without knowledge of prior conversational context.

Grounding is not determined by a single utterance alone. The accumulated grounding state e_t integrates turn-local evidence with prior grounding through a revision-weighted interpolation:

$$e_t = (1 - \rho_t)\, e_{t-1} + \rho_t\, \hat{e}_t \tag{2}$$

initialized from a uniform prior $e_0 = \frac{1}{49}\mathbf{1}$. The scalar $\rho_t \in (0, 1)$ controls how strongly new evidence revises grounding at each turn. Different assumptions about ρ_t therefore capture different ways listeners may preserve or revise over time.

Coherence as an observable grounding trajectory. The evidence state e_t captures what is currently grounded, but does not by itself indicate whether grounding remains coherent across turns. We therefore track grounding coherence as an evolving trajectory. An evidence-weighted visual summary $c_t = e_t^\top V$ combines the current grounding state with image features, and turn-level coherence is estimated as:

$$\tilde{a}_t = \sigma\Big(W_a[q_t;\, c_t;\, v^{\mathrm{cls}}]\Big) \tag{3}$$

We then update coherence over time:

$$a_t = (1 - \lambda_t)\, a_{t-1} + \lambda_t\, \tilde{a}_t, \quad \lambda_t = 1 - \rho_t \tag{4}$$

This coupling allows local evidence and longer-term coherence to be tracked separately. The resulting signal a_t provides an interpretable trajectory of grounding coherence across dialogue turns.

Different policies about grounding revision. Rather than assuming a single mechanism for preserve-or-revise decisions, we compare four competing policies about how listeners should revise grounding over time. Each hypothesis represents a different assumption about the role of mismatch, uncertainty, and prior grounding stability under identical interaction settings. Comparing their behavior allows us to measure not only whether grounding succeeds, but what kinds of revision dynamics support coherent grounding.

- PURE-CORR: **Grounding as conservative accumulation.** Grounding evolves through weak, fixed accumulation ($\rho = 0.1$) without explicit preserve-or-revise supervision. This policy assumes that understanding changes gradually and serves as a minimal baseline for grounding dynamics.

- FIXED: **Grounding as constant-rate revision.** Listeners revise grounding at a fixed pace regardless of conversational context, implemented as a constant revision rate ($\rho = 0.1$) across all turns. Unlike PURE-CORR, preserve-or-revise objectives remain active, allowing grounding coherence to shape learning while revision strength itself stays fixed. This reflects the assumption that grounding should evolve steadily rather than adaptively.

- DCP-B: **Grounding as surprise-driven revision.** Revision is governed by mismatch between prior grounding and incoming evidence:

$$d_t = \mathrm{KL}(e_{t-1} \| \hat{e}_t) \tag{5}$$

  Larger divergence increases the revision signal ρ_t, shifting grounding toward new evidence. This reflects a common computational intuition: local mismatch should trigger stronger updating.

- MULTI-C: **Grounding as uncertainty-sensitive revision.** Rather than treating local mismatch as sufficient for revision, this policy assumes that preserve-or-revise decisions depend on broader contextual signals: prior grounding stability, uncertainty in current understanding, and whether trusting new evidence would improve coherence. Revision is therefore modeled as a learned function of five conversational signals:

$$\rho_t = \sigma(W_p\,[d_t,\, a_{t-1},\, \Delta\hat{a}_t,\, H(e_{t-1}),\, H(\hat{e}_t)]) \tag{6}$$

  where d_t measures local evidence divergence, a_{t-1} prior grounding coherence, $\Delta\hat{a}_t$ the expected coherence change if new evidence were trusted, and $H(\cdot)$ uncertainty in prior and incoming grounding. To avoid circular dependence, $\Delta\hat{a}_t$ is estimated from a counterfactual fully-trusted grounding state before updating occurs:

$$\hat{a}_t = \sigma\big(W_a[q_t;\, \hat{c}_t;\, v^{\mathrm{cls}}]\big)\,, \qquad \Delta\hat{a}_t = \hat{a}_t - a_{t-1} \tag{7}$$

  This parameterization evaluates whether revision should depend on multiple contextual signals rather than local divergence alone.

Learning grounding dynamics. Training is guided by preserve-or-revise constraints rather than explicit repair supervision. The objectives encourage stable grounding during low-revision turns and measurable grounding change during high-revision turns, while keeping perceptual evidence fixed.

- **Grounding coherence through contrastive alignment.** To ensure that grounding remains anchored to the shared visual scene over time, we optimize a cross-dialogue contrastive objective. At each turn, an utterance query q_t is encouraged to remain aligned with its corresponding dialogue image while staying distinguishable from images in other dialogues:

$$L_{\mathrm{corr}} = -\sum_t \log \frac{\exp(s_t^+/\tau)}{\sum_b \exp(s_b/\tau)} \tag{8}$$

  where $s_t^+ = \mathrm{sim}(q_t, v^{\mathrm{cls}}_{d(t)})$ demonstrates similarity to the correct image, $s_b = \mathrm{sim}(q_t, v^{\mathrm{cls}}_b)$ to negatives, and τ is a temperature parameter. This objective stabilizes grounding against the shared visual context, ensuring that preserve-or-revise dynamics unfold relative to a consistent referential target.

- **Preserve when evidence remains consistent.** The preserve objective penalizes unnecessary shifts during low-revision turns. When revision pressure is low, accumulated grounding is encouraged to remain stable rather than drift unnecessarily. We operationalize this through a preserve constraint that penalizes changes in grounding during low-revision turns:

$$L_{\mathrm{preserve}} = \tfrac{1}{T-1} \textstyle\sum_{t\geq 2} (1-\rho_t)_{\mathrm{detach}}\, \mathrm{KL}(e_{t-1} \| e_t) \tag{9}$$

  Here, ρ_t is detached so that revision acts as a state regulator rather than a directly optimized objective.

- **Revise when accumulated evidence warrants change.** The revise objective encourages grounding to change when revision pressure is high. When revision pressure is high, grounding is encouraged to change rather than remain artificially stable. We measure this through a revise constraint that penalizes turns

where accumulated grounding changes too little despite strong revision signals:

$$L_{\text{revise}} = \frac{1}{T-1} \sum_{t \geq 2} \rho_{t,\text{detach}} \max\Big(0,\, \delta - \text{KL}(e_{t-1} \| e_t)\Big) \quad (10)$$

where δ defines the minimum grounding shift expected when revision pressure is high. As in the preserve objective, ρ_t is detached so that revision remains a state-dependent regulator rather than a trivially optimized target.

- **Selective revision rather than constant updating.** We regularize the mean revision rate toward a target frequency r to encourage sparse revision. We therefore regularize the mean revision rate toward a target frequency r, encouraging revision to remain sparse rather than constant:

$$L_{\text{rate}} = \big(\text{mean}_t(\rho_t) - r\big)^2 \quad (11)$$

This objective operationalizes a preserve-by-default assumption, in which revision remains available but is expected to occur selectively rather than continuously.

The full objective combines these terms:

$$\mathcal{L} = L_{\text{corr}} + \alpha L_{\text{preserve}} + \beta L_{\text{revise}} + \gamma L_{\text{rate}} \quad (12)$$

encouraging stable grounding under consistent evidence with selective revision when needed.

4 Experimental Setup

Task and Dataset. We evaluate conversational grounding on PhotoChat (Zang et al., 2021), a grounded dialogue dataset in which one participant privately sees an image and reveals it only after a conversation unfolds. This creates referential uncertainty: interlocutors must establish shared understanding through dialogue alone, without direct access to the image. Because reference is refined across turns, successful interaction depends on coherent grounding over time rather than isolated utterance matching, making PhotoChat a natural setting for studying how grounding is preserved and revised under referential uncertainty.

To isolate incremental grounding, we retain only turns before image disclosure and treat the hidden image as the grounding target. This prevents trivial matching from explicit visual descriptions and requires grounding to evolve through accumulated conversational evidence across turns. The dataset contains 21 training shards. We use a deterministic shard-level split (19 train, 2 test; random seed 42), fixed across all experiments to ensure comparability across revision policies. Full hyperparameter settings are provided in Appendix A.

5 Evaluation Protocol

We evaluate grounding quality through visual retrieval. Given a dialogue trajectory, the model retrieves the target image from a gallery of candidates. Retrieval serves as a behavioral test of grounding quality rather than grounding itself: a model may retrieve the correct image, yet rely on unstable or incoherent revision dynamics. At the same time, grounding that fails to remain coherent should eventually impair retrieval. We therefore evaluate both behavioral performance and grounding dynamics, asking not only whether a model succeeds, but how that success is achieved. We report Recall@K (R@1, R@5, R@10), median rank (MedR), and mean reciprocal rank (MRR), together with revision behavior and coherence trajectories.

Reference Conditions. We include two reference conditions to contextualize grounding behavior under the four revision policies (PURE-CORR, FIXED, DCP-B, MULTI-C):

- **Random**: chance-level retrieval without grounding, providing a lower-bound.
- **CLIP-only**: frozen CLIP representations without grounding updates or task-specific training. Utterances are encoded independently and mean-pooled into a dialogue-level query, then matched to frozen image embeddings via cosine similarity. This provides a strong static reference without incremental grounding dynamics.

These conditions serve as behavioral anchors for interpreting how different revision policies shape grounding and retrieval.

5.1 Intrinsic Grounding Dynamics

To characterize how grounding evolves over time, we analyze three intrinsic metrics capturing complementary aspects of preserve-or-revise behavior. All metrics are computed over evaluation turns and aggregated across random seeds.

Preservation index (Pres). We measure grounding stability during low-revision turns ($\rho_t < 0.5$) as the mean evidence shift between consecutive

	Retrieval					*Grounding Dynamics*		
Regime	**R@1**	**R@5**	**R@10**	**MRR**	**MedR↓**	**Pres.↓**	**Rev. Sens.↑**	**Mean ρ**
Random	0.0011	0.0056	0.0112	0.0083	447	—	—	—
CLIP-ONLY	0.0873	0.2066	0.2630	0.1450	73	—	—	—
PURE-CORR	$0.118 \pm .005$	$0.314 \pm .005$	$0.424 \pm .006$	$0.219 \pm .006$	16.0 ± 1.0	$0.050 \pm .000$	—	0.100
FIXED	$0.117 \pm .006$	$0.315 \pm .003$	$0.426 \pm .006$	$0.218 \pm .007$	16.0 ± 1.0	$0.049 \pm .001$	—	0.100
DCP-B	$0.009 \pm .006$	$0.052 \pm .032$	$0.086 \pm .048$	$0.038 \pm .019$	184.0 ± 121.3	$0.001 \pm .001$	$0.900 \pm .015$	0.203
MULTI-C	$0.112 \pm .004$	$0.304 \pm .012$	$0.418 \pm .009$	$0.213 \pm .003$	16.0 ± 0.0	$0.083 \pm .001$	$0.598 \pm .097$	0.148

Table 1: Retrieval performance and intrinsic grounding dynamics across policies (mean $\pm$ std, three seeds). PURE-CORR and FIXED remain within one standard deviation of MULTI-C on all retrieval metrics. DCP-B exhibits the strongest intrinsic dynamics, including large mismatch-triggered switches (Switch KL $= 12.31 \pm 0.88$), yet substantially underperforms on retrieval.

grounding states: $\mathrm{KL}(e_{t-1} \| e_t)$. Lower values indicate stronger preservation of prior grounding when revision pressure is low.

Revision sensitivity (Rev Sens). We measure how strongly revision signals correspond to realized grounding updates by computing the Pearson correlation between ρ_t and the evidence shift $\mathrm{KL}(e_{t-1} \| e_t)$. Higher values indicate that larger revision signals are associated with larger grounding changes. This metric is undefined for PURE-CORR and FIXED, which use constant ρ_t.

Mean revision rate (Mean ρ). We report the mean revision rate ρ_t across turns, characterizing whether grounding evolves selectively or continuously. For DCP-B, we additionally report the average grounding shift (Switch KL) during high-revision turns ($\rho_t \geq 0.5$), capturing how strongly grounding changes once mismatch-driven revision is activated.

5.2 Modality Dependence

To verify that learned grounding dynamics reflect actual multimodal coordination rather than single-modality shortcuts, we perform inference-time perturbation experiments on MULTI-C. We selectively replace text or image representations with zero vectors, random Gaussian vectors, or empirical mean embeddings without retraining. If grounding emerges from coordinated preserve-or-revise behavior across modalities, perturbing either modality should impair retrieval toward chance. Differences across perturbation types further reveal how strongly grounding depends on linguistic versus visual evidence under different conditions.

6 Results

Retrieval Performance and Grounding Dynamics. Table 1 reveals an important contrast. Despite different grounding dynamics, PURE-CORR, FIXED, and MULTI-C achieve comparable retrieval performance across metrics. This suggests that successful grounding does not require aggressive revision, as long as accumulated understanding remains coherent over time.

DCP-B, however, collapses on retrieval. Although its revision signal closely tracks local evidence change (Rev. Sens. $= 0.900$), retrieval performance deteriorates substantially. Its intrinsic dynamics reveal why: once local mismatch triggers revision, grounding shifts abruptly (Switch KL $= 12.31 \pm 0.88$), repeatedly disrupting accumulated understanding. Rather than supporting coherent grounding, local mismatch drives overly reactive belief updates. High sensitivity to conversational change alone is therefore insufficient for successful grounding.

MULTI-C resolves this tradeoff. Retrieval remains within standard deviation of PURE-CORR and FIXED, while maintaining structured revision behavior (Rev. Sens. $= 0.598$, Mean $\rho = 0.148$). Rather than reacting strongly to isolated mismatch, revision remains selective and distributed across turns, preserving accumulated grounding while remaining responsive to new evidence.

Local divergence is associated with reduced revision. Table 2 reveals a counterintuitive pattern. Contrary to the common intuition that local mismatch should trigger stronger updating, local divergence d_t correlates negatively with revision in MULTI-C ($r = -0.935$), despite contributing strongly to the model (0.399). Higher divergence is therefore associated with lower revision rates rather than stronger updating. In contrast, incoming uncertainty $H(\hat{e}_t)$ shows the strongest positive association with revision ($r = +0.562$, contribution 0.103), suggesting that revision is influenced by ambiguity in new evidence rather than local mismatch alone. More broadly, the learned controller

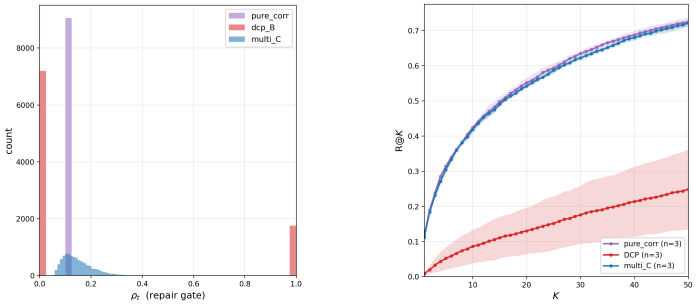

Figure 2: *Left*: Learned revision policies. PURE-CORR remains low and deterministic, DCP-B collapses into near-binary switching, and MULTI-C maintains bounded revision. *Right*: R@K retrieval performance. MULTI-C closely matches PURE-CORR, while DCP-B consistently underperforms.

Signal	Contrib.	Corr
d_t	0.399	-0.935
a_{t-1}	0.034	$+0.040$
$\Delta\hat{a}_t$	0.025	-0.008
$H(e_{t-1})$	0.054	-0.176
$H(\hat{e}_t)$	0.103	$+0.562$

Table 2: Signal contributions to MULTI-C revision and correlation with ρ_t (mean across seeds). Contribution is computed as scaled importance, $|W_i| \cdot \mathrm{std}(x_i)$.

distributes revision across multiple conversational signals rather than relying primarily on divergence.

Different policies produce distinct revision behavior. Figure 2 (Left) shows that different competing policies produce distinct revision behavior. By design, PURE-CORR maintains fixed low revision, DCP-B collapses into near-binary switching driven by local mismatch, and MULTI-C learns smooth bounded revision across turns. These differences reveal fundamentally different revision assumptions: stable accumulation (PURE-CORR), surprise-driven switching (DCP-B), and uncertainty-sensitive revision (MULTI-C).

Trajectories confirm the dissociation. Figure 3 presents two randomly sampled dialogues truncated before image reveal. The first involves a personal disclosure (*"t5: She got divorced"*); the second establishes a volleyball-related reference early and revisits it through later conversational details before photo sharing.

In both dialogues, grounding assumptions produce distinct behaviors. By design, PURE-CORR and FIXED maintain nearly constant revision and flat coherence trajectories, reflecting stable accumulation with limited adaptation. DCP-B instead exhibits a spike-then-freeze pattern: ρ_t rises sharply at the first divergent turn and quickly collapses toward 0. Coherence drops and fails to recover despite continued evidence, suggesting that local mismatch triggers abrupt revision followed by disengagement from grounding.

In contrast, MULTI-C remains responsive without collapsing. In both dialogues, coherence temporarily decreases at conversational turning points but recovers as additional evidence accumulates, producing trajectories aligned with the evolving dialogue. Rather than collapsing after local mismatch, revision remains active but bounded, preserving accumulated grounding while adapting to new information. These trajectories provide an interpretable view of grounding pressure and recovery that is not visible from retrieval metrics alone.

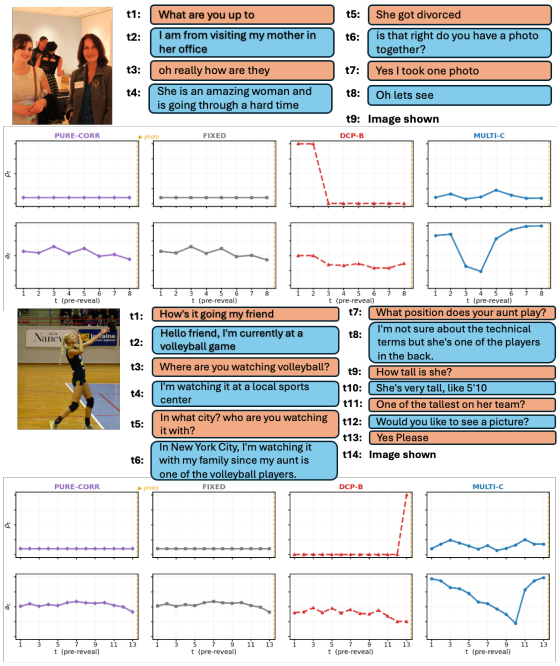

Figure 3: Turn-level grounding trajectories in two pre-reveal dialogues. PURE-CORR and FIXED remain stable, DCP-B exhibits spike–freeze revision, and MULTI-C shows bounded adaptation to conversational change.

Both modalities are necessary. Inference-time perturbation experiments confirm that MULTI-C relies on both visual and linguistic information. Replacing either modality reduces retrieval to near-random levels across metrics (Appendix B). Grounding dynamics therefore cannot emerge from either modality alone.

7 Discussion

Successful grounding is conservative by default. MULTI-C converges to an assumption in which accumulation dominates and revision remains selec-

tive (Mean $\rho_t = 0.148$). Contrary to the common computational intuition that larger mismatch should trigger stronger updating, local divergence tends to reduce revision rather than amplify it. Revision instead appears to be more strongly associated with uncertainty in incoming evidence than with local conflict alone. This behavior resembles how interlocutors maintain common ground, preserving interpretations unless ambiguity becomes sufficiently high to motivate reinterpretation. Notably, the model recovers this pattern without explicit supervision. The result is broadly consistent with psycholinguistic accounts of conceptual pacts (Brennan and Clark, 1996), in which partner-specific interpretations are preserved by default rather than continuously renegotiated.

Repair is not reducible to local mismatch. DCP-B suggests that sensitivity to conversational change alone may be insufficient for successful grounding. Although the model closely tracks local divergence (Rev Sens. = 0.900), grounding becomes unstable once mismatch triggers abrupt revision. This pattern suggests a limitation of relying on local divergence alone as signal: mismatch does not always warrant reinterpretation. Divergence may instead reflect ambiguity, incomplete reference, or temporary conversational repair that still benefits from preserving prior grounding. Detecting mismatch and deciding whether mismatch justifies revision therefore appear to be distinct operations.

Grounding dynamics are not visible in retrieval alone. MULTI-C and PURE-CORR achieve statistically similar retrieval performance while exhibiting qualitatively different revision behavior. Retrieval indicates whether grounding succeeds, but not how it succeeds. Compared with PURE-CORR, MULTI-C exhibits selective preserve-or-revise behavior despite comparable retrieval, suggesting that similar behavioral outcomes can emerge from different grounding mechanisms. The case studies further suggest that these dynamics are behaviorally meaningful: coherence trajectories temporarily destabilize at conversational shifts and recover as evidence accumulates. The mechanisms separating successful grounding from failure therefore provide additional evidence about how conversational grounding unfolds over time. While these findings are consistent with uncertainty-sensitive belief revision, whether human listeners implement revision in the same form remains an open question.

8 Conclusion

We introduced a computational framework for studying conversational grounding as a preserve-or-revise problem. Across revision assumptions, a clear pattern emerged: successful grounding depends less on reacting to local mismatch than on selective revision under uncertainty. Counter to a common computational intuition, mismatch is associated with reduced revision, whereas uncertainty permits accumulated understanding to change. These findings suggest that conversational grounding is better understood as selective belief modulation than purely reactive repair, broadly consistent with conceptual pact theory. The framework provides a computational instrument for studying how understanding evolves through interaction.

9 Limitations

Single dataset. All experiments are conducted on PhotoChat, one of the few large-scale grounded dialogue corpora in which common ground must be constructed through conversation before image disclosure. While additional datasets would strengthen generalizability, few alternatives provide the combination of turn-level dialogue, hidden visual context, and progressive referential grounding required to evaluate preserve-or-revise dynamics. Because PhotoChat interactions are cooperative and goal-oriented, coherent grounding is typically maintained and strong repair events are relatively infrequent, potentially favoring persistence-based strategies. Whether the separation between local mismatch and successful revision generalizes to less stable communicative settings therefore remains open. As a next step, we plan to evaluate these revision policies on I-CONECT dataset (Dodge et al., 2024), a caregiver-guided reminiscence dialogue corpus in which older adults, including individuals with mild cognitive impairment, discuss shared photographs. Compared with PhotoChat, this setting introduces less stable grounding and greater conversational scaffolding, providing a natural stress test for preserve-or-revise dynamics under cognitive and communicative pressure.

Interpretability versus expressivity. Our design intentionally prioritizes interpretability over capacity, making preserve-or-revise dynamics explicit rather than maximizing performance. Whether richer architectures can retain this transparency while capturing finer-grained grounding remains an open question.

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Hiroko H Dodge, Kexin Yu, Chao-Yi Wu, Patrick J Pruitt, Meysam Asgari, Jeffrey A Kaye, Benjamin M Hampstead, Laura Struble, Kathleen Potempa, Peter Lichtenberg, and 1 others. 2024. Internet-based conversational engagement randomized controlled clinical trial (i-conect) among socially isolated adults 75+ years old with normal cognition or mild cognitive impairment: topline results. *The Gerontologist*, 64(4):gnad147.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640.

Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Young-Jun Lee, Dokyong Lee, Junyoung Youn, Kyeong-Jin Oh, Byungsoo Ko, Jonghwan Hyeon, and Ho-Jin Choi. 2024. Stark: Social long-term multi-modal conversation with persona commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12137–12162.

Xiaolong Li and Kristy Boyer. 2016. Reference resolution in situated dialogue with learned semantics. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 329–338.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2:83–111.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the association for computational linguistics: NAACL 2022*, pages 2060–2070.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Alberto Testoni and Raquel Fernández. 2024. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–275.

David Richard Traum. 1994. A computational theory of grounding in natural language conversation.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. Resolving references in visually-grounded dialogue via text generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–469.

George Yule. 2013. *Referential communication tasks*. Routledge.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6142–6152.

A Appendix A: Implementation Details

A.1 Training Protocol

All grounding regimes share an identical frozen CLIP encoder (ViT-B/32) and optimization protocol, trained end-to-end with the following settings:

- Optimizer: AdamW
- Learning rate: 5×10^{-4}
- Batch size: 16 dialogues
- Epochs: 12
- Temperature: $\tau = 0.07$
- Random seeds: $\{0, 1, 2\}$

The CLIP backbone remains frozen throughout. Results are reported as mean $\pm$ std over three seeds. The train–test split is fixed independently of the training seed (split seed $= 42$).

A.2 Loss Hyperparameters

All dynamic regimes (FIXED, DCP-B, MULTI-C) share the following loss weights:

- α (preserve): 0.1
- β (revise): 0.1
- γ_{rate} (sparsity): 10.0
- Target revision rate r: 0.15
- DCP sharpness initialization a: 2.0

PURE-CORR sets $\alpha = \beta = 0$, disabling all grounding dynamics losses.

A.3 Data Split

PhotoChat contains 21 training shards. We construct a deterministic split at the shard level: 19 shards for training, 2 for testing (split seed $= 42$). The split is created once and shared across all training seeds and regime variants.

A.4 Reproducibility

All six regime variants are trained in two parallel waves across three GPUs per wave. Wave 1 trains FIXED, DCP-B, and PURE-CORR in parallel; Wave 2 trains MULTI-C and two ablated variants. Evaluation runs for all six trained models and two untrained baselines (Random, CLIP-ONLY) are executed sequentially per seed. Per-seed results are aggregated into mean $\pm$ std tables using a fixed aggregation script. The full pipeline is idempotent: interrupted runs resume from the last completed checkpoint.

B Appendix B: Modality Ablation

Table 3 reports inference-time perturbation results for MULTI-C. Each modality is replaced independently with zero vectors, random Gaussian vectors, or empirical mean embeddings computed over the test set. No retraining is performed.

Condition	R@1	R@5	R@10	MRR
Full	$0.112 \pm .004$	$0.303 \pm .011$	$0.418 \pm .009$	$0.212 \pm .002$
Img-Zero	$0.000 \pm .000$	$0.000 \pm .000$	$0.000 \pm .000$	$0.002 \pm .000$
Img-Random	$0.001 \pm .000$	$0.004 \pm .003$	$0.010 \pm .004$	$0.008 \pm .001$
Img-Mean	$0.000 \pm .000$	$0.000 \pm .000$	$0.000 \pm .000$	$0.002 \pm .000$
Txt-Zero	$0.000 \pm .000$	$0.000 \pm .000$	$0.000 \pm .000$	$0.002 \pm .000$
Txt-Random	$0.003 \pm .002$	$0.006 \pm .002$	$0.013 \pm .001$	$0.010 \pm .001$
Txt-Mean	$0.001 \pm .000$	$0.006 \pm .000$	$0.011 \pm .001$	$0.008 \pm .000$

Table 3: Inference-time modality ablation for MULTI-C (mean $\pm$ std, three seeds). Replacing either modality collapses retrieval to near-random, confirming grounding depends jointly on both modalities.