REFINE BEFORE YOU REGRESS: PERSISTENT BELIEF REFINEMENT FOR GEOMETRIC REASONING WITH FROZEN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Iterative geometric reasoning raises a basic design question: should the numerical answer also serve as the state for subsequent refinement? We study this question through full-scene projective cuboid prediction from grayscale observations of a single RGB image with frozen vision–language models (VLMs). We introduce *persistent belief refinement*, a training-free inference protocol that maintains persistent object identities and revisable hypotheses about object support, shape, camera and scene relations, and uncertainty, while generating cuboid coordinates only at the final readout. A recurrent-sufficiency analysis motivates this separation by characterizing when a terminal output can replace a richer state without changing subsequent task-level behavior. On 90 held-out SUN RGB-D scenes, the protocol improves coverage-aware Cuboid-F1 AUC by 0.114 under Raw observations and 0.148 under coarse-to-fine observations relative to coordinate regeneration, with both protocols using six logical calls. Both configurations also substantially reduce false positives. On a 20-scene subset, adding coordinate feedback while retaining the belief-state schema and persistence rules yields lower F1 at unchanged aggregate recall under approximately matched token usage. A 17-scene cross-backbone comparison finds higher Cuboid-F1 AUC for the complete belief-refinement protocol in all three tested VLM families. These findings support separating recurrent hypothesis refinement from terminal numerical prediction in frozen-VLM cuboid inference.

1 INTRODUCTION

Vision–language models (VLMs) contain substantial semantic and commonsense knowledge, yet translating it into precise geometry remains difficult. We study this gap through full-scene monocular cuboid prediction with a frozen VLM. Given a single RGB source image, the model uses grayscale observations derived from it to identify major cuboid-like objects and predict eight projected image-space corners for each object. This requires reasoning about viewpoint, occlusion, hidden structure, and object extent without camera intrinsics, depth maps, externally supplied object crops, external detectors, task-specific training, or inference-time ground truth. The target is an image-grounded projective cuboid hypothesis rather than uniquely identifiable metric 3D reconstruction.

A natural iterative strategy repeatedly predicts and revises cuboid coordinates. This raises a design question: *should the final-answer representation also serve as the recurrent reasoning state?* Projected coordinates specify an output but may leave implicit distinctions needed for subsequent revision: a thin side face may reflect shallow object depth or foreshortening, while hidden extent may be supported by image evidence or only by category knowledge. The same current prediction can therefore conceal different spatial hypotheses that respond differently to later observations.

We introduce *persistent belief refinement*, a training-free inference protocol that maintains revisable hypotheses about object support, shape, camera relation, scene relation, and uncertainty. Object identities persist across rounds, removal requires an explicit decision, and numerical cuboid coordinates are generated only at the final readout. We formalize the motivating distinction through *recurrent sufficiency*: replacing a hypothesis state by its output is lossless for a given update rule

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

only if states sharing that output remain indistinguishable at the task-output level under every common admissible subsequent observation. Violations constitute *output-state aliasing*. This criterion motivates separating state refinement from readout; the effectiveness of our concrete protocol is evaluated empirically.

We study this protocol under repeated unblurred grayscale observations (*Raw*) and a deterministic coarse-to-fine schedule. Both use the same unblurred grayscale image for the terminal readout. All observations derive from the same source image and introduce no additional scene information. The schedule changes how existing evidence is presented, while persistent belief refinement and deferred readout do not depend on a particular evidence order.

On 90 held-out SUN RGB-D scenes, belief refinement improves the primary coverage-aware geometry metric, Cuboid-F1 AUC, by 0.114 under Raw observations and 0.148 under coarse-to-fine observations relative to coordinate regeneration with the same number of logical calls. These improvements are accompanied by substantially fewer false positives, while the two belief variants exhibit different precision–recall trade-offs. Targeted controls on 20 scenes examine free-text memory, numerical feedback, and evidence order. An approximately token-matched intervention that retains the belief-state schema and persistence rules but adds intermediate coordinate feedback yields lower F1 at unchanged aggregate recall. A 17-scene cross-backbone comparison likewise favors the belief-refinement protocol over coordinate regeneration in all three tested VLM families. Together, these results support persistent hypothesis refinement with deferred readout as an effective inference protocol in this setting.

Our contributions are threefold:

- We introduce a training-free persistent belief refinement protocol that maintains revisable spatial hypotheses across iterations and generates numerical cuboids only at the final readout.
- We demonstrate improvements in coverage-aware geometry and false-positive control under both Raw and coarse-to-fine observations, complemented by targeted memory, feedback, and evidence-order controls and a cross-backbone evaluation.
- We formalize *recurrent sufficiency* and *output-state aliasing* to characterize when a terminal output can replace a richer recurrent state without changing subsequent task-level behavior.

2 Related Work

Spatial grounding and multimodal reasoning. Recent multimodal large language models increasingly connect language reasoning with localized visual evidence. KOSMOS-2 and Ferret introduce explicit region-level grounding (Peng et al., 2024; You et al., 2024), while SpatialVLM and SpatialRGPT strengthen spatial reasoning through large-scale spatial supervision and 3D-aware grounding (Chen et al., 2024a; Cheng et al., 2024). Related work further improves grounding through vision-centric architectures, frozen-model localization, attention-based grounding, and multimodal agent representations (Tong et al., 2024; Wu et al., 2025; Kang et al., 2025a; Yang et al., 2025b). These methods primarily improve access to spatial information through training, specialized modules, or grounding supervision. We instead study how an *already frozen* VLM should represent and propagate spatial hypotheses during iterative inference.

3D understanding and spatial mental models. A broad line of work equips language models with explicit 3D understanding using point clouds, multi-view observations, depth, or large-scale 3D-language supervision (Hong et al., 2023; Chen et al., 2024b; Huang et al., 2024; Cho et al., 2025; Zheng et al., 2025; Yang et al., 2025a; Daxberger et al., 2025; Ning et al., 2025; Kang et al., 2025b). Meanwhile, recent benchmarks reveal persistent limitations in viewpoint reasoning, occlusion, hidden-scene completion, and 3D spatial understanding (Zhang et al., 2025b; Pothiraj et al., 2025; Ma et al., 2025; Lee et al., 2025; Ma et al., 2026). Thinking in Space, MindCube, and Theory of Space further emphasize the value of constructing and maintaining internal spatial representations when relevant geometry is only partially observed (Yang et al., 2025c; Wang et al., 2026; Zhang et al., 2026). In contrast, we use projective cuboid prediction from a single RGB image with an unchanged VLM as an observable testbed for how spatial hypotheses are maintained and revised.

Iterative inference and refinement. Iterative reasoning improves model outputs through self-feedback, visual search, reflective retrieval, generative feedback, and task-level interaction (Madaan et al., 2023; Wu & Xie, 2024; Cocchi et al., 2025; Zhang et al., 2025a; Zhai et al., 2024). These approaches largely focus on what is revised—an answer, feedback signal, retrieved evidence, or action trajectory. Our focus is instead on *what representation is propagated between revisions*. We distinguish the terminal output from the recurrent reasoning state and ask when an output representation is sufficiently expressive for continued inference. Rather than repeatedly feeding projected coordinates back into the model, we maintain a revisable spatial hypothesis and defer geometric readout until refinement terminates, preserving distinctions in shape, viewpoint, evidence provenance, and uncertainty that may remain relevant to later updates.

3 METHOD

3.1 PROBLEM FORMULATION AND DESIGN PRINCIPLE

We study full-scene monocular cuboid prediction with a frozen vision–language model (VLM). A single RGB image $I \in \mathbb{R}^{H\times W\times 3}$ is the source; inference uses grayscale observations derived from $\bar{I} = \mathrm{Gray}(I)$. The model identifies major cuboid-like objects and predicts, for each object i, eight projected image-space corners:

$$B_i = [\mathbf{p}_{i,0}, \ldots, \mathbf{p}_{i,7}], \qquad \mathbf{p}_{i,k} \in [0, 1000]^2. \tag{1}$$

Corners 0–3 and 4–7 form corresponding cyclic faces, with cross-face edges $(0, 4)$, $(1, 5)$, $(2, 6)$, and $(3, 7)$. No camera intrinsics, depth map, externally supplied object crop, detector output, or inference-time ground truth is provided. The target is an image-grounded *projective cuboid hypothesis*, not uniquely identifiable metric 3D reconstruction.

Our protocol separates maintaining revisable spatial hypotheses from producing the final numerical geometry. A coordinate tuple specifies a prediction but need not explicitly retain its supporting evidence, alternative interpretations, or unresolved uncertainty. We propagate a structured hypothesis and defer cuboid generation to the terminal call:

$$(O_0, \ldots, O_T) \xrightarrow{\text{hypothesis refinement}} \mathcal{M}_{0:T} \xrightarrow{\text{readout with } I^{(T)}} \mathcal{B}, \qquad \mathcal{B} = \{B_i\}_{i=1}^{N}. \tag{2}$$

Here O_t is the observation supplied at refinement step t and $\mathcal{M}_t$ is the persistent state. The depicted C2F configuration uses $O_t = I^{(t)}$, ending at $I^{(T)} = \bar{I}$; Raw repeats this unblurred endpoint. Sec. 3.4 specifies both schedules. Projected cuboid coordinates remain the task output but do not enter the explicit recurrent state.

3.2 PERSISTENT REVISABLE SPATIAL BELIEF

The state contains scene-level structure and persistent object hypotheses:

$$\mathcal{M}_t = \left(\mathcal{S}_t, \{\mathbf{m}_{i,t}\}_{i=1}^{N_t}\right), \qquad \mathbf{m}_{i,t} = (s_{i,t}, q_{i,t}, c_{i,t}, r_{i,t}, u_{i,t}), \tag{3}$$

where $\mathcal{S}_t$ describes the scene and each object record contains image support s, object-centric shape q, camera relation c, scene relation r, and uncertainty/evidence provenance u. Here *belief* denotes a structured, revisable model-generated hypothesis, not a calibrated probability distribution. Image support records the hypothesized object region and bounds (the "crop & bounds" component in Fig. 1); it is part of the inferred state, not an externally supplied object annotation.

The shape component couples front, side, and top descriptions through shared semantic width, depth, and height. A depth revision must remain consistent across side and top descriptions. Separating shape from camera relation allows a thin projected side face to be attributed to shallow depth or foreshortening without immediately committing to cuboid coordinates. The state also distinguishes image-supported evidence from prior-supported completion and permits unresolved alternatives. The figure's "RGB check" denotes evidence attributed to the source image through its grayscale observations, not a separate color-image call. Appendix A.2 specifies the factorization.

The first observation initializes the state:

$$\mathcal{M}_0 = \mathcal{N}\left(F_\theta^{\text{init}}(O_0, P_{\text{init}})\right), \tag{4}$$

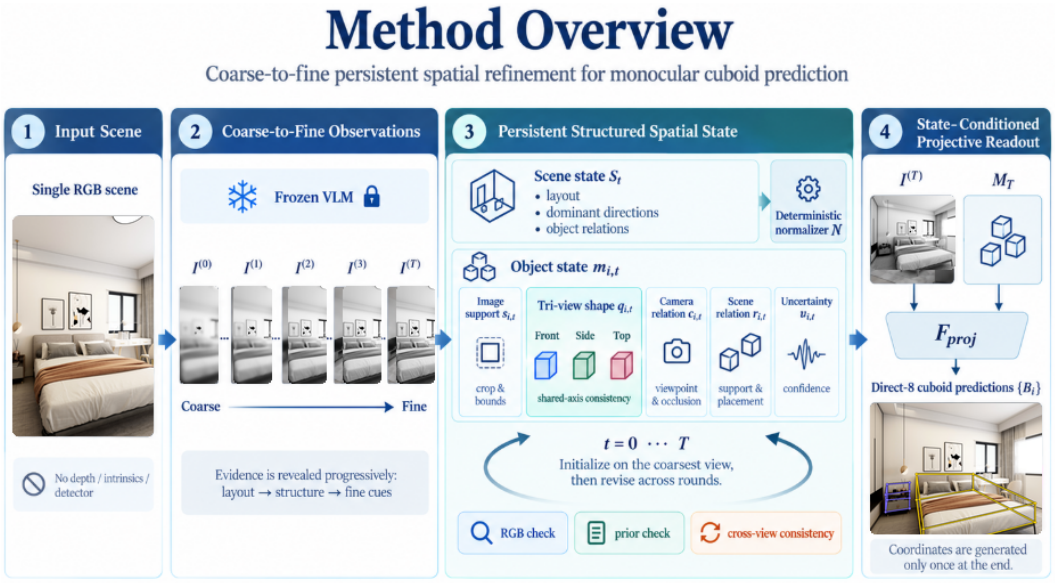


Figure 1: **Persistent belief refinement with deferred geometric readout (C2F configuration).** A single RGB source image is converted into grayscale observations $I^{(0)}, \ldots, I^{(T)}$. A frozen VLM revises a persistent spatial hypothesis, while a deterministic normalizer maintains object identities and state consistency. The terminal call uses the final state $\mathcal{M}_T$ and the unblurred grayscale endpoint $I^{(T)}$. Color images show the source scene and the output overlay for visualization. Raw uses the same state and readout protocol with the unblurred endpoint repeated throughout refinement.

where F_θ is the frozen VLM and $\mathcal{N}$ is a deterministic normalizer. Subsequent calls revise the previous state:

$$\mathcal{M}_t = \mathcal{N}\left(F_\theta^{\text{update}}(O_t, \mathcal{M}_{t-1}, P_{\text{update}}), \mathcal{M}_{t-1}\right), \qquad t = 1, \ldots, T. \tag{5}$$

All roles use the same frozen model with fixed role-specific instructions. Image-supported evidence may revise prior-supported interpretations, while fields without contradictory evidence persist. If a previously active object is omitted, its previous state is retained and marked uncertain; removal requires an explicit false-positive decision from the VLM. The normalizer canonicalizes fields, preserves identities, and enforces shared semantic constraints, without visual recognition or geometric optimization. These rules are part of the evaluated protocol; operational details are provided in Appendix A.3.

3.3 Deferred Readout and Recurrent Sufficiency

After refinement, one terminal call generates numerical cuboids:

$$\mathcal{B} = F_\theta^{\text{proj}}\left(I^{(T)}, \mathcal{M}_T, P_{\text{proj}}\right), \qquad I^{(T)} = \bar{I}. \tag{6}$$

The unblurred grayscale endpoint supplies localization evidence, while $\mathcal{M}_T$ supplies the accumulated object, shape, viewpoint, and uncertainty hypotheses. The trajectory is

$$\mathcal{M}_0 \to \cdots \to \mathcal{M}_T \to \mathcal{B},$$

rather than repeated cuboid regeneration. One initialization, T updates, and one readout require $T + 2$ logical calls; the six-call setting uses $T = 4$.

For a representation-level analysis, fix the terminal observation $J = I^{(T)}$ and consider deterministic readout and update operators. Let

$$\mathcal{R}_J : \mathsf{M} \to \mathsf{B} \tag{7}$$

map hypothesis states to cuboid outputs, and let

$$\mathcal{M}_{t+1} = \Phi_t(\mathcal{M}_t, O_{t+1}) \tag{8}$$

denote a recurrent update. We call $\mathcal{R}_J$ *recurrently sufficient* for Φ_t if there exists an answer-space transition G_t satisfying

$$\mathcal{R}_J\left(\Phi_t(\mathcal{M}, O)\right) = G_t\left(\mathcal{R}_J(\mathcal{M}), O\right) \quad \text{for all admissible } \mathcal{M} \text{ and } O. \tag{9}$$

This condition asks whether the current output determines the next task output under the specified update rule and common observation. If it holds at every step, answer-space transitions can reproduce the corresponding output trajectory without reconstructing the richer state.

The condition fails under *output-state aliasing*: two states share the current readout,

$$\mathcal{R}_J(\mathcal{M}^{(1)}) = \mathcal{R}_J(\mathcal{M}^{(2)}), \tag{10}$$

but yield different readouts after a common observation O:

$$\mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(1)}, O)\right) \neq \mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(2)}, O)\right). \tag{11}$$

No deterministic transition receiving only the shared readout and O can reproduce both continuations. Appendix B.1 gives the characterization and proof. The analysis motivates state–output separation, without proving our schema sufficient or guaranteeing improved predictions. Adding coordinate feedback while retaining the belief state is a different operation from state replacement; the feedback control evaluates that additional conditioning empirically.

3.4 Complementary Evidence Ordering

C2F constructs a fixed sequence from the grayscale source:

$$I^{(t)} = \mathcal{T}_{\lambda_t}(I), \qquad \lambda_0 \succ \lambda_1 \succ \cdots \succ \lambda_T, \tag{12}$$

with the smoothing-based instantiation

$$I^{(t)} = G_{\sigma_t}(\bar{I}), \qquad \bar{I} = \mathrm{Gray}(I), \qquad \sigma_0 > \sigma_1 > \cdots > \sigma_T = 0. \tag{13}$$

The endpoint $I^{(T)} = \bar{I}$ is therefore unblurred grayscale. C2F presents $O_t = I^{(t)}$, whereas Raw presents $O_t = \bar{I}$ at every step; both read out from the same endpoint $I^{(T)}$. The F2C control reverses only the refinement observations, $O_t = I^{(T-t)}$, retaining this readout input. Here superscripts label C2F levels, not presentation positions in F2C. Appendix A.1 lists the call allocation.

Smoothing attenuates high-frequency appearance variation; reducing its strength exposes finer boundaries in later observations. The schedule changes how existing evidence is presented, without supplying another view or a new scene measurement. The complete sequence is a deterministic function of the source image and introduces no additional scene information (Appendix B.3). Persistent state refinement and deferred readout apply under either schedule; C2F is a complementary configuration rather than a required component. The procedure uses only the frozen VLM and deterministic image/state transformations, with no external detector, depth estimator, numerical 3D optimizer, task-specific training, or inference-time ground truth.

4 Experiments

We evaluate persistent belief refinement as an inference protocol for projective cuboid prediction. Three questions guide the experiments. First, does the protocol improve coverage-aware geometry and false-positive control relative to repeated coordinate regeneration? Second, how do targeted changes to recurrent memory, numerical feedback, and observation order affect its behavior? Third, does the advantage of the complete protocol appear across frozen VLM backbones? The primary comparison establishes protocol-level performance; the smaller controls examine specific design choices rather than identifying every component's contribution.

4.1 Experimental Setup

Data and inference protocol. The evaluation starts from 100 SUN RGB-D validation images with Omni3D 3D annotations. We exclude the 10 scenes used during pilot development and evaluate on the remaining 90 held-out scenes, containing 213 eligible ground-truth (GT) objects. Eligible

Table 1: Primary results on 90 held-out SUN RGB-D scenes (213 eligible GT objects). All iterative conditions use six logical calls; token usage is reported in Sec. 4.1. GI-NCE is matched-only, C-F1 AUC combines geometry with precision and recall, and C-P AUC evaluates precision over all predictions. Bold denotes the best reported value in each metric column.

Method	P↑	R↑	F1↑	FPPI↓	GI-NCE↓	C-F1 AUC↑	C-P AUC↑	OS-P↑
DIRECT-1	43.4	75.6	55.1	2.33	0.154	0.288	0.227	0.339
COORD-RAW$_6$	37.9	**80.3**	51.5	3.11	0.146	0.284	0.209	0.334
COORD-C2F	38.2	60.6	46.8	2.32	0.160	0.238	0.194	0.356
BELIEF-RAW$_6$	65.7	72.8	**69.0**	0.90	**0.139**	**0.398**	0.378	0.616
BELIEF-C2F	**70.1**	66.2	68.1	**0.67**	0.143	0.386	**0.398**	**0.659**

objects belong to one of ten cuboid-like indoor categories, have valid 3D annotations and positive depth, have at most 50% known truncation, and occupy at least 1% of the image. GT geometry and camera calibration are used only by the evaluator and are never exposed to the VLM.

Gemini 3.1 Pro Preview is the primary frozen backbone. Unless otherwise stated, iterative conditions use the same input resolution, cuboid convention, per-call output cap, temperature 0, and six logical calls. Targeted controls use a fixed 20-scene subset containing 41 GT objects. Cross-backbone comparisons use the same fixed 17 scenes (55 GT objects) for Gemini, GPT-5.6-Sol, and Qwen3.8-Flash. These subsets were fixed without consulting GT-dependent method scores.

The main experiment follows a 2×2 design over the recurrent protocol and observation schedule. COORD re-emits a complete object set and projected cuboid coordinates at every round. BELIEF uses one state initialization, four state updates, and a terminal numerical readout in the sixth call. Its schema, persistence rules, normalization, and deferred readout constitute the evaluated protocol. All conditions use grayscale image inputs. RAW repeats the unblurred image $\bar{I}$ at every call. For C2F, the first five calls use $I^{(0)}, \ldots, I^{(4)}$ and the sixth repeats $I^{(4)} = \bar{I}$. DIRECT-1 predicts from $\bar{I}$ in one call and is a reference rather than a factorial condition. Appendix A.1 summarizes the image inputs. Logical calls do not imply equal realized token usage: COORD-RAW$_6$, COORD-C2F, BELIEF-RAW$_6$, and BELIEF-C2F use 47.8k, 41.3k, 66.6k, and 60.9k tokens per scene, respectively. The main comparison therefore evaluates complete protocols under a shared call budget, not strict compute equivalence.

Metrics. Predictions are assigned to eligible GT objects by class-aware Hungarian matching of projected envelopes at IoU ≥ 0.25. We report precision (P), recall (R), F1, and false positives per image (FPPI). Tables express P, R, and F1 in percent; their reported differences below use the $[0, 1]$ scale unless stated otherwise. For matched pairs, graph-invariant normalized corner error is

$$\text{GI-NCE}(\hat{P}, P^\star) = \min_{\pi \in \text{Aut}(Q_3)} \frac{1}{8d^\star} \sum_{k=0}^{7} \left\| \hat{p}_k - p^\star_{\pi(k)} \right\|_2, \tag{14}$$

where $d^\star$ is the diagonal of the GT projected envelope and $\text{Aut}(Q_3)$ contains the 48 cube-graph automorphisms.

Because matched-only error does not account for coverage, our primary geometry metric is Cuboid-F1 AUC (C-F1 AUC). For each GI-NCE threshold $\tau \in [0.02, 0.30]$, a prediction is correct only if it is matched and has GI-NCE $\leq \tau$; we report the normalized area under the resulting F1 curve. Cuboid-Precision AUC (C-P AUC) is defined analogously with all predictions in the denominator. Orientation-similarity precision is

$$\text{OS-P} = \frac{1}{N_{\text{pred}}} \sum_{i \in \mathcal{M}} \frac{1 + \cos \theta_i}{2}, \tag{15}$$

where $\mathcal{M}$ contains the matched pairs, θ_i is the symmetry-aware cuboid-axis error, and unmatched predictions contribute zero. OS-P therefore reflects both match precision and orientation agreement, rather than orientation accuracy alone. We use *reliability* to refer to false-positive control and coverage-aware geometry under this evaluator, not calibrated confidence. Reported 95% confidence intervals use 10,000 paired scene-level bootstrap resamples.

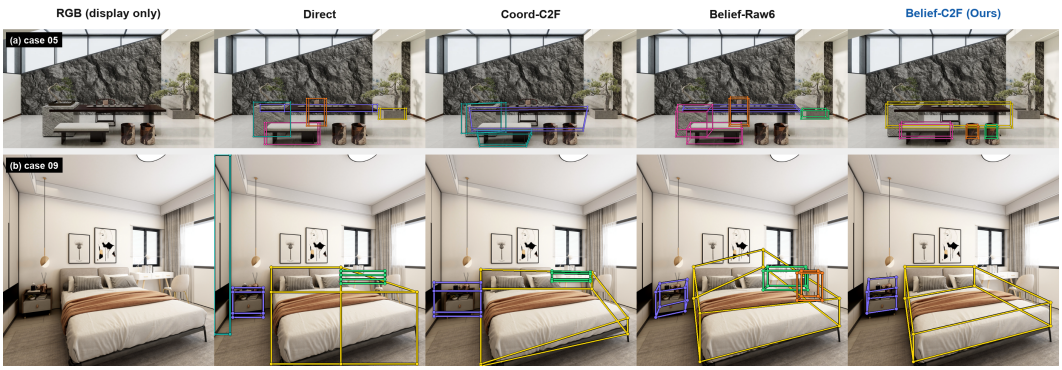


Figure 2: **Qualitative comparison on selected indoor scenes.** The RGB column and color backgrounds are for visualization only; model inputs follow the grayscale protocol in Sec. 4.1. The remaining columns show final projected cuboids from the respective strategies, illustrating differences in object multiplicity, overlap, and projected extent. The figure labels DIRECT and BELIEF-RAW correspond to DIRECT-1 and BELIEF-RAW$_6$. No GT cuboids are overlaid, so these panels illustrate prediction structure rather than independently establish geometric accuracy. Additional examples are provided in Appendix Fig. 3.

4.2 PERSISTENT BELIEF REFINEMENT IMPROVES RELIABILITY

Table 1 shows that additional coordinate-generation calls do not necessarily improve prediction-set quality. Relative to DIRECT-1, COORD-RAW$_6$ increases recall from 75.6% to 80.3%, but also increases FPPI from 2.33 to 3.11 and reduces F1 from 55.1% to 51.5%. COORD-C2F likewise reaches only 46.8% F1 and 0.238 C-F1 AUC. Repeatedly generating the answer is therefore not, by itself, an effective refinement strategy in these comparisons.

Both belief configurations improve the primary coverage-aware metric over their corresponding coordinate protocols. C-F1 AUC increases by $+0.114$ under Raw observations (95% CI $[0.065, 0.160]$) and by $+0.148$ under C2F ($[0.101, 0.193]$). F1 increases by $+0.175$ ($[0.090, 0.257]$) and $+0.213$ ($[0.139, 0.286]$), while FPPI decreases by 2.211 ($[1.644, 2.867]$) and 1.656 ($[1.244, 2.100]$), respectively. Among the evaluated conditions, BELIEF-RAW$_6$ has the best F1 (69.0%), matched GI-NCE (0.139), and C-F1 AUC (0.398). BELIEF-C2F is more selective, with the best precision (70.1%), FPPI (0.67), C-P AUC (0.398), and OS-P (0.659). These are two observation configurations of the same belief-refinement protocol, with different precision–recall trade-offs.

Figure 2 illustrates differences in the final hypothesis sets; the quantitative trajectories provide a separate diagnostic. For COORD-RAW$_6$, false positives increase from 210 at round 0 to 280 at round 5 even as matched GI-NCE decreases from 0.154 to 0.146. For COORD-C2F, false positives increase from 113 to 209. Thus, lower error on the current matched subset can coexist with poorer prediction-set reliability; it does not imply that the same objects improve throughout refinement.

On the 97 GT objects matched by all five primary methods, mean GI-NCE is 0.130 for BELIEF-RAW$_6$, 0.139 for BELIEF-C2F, 0.143 for COORD-RAW$_6$, 0.152 for DIRECT-1, and 0.154 for COORD-C2F. Both belief configurations therefore also improve mean geometry on common support, so the results are not solely an artifact of different matched sets. Together with the FPPI reductions, this supports evaluating hypothesis management alongside geometric accuracy, without implying a uniform improvement for every object.

Within the belief protocol, switching from Raw to C2F decreases FPPI from 0.90 to 0.67 and increases OS-P from 0.616 to 0.659, but reduces recall from 72.8% to 66.2%. The paired F1 difference is -0.009 (95% CI $[-0.054, 0.037]$). The observed effect is therefore a precision–recall trade-off, not consistent superiority of C2F across metrics. Importantly, the improvement over coordinate regeneration is present under both observation schedules.

Table 2: Targeted controls on the same 20 scenes (41 GT objects). BELIEF+COORD-C2F shares the belief-state schema and persistence rules with BELIEF-C2F, but adds intermediate cuboid proposals to subsequent conditioning; their token usage is approximately matched. Bold denotes the best reported value in each metric column, including ties.

Method	Pred.	Match	P↑	R↑	F1↑	FPPI↓	C-F1 AUC↑	C-P AUC↑	OS-P↑
DIRECT-1	77	31	40.3	75.6	52.5	2.30	0.280	0.215	0.360
COORD-RAW$_6$	101	32	31.7	78.0	45.1	3.45	0.251	0.177	0.305
FREE-TEXT	107	35	32.7	**85.4**	47.3	3.60	0.272	0.188	0.244
BELIEF-RAW$_6$	53	35	66.0	**85.4**	**74.5**	0.90	**0.402**	0.356	0.612
BELIEF-F2C	53	34	64.2	82.9	72.3	0.95	0.392	0.348	0.562
BELIEF+COORD-C2F	55	30	54.5	73.2	62.5	1.25	0.330	0.288	0.485
BELIEF-C2F	45	30	**66.7**	73.2	69.8	**0.75**	0.401	**0.383**	**0.635**

4.3 TARGETED CONTROLS ON THE INFERENCE PROTOCOL

We next examine recurrent memory, intermediate coordinate feedback, and evidence order on the fixed 20-scene subset. Each comparison changes a specified part of the protocol; the controls do not constitute a field-by-field decomposition of the main effect.

Coordinate feedback within the belief protocol. BELIEF+COORD-C2F retains the belief-state schema, object-identity and lifecycle rules, uncertainty fields, C2F observations, and final output convention of BELIEF-C2F. It additionally conditions intermediate updates on numerical cuboid proposals. The two conditions use approximately matched tokens per scene (61.2k for BELIEF+COORD-C2F versus 62.2k for BELIEF-C2F). Both yield 30 matches and 73.2% aggregate recall; this equality concerns counts, not necessarily the identities of matched GT objects. Adding coordinate feedback increases predictions from 45 to 55, reduces precision from 66.7% to 54.5%, and increases FPPI from 0.75 to 1.25. F1 decreases from 69.8% to 62.5%, C-F1 AUC from 0.401 to 0.330, and OS-P from 0.635 to 0.485. These point estimates support sensitivity to explicit numerical feedback in this protocol. Because the belief state remains available, this intervention concerns additional conditioning, not the state-replacement operation characterized by recurrent sufficiency.

Unrestricted textual memory. FREE-TEXT uses five Raw reasoning calls followed by one final readout, but propagates an unrestricted textual scratchpad instead of the persistent structured state. It yields 35 matches, as does BELIEF-RAW$_6$, while emitting 107 predictions rather than 53. Relative to FREE-TEXT, BELIEF-RAW$_6$ improves precision by $+0.333$ (95% CI $[0.194, 0.458]$), F1 by $+0.272$ ($[0.138, 0.388]$), and C-F1 AUC by $+0.130$ ($[0.063, 0.184]$), while the FPPI difference is -2.70 ($[-4.15, -1.45]$). Thus, deferring readout while propagating the evaluated unrestricted scratchpad does not reproduce the gains of the structured protocol. This comparison concerns the implemented memory designs, not the impossibility of effective free-text recurrence or the necessity of every state field.

Reversing evidence order. BELIEF-F2C receives the same five transformed observations as BELIEF-C2F, but in reverse order during belief refinement. Both terminal readouts use the same unblurred grayscale image (Appendix A.1). C2F emits fewer predictions (45 vs. 53), increases C-P AUC from 0.348 to 0.383 and OS-P from 0.562 to 0.635, but reduces recall from 82.9% to 73.2%. The paired OS-P improvement is $+0.073$ (95% CI $[0.004, 0.153]$), whereas the intervals for precision, F1, FPPI, GI-NCE, and the AUC metrics cross zero. These results are consistent with an operating-point change, not a uniform advantage of C2F. Together with the Raw results, they support treating evidence ordering as a complementary configuration rather than a prerequisite for the protocol's benefit.

State logs provide additional diagnostics. In the primary evaluation, every object active in the final belief is emitted by the numerical readout (236/236 for Raw and 201/201 for C2F). The terminal readout therefore does not obtain its selectivity by omitting active objects in these runs. Unresolved hypotheses match GT less often than resolved ones (58.2% vs. 77.8% under Raw; 67.1% vs. 83.8% under C2F). Hypotheses tagged as RGB-dominant also match GT more often than those tagged as prior-dominant (72.6% vs. 31.6%; 74.8% vs. 20.0%). "RGB-dominant" is the schema label

Table 3: Cross-backbone comparison on the same 17 scenes (55 GT objects). Both methods receive identical C2F observations and use six logical calls. Bold denotes the better reported value within each backbone.

Backbone	Method	P↑	R↑	F1↑	FPPI↓	GI-NCE↓	C-F1 AUC↑
Gemini	COORD-C2F	60.0	60.0	60.0	1.29	0.160	0.300
	BELIEF-C2F	**76.1**	**63.6**	**69.3**	**0.65**	**0.125**	**0.432**
GPT	COORD-C2F	50.0	50.9	50.5	1.65	0.230	0.146
	BELIEF-C2F	**70.4**	**69.1**	**69.7**	**0.94**	**0.179**	**0.324**
Qwen	COORD-C2F	46.4	47.3	46.8	1.76	0.202	0.176
	BELIEF-C2F	**74.4**	**52.7**	**61.7**	**0.59**	**0.155**	**0.327**

for image-supported evidence, including evidence from grayscale observations, rather than a color-input condition. These associations are diagnostics of model-reported uncertainty and provenance, not calibrated confidence estimates or causal tests of individual fields.

4.4 CROSS-BACKBONE EVALUATION

Table 3 compares the complete belief and coordinate C2F protocols across three frozen backbones. C-F1 AUC improves by 0.132 for Gemini, 0.178 for GPT, and 0.151 for Qwen. F1 increases from 60.0% to 69.3%, 50.5% to 69.7%, and 46.8% to 61.7%, respectively. On GT objects matched by both conditions within each backbone, GI-NCE decreases by 0.038 for GPT (95% CI $[0.016, 0.063]$) and by 0.028 for Qwen ($[0.010, 0.048]$). This common-support analysis is distinct from the matched-only averages in the table. Although limited to 17 scenes, the comparison shows a consistent direction of benefit beyond the primary model. It evaluates cross-backbone consistency of the full protocol, not replication of the coordinate-feedback intervention in every backbone.

Taken together, the primary results establish improved coverage-aware performance under both observation schedules in the evaluated setting. The targeted controls provide additional evidence about the implemented memory and feedback choices, while evidence ordering changes the observed precision–recall trade-off. The evidence supports the concrete persistent-refinement protocol, without requiring that numerical feedback explain every performance difference.

5 DISCUSSION AND LIMITATIONS

Persistent belief refinement with deferred readout improves coverage-aware cuboid prediction and false-positive control in the studied frozen-VLM setting. The benefit appears under both Raw and C2F observations, while the two configurations favor different precision–recall trade-offs. The targeted controls further show that the evaluated free-text scratchpad does not reproduce the structured protocol's gains, and that adding numerical feedback yields lower observed F1 and C-F1 AUC despite similar token usage and unchanged aggregate recall. These findings support treating state maintenance and readout timing as explicit inference-design choices. They evaluate concrete protocols, however, rather than isolating every contribution of the schema, object persistence, normalization, and readout. The particular factorization is an effective realization, but its optimality is not established; C2F is a complementary configuration rather than a necessary component.

The empirical scope is projective cuboid prediction on indoor scenes using grayscale observations derived from a single RGB source, not metric 3D reconstruction or general multimodal reasoning. Color-enabled inference is outside the evaluated protocol. The 20-scene controls and 17-scene cross-backbone comparison provide supporting evidence on smaller samples than the 90-scene primary evaluation. The main comparison matches logical calls rather than tokens; approximate token matching in the feedback control constrains that particular comparison but does not establish a general equal-compute advantage. Likewise, Appendix B.1 characterizes lossless replacement of a richer recurrent state by its output, rather than establishing output-state aliasing as the empirical cause of the observed gaps or guaranteeing convergence. Within these boundaries, the results support a practical design principle: maintain a revisable spatial hypothesis during refinement, and generate the numerical answer when that refinement is complete.

REPRODUCIBILITY STATEMENT

We provide the task definition, evaluation protocol, matching procedure, metrics, and inference settings in Sec. 4.1. The recurrent state representation, update and normalization rules, theoretical statements and proofs, and additional experimental details are provided in the appendix. The supplementary implementation materials contain the prompts, state schemas, evaluation code, and configuration details for the reported experiments. The distinction between the RGB source, grayscale model observations, and color visualization backgrounds is specified in Appendix A.1. All reported comparisons use the frozen VLMs and deterministic processing described in the paper, with ground-truth annotations used only by the evaluator.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and revision, literature discovery, research ideation, conceptual framing, implementation and code editing, and interpretation of experimental results. They also assisted with the formulation and exposition of mathematical statements and proofs; their outputs are not treated as independent verification. The VLMs evaluated in this paper are additionally the inference engines of the proposed and comparison protocols. Reported quantitative results are measurements from executed experiments, not illustrative values generated for the manuscript. The authors take responsibility for the research decisions, mathematical arguments, citations, implementations, reported results, and final text.

REFERENCES

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual interactive instruction tuning for omni-3D understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, 2024.

Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, and Marco Pavone. Language-image models with 3D understanding. In *International Conference on Learning Representations*, 2025.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal LLMs with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems*, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3D: Improving 3D large language model via robust instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Phillip Y. Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. CAPTURE: Evaluating spatial reasoning in vision language models via occluded object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Spatial mental modeling from limited views. In *International Conference on Learning Representations*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F. Fouhey, and Joyce Chai. 3D-GRAND: A million-scale dataset for 3D-LLMs with better grounding and less hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, and Jianfeng Gao. Magma: A foundation model for multimodal AI agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *International Conference on Learning Representations*, 2025a.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations*, 2026.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *International Conference on Learning Representations*, 2025b.

Henry Zheng, Hao Shi, Qihang Peng, Yong Xien Chng, Rui Huang, Yepeng Weng, Zhongchao Shi, and Gao Huang. Densegrounding: Improving dense language-vision semantics for ego-centric 3D visual grounding. In *International Conference on Learning Representations*, 2025.

A Additional Method Details

This appendix specifies the recurrent representation and update rules used in our persistent belief-refinement protocol. The representation makes spatial hypotheses available for revision before a terminal numerical readout; its effectiveness is evaluated together with the object-persistence and normalization rules below. We first make the image and call conventions explicit, then describe the state factorization and persistence rules.

A.1 Image Inputs and Call Allocation

We distinguish the source image I, the unblurred grayscale image $\bar{I} = \mathrm{Gray}(I)$, and the scheduled observations $I^{(t)} = G_{\sigma_t}(\bar{I})$. "Raw" denotes $\bar{I}$ without additional smoothing; it does not denote an RGB input. For the six-call conditions, the five C2F levels end at $I^{(4)} = \bar{I}$. The observed inputs $O_0, \ldots, O_4$ depend on presentation order, while the sixth-call image remains $\bar{I}$. Table 4 makes this distinction explicit.

Direct-1 uses one unblurred grayscale input. Free-Text follows the Raw schedule, and Belief+Coord-C2F follows the C2F schedule with additional numerical conditioning. The final image in F2C is the fixed unblurred readout input, not the last (coarsest) refinement observation. RGB backgrounds in the qualitative figures are used to display the scene and predicted cuboids; their colors are not additional model evidence.

Table 4: Image inputs in the evaluated schedules. C2F-level superscripts retain their meaning under order reversal. For belief conditions, calls 1–5 initialize/update the state and call 6 performs the numerical readout. Coordinate conditions use Raw or C2F inputs and generate cuboids at every call; F2C is a belief control. No terminal call adds an RGB observation.

Schedule	Call 1	Call 2	Call 3	Call 4	Call 5	Call 6
Raw	$\bar{I}$	$\bar{I}$	$\bar{I}$	$\bar{I}$	$\bar{I}$	$\bar{I}$
C2F	$I^{(0)}$	$I^{(1)}$	$I^{(2)}$	$I^{(3)}$	$I^{(4)}$	$I^{(4)}$
F2C	$I^{(4)}$	$I^{(3)}$	$I^{(2)}$	$I^{(1)}$	$I^{(0)}$	$I^{(4)}$

A.2 Spatial-Belief Parameterization

The persistent spatial state is

$$\mathcal{M}_t = \left(\mathcal{S}_t, \{\mathbf{m}_{i,t}\}_{i=1}^{N_t}\right), \qquad \mathbf{m}_{i,t} = (s_{i,t}, q_{i,t}, c_{i,t}, r_{i,t}, u_{i,t}), \tag{16}$$

where $\mathcal{S}_t$ denotes scene-level structure and each object record contains image support s, object-centric shape q, camera relation c, scene relation r, and uncertainty/evidence provenance u. As in Sec. 3.2, these records are model-generated hypotheses, not calibrated probability distributions.

The object-centric shape component is

$$q_{i,t} = \left(V_{i,t}^{\text{front}}, V_{i,t}^{\text{side}}, V_{i,t}^{\text{top}}, \boldsymbol{\rho}_{i,t}, \pi_{i,t}^{\text{comp}}\right), \tag{17}$$

where $\boldsymbol{\rho}_{i,t}$ describes relative width–depth–height proportions and $\pi_{i,t}^{\text{comp}}$ records the current hypothesis about unobserved geometry. These quantities describe semantic relative extents and completion hypotheses rather than recovered metric dimensions.

The three canonical views share semantic axes:

$$\begin{aligned} V_{i,t}^{\text{front}} &\leftrightarrow (W_{i,t}, H_{i,t}), \\ V_{i,t}^{\text{side}} &\leftrightarrow (D_{i,t}, H_{i,t}), \\ V_{i,t}^{\text{top}} &\leftrightarrow (W_{i,t}, D_{i,t}). \end{aligned} \tag{18}$$

Revising one semantic extent must remain consistent across the views that share it: changing $D_{i,t}$ affects both the side and top descriptions. This coupling distinguishes object-centric shape from image-space appearance. For example, a narrow projected side face may be attributed to shallow depth or to viewpoint-induced foreshortening, without immediately assigning cuboid corner coordinates.

The camera-dependent component is

$$c_{i,t} = \left(\mathcal{F}_{i,t}^{\text{vis}}, v_{i,t}, b_{i,t}, \ell_{i,t}, \delta_{i,t}, o_{i,t}\right), \tag{19}$$

where $\mathcal{F}_{i,t}^{\text{vis}}$ denotes visible semantic faces, $v_{i,t}$ the coarse viewpoint, $b_{i,t}$ the horizontal orientation branch, $\ell_{i,t}$ the dominant screen-space axis, $\delta_{i,t}$ projected-depth visibility, and $o_{i,t}$ the occlusion state. These fields describe how the current shape hypothesis is viewed and projected, rather than redefining its object-centric extents.

Evidence provenance and unresolved uncertainty are represented by

$$u_{i,t} = \left(\mathcal{E}_{i,t}^{\text{RGB}}, \mathcal{E}_{i,t}^{\text{prior}}, \mathcal{U}_{i,t}, f_{i,t}^{\star}\right), \tag{20}$$

where $\mathcal{E}^{\text{RGB}}$ is the schema's label for hypotheses attributed to visible image evidence, $\mathcal{E}^{\text{prior}}$ records hypotheses attributed mainly to physical or category knowledge, $\mathcal{U}$ stores unresolved alternatives, and $f^{\star}$ identifies the currently most uncertain factor. The superscript "RGB" refers to the source-image provenance category, not a requirement that the VLM receive color pixels: the evidence is read from O_t, which is grayscale in the evaluated protocol. The figure's "confidence" field records this model-reported uncertainty, and "cross-view consistency" refers to the coupled canonical shape views, not additional camera observations. The provenance distinction keeps prior-supported completion revisable when later image evidence conflicts with it.

A.3 State Update and Normalization

At update round $t = 1, \ldots, T$, with the scheduled observation O_t from Appendix A.1, the frozen VLM proposes

$$\widehat{\mathcal{M}}_t = F_\theta^{\text{update}}\left(O_t, \mathcal{M}_{t-1}, P_{\text{update}}\right), \tag{21}$$

which is passed through the deterministic normalizer,

$$\mathcal{M}_t = \mathcal{N}\left(\widehat{\mathcal{M}}_t, \mathcal{M}_{t-1}\right). \tag{22}$$

Initialization applies the same representation conventions to the first model-generated state, as specified in Algorithm 1.

The normalizer performs no visual recognition, geometric estimation, or optimization. It canonicalizes field values, validates the allowed categorical and normalized values, preserves persistent object identities, and enforces the shared semantics of coupled variables. These are representation-level operations, not an independent check that the hypothesized objects or geometry are correct.

Object persistence follows an explicit lifecycle rule. If a previously active object is omitted from a later model response, its previous record is retained and marked uncertain rather than silently deleted. Removal requires an explicit false-positive decision from the VLM. Omission and rejection therefore have different effects on the state, and the normalizer does not infer object absence from a missing record.

The update instructions allow new image-supported evidence to revise an earlier prior-supported interpretation. Fields without contradictory evidence persist, and unresolved alternatives may remain explicit rather than being forced into an intermediate coordinate estimate. The VLM supplies the semantic judgments; the normalizer maintains the specified representation and persistence rules. Together these operations define the evaluated protocol, without an external detector, geometric solver, or task-specific parameter update. The VLM supplies the visual interpretation; deterministic normalization organizes the resulting state rather than independently judging the scene.

B Recurrent Sufficiency and Deferred Readout

The following analysis characterizes when a terminal task output can replace a richer recurrent state while preserving the behavior of a given update process. It motivates separating state refinement from readout, rather than proving a performance advantage of the particular schema above. We use deterministic abstract update and readout operators; this specifies the scope of the analysis, not an empirical determinism claim about model calls. No contraction, convergence, or correctness relative to GT is assumed.

B.1 Recurrent Sufficiency of a Task Output

Fix the terminal grayscale observation $J = I^{(T)} = \bar{I}$. Let M be an admissible state space closed under the updates under consideration, and let B be the task-output space. The readout is

$$\mathcal{R}_J : \mathsf{M} \to \mathsf{B}. \tag{23}$$

At step t, observations belong to a common admissible set O_t, and Φ_t is defined on $\mathsf{M} \times \mathsf{O}_t$. For $O \in \mathsf{O}_t$, one normalized state update is

$$\mathcal{M}^+ = \Phi_t(\mathcal{M}, O), \tag{24}$$

where t indexes the transition from state t to state $t+1$. Fixed instructions and normalization are included in Φ_t. The states compared below share both the terminal readout image J and the subsequent observation O.

Recurrent sufficiency. We call $\mathcal{R}_J$ *recurrently sufficient* for Φ_t if there exists a transition G_t, defined on attainable readouts and admissible observations, such that

$$\mathcal{R}_J\left(\Phi_t(\mathcal{M}, O)\right) = G_t\left(\mathcal{R}_J(\mathcal{M}), O\right) \quad \text{for all admissible } \mathcal{M} \text{ and } O. \tag{25}$$

The condition asks whether the current readout determines the next task output under the specified update rule. It concerns exact reproduction of that process, not whether the reproduced output is correct.

Define the readout-induced equivalence relation

$$\mathcal{M}^{(1)} \sim_{\mathcal{R}} \mathcal{M}^{(2)} \quad \Longleftrightarrow \quad \mathcal{R}_J(\mathcal{M}^{(1)}) = \mathcal{R}_J(\mathcal{M}^{(2)}). \tag{26}$$

Proposition 1 (Characterization of recurrent sufficiency)**.** *The readout $\mathcal{R}_J$ is recurrently sufficient for Φ_t if and only if, for every pair of admissible states $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}$ and every common admissible subsequent observation O,*

$$\mathcal{M}^{(1)} \sim_{\mathcal{R}} \mathcal{M}^{(2)} \quad \Longrightarrow \quad \Phi_t(\mathcal{M}^{(1)}, O) \sim_{\mathcal{R}} \Phi_t(\mathcal{M}^{(2)}, O). \tag{27}$$

Thus, states with the same current readout must remain indistinguishable at the task-output level after the same update and observation.

Proof. Suppose G_t satisfies Eq. (25). For $\mathcal{M}^{(1)} \sim_{\mathcal{R}} \mathcal{M}^{(2)}$ and a common observation O,

$$\begin{aligned} \mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(1)}, O)\right) &= G_t\left(\mathcal{R}_J(\mathcal{M}^{(1)}), O\right) \\ &= G_t\left(\mathcal{R}_J(\mathcal{M}^{(2)}), O\right) \\ &= \mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(2)}, O)\right), \end{aligned} \tag{28}$$

which gives Eq. (27).

Conversely, suppose Eq. (27) holds. For an attainable output $b \in \mathcal{R}_J(\mathsf{M})$, choose a representative $\mathcal{M}_b$ with $\mathcal{R}_J(\mathcal{M}_b) = b$ and set

$$G_t(b, O) = \mathcal{R}_J\left(\Phi_t(\mathcal{M}_b, O)\right). \tag{29}$$

The assumed condition makes this value independent of the chosen representative for every common admissible O. Consequently, G_t is well defined on attainable readouts and satisfies Eq. (25). □

The proposition is a one-step characterization. If its condition holds for every transition $t = 0, \ldots, T-1$, applying the corresponding G_t recursively reproduces the readout trajectory under any common admissible observation sequence, by induction. Non-injectivity alone is therefore not an obstruction: a readout may merge states as long as the merged distinctions do not affect subsequent task outputs under the update process.

Output-state aliasing. Suppose instead that two admissible states satisfy

$$\mathcal{R}_J(\mathcal{M}^{(1)}) = \mathcal{R}_J(\mathcal{M}^{(2)}), \tag{30}$$

but, for some common subsequent observation O,

$$\mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(1)}, O)\right) \neq \mathcal{R}_J\left(\Phi_t(\mathcal{M}^{(2)}, O)\right). \tag{31}$$

We call this *output-state aliasing*. It violates Eq. (27), so no deterministic transition receiving only the shared current readout and O can reproduce both continuations exactly.

Spatial hypotheses provide an intuitive setting for such a distinction. A thin projected side may be interpreted as shallow object depth or as foreshortening; a hidden extent may be attributed to image evidence or a category prior. Two states can record different interpretations while emitting the same current cuboid proposal. If a later observation leads the specified update process to revise those states into different outputs, the aliasing condition holds. This is an illustrative scenario, not an observed aliasing event: similar coordinates alone do not establish the exact-equality premise of the proposition.

The result does not imply that coordinate-based refinement must fail. A coordinate-generating model may recover useful distinctions from the image, or an alternative update process may produce accurate outputs without reproducing the richer-state trajectory. Conversely, the proposed belief schema is not proved to retain every future-relevant distinction. The characterization clarifies the requirement for lossless state replacement; experiments establish the performance of the concrete inference protocol.

B.2 Deferred Readout Excludes an Explicit Numerical-Feedback Path

An answer-state recurrence can be written schematically as

$$\mathcal{B}_{t+1} = \Gamma_t(\mathcal{B}_t, O_{t+1}), \tag{32}$$

where Γ_t is a coordinate-generating update operator. The current cuboid proposal explicitly conditions the next update, providing a path by which numerical errors or unsupported proposals can influence later predictions. The existence of this path does not determine whether feedback helps or hurts.

The belief protocol instead propagates

$$\mathcal{M}_{t+1} = \Phi_t(\mathcal{M}_t, O_{t+1}), \qquad t = 0, \ldots, T-1, \tag{33}$$

and reads out the result only after refinement:

$$\mathcal{B} = \mathcal{R}_J(\mathcal{M}_T). \tag{34}$$

For analysis, consider the auxiliary readout

$$\widetilde{\mathcal{B}}_s = \mathcal{R}_J(\mathcal{M}_s), \qquad 0 \le s < T. \tag{35}$$

This is a conceptual readout, not an extra model call in the evaluated loop. It is never supplied to a later update. An external edit to this unused output, holding $\mathcal{M}_s$ and subsequent observations fixed, cannot change

$$\mathcal{M}_{s+1} = \Phi_s(\mathcal{M}_s, O_{s+1}),$$

since the update has no argument for $\widetilde{\mathcal{B}}_s$. The same independence holds inductively for the remaining states and the terminal output.

This is an architectural exclusion of an explicit conditioning path, not a claim that belief-state errors vanish or that inference converges. The coordinate-feedback control in Sec. 4.3 addresses a separate empirical question. It retains the belief-state schema, persistence rules, and observation schedule while adding numerical cuboid proposals to subsequent conditioning. Sharing these rules does not hold the generated belief trajectory fixed. In particular, adding coordinates while retaining the belief state is not the state-replacement operation characterized in Sec. B.1. The formal analysis therefore does not predict the direction or magnitude of the feedback control's performance difference.

B.3 Deterministic Evidence Ordering Does Not Create Information

Evidence ordering is separate from recurrent sufficiency. For this information-theoretic statement, I denotes a random RGB source image and $\bar{I} = \mathrm{Gray}(I)$ its grayscale version. This differs from fixing the terminal observation J in the previous analysis. Let Z denote latent scene variables relevant to the target cuboids, and let

$$I^{(t)} = \mathcal{T}_{\lambda_t}(I) \tag{36}$$

be an observation produced by a fixed deterministic transformation. Write $\mathcal{I}(\cdot;\cdot)$ for mutual information.

Proposition 2 (Information non-creation)**.** *For every deterministic observation $I^{(t)} = \mathcal{T}_{\lambda_t}(I)$,*

$$\mathcal{I}(Z; I^{(t)}) \le \mathcal{I}(Z; I). \tag{37}$$

For the grayscale-based schedules used here, the complete sequence satisfies

$$\mathcal{I}(Z; I^{(0:T)}) \le \mathcal{I}(Z; \bar{I}) \le \mathcal{I}(Z; I). \tag{38}$$

Proof. The deterministic dependence of $I^{(t)}$ on I gives the Markov chain

$$Z \longrightarrow I \longrightarrow I^{(t)}.$$

Equation (37) follows from the data processing inequality. For the evaluated schedules, the entire tuple is a deterministic function of $\bar{I}$, which is itself a deterministic function of I. Applying the same inequality to $Z \to I \to \bar{I} \to I^{(0:T)}$ gives Eq. (38). □

Algorithm 1 Persistent Belief Refinement with Deferred Readout

Require: RGB source image I; frozen VLM F_θ; C2F scales $\sigma_0 > \cdots > \sigma_T = 0$; schedule ω; fixed instructions $P_{\text{init}}, P_{\text{update}}, P_{\text{proj}}$; normalizer $\mathcal{N}$
Ensure: Projective cuboids $\mathcal{B}$

$\bar{I} \leftarrow \text{Gray}(I)$; $I^{(T)} \leftarrow \bar{I}$
if $\omega = \text{Raw}$ **then**
  Set $O_t \leftarrow \bar{I}$ for $t = 0, \ldots, T$
else
  Construct $I^{(t)} \leftarrow G_{\sigma_t}(\bar{I})$ for $t = 0, \ldots, T$
  Set $O_t \leftarrow I^{(t)}$ for C2F or $O_t \leftarrow I^{(T-t)}$ for F2C
end if
$\mathcal{M}_0 \leftarrow \mathcal{N}\big(F_\theta^{\text{init}}(O_0, P_{\text{init}})\big)$
for $t = 1, \ldots, T$ **do**
  $\widehat{\mathcal{M}}_t \leftarrow F_\theta^{\text{update}}(O_t, \mathcal{M}_{t-1}, P_{\text{update}})$
  $\mathcal{M}_t \leftarrow \mathcal{N}(\widehat{\mathcal{M}}_t, \mathcal{M}_{t-1})$
end for
$\mathcal{B} \leftarrow F_\theta^{\text{proj}}(I^{(T)}, \mathcal{M}_T, P_{\text{proj}})$
return $\mathcal{B}$

This direct application of the data processing inequality rules out an information-creation interpretation of the transformed observations. The terminal readout reuses the unblurred grayscale endpoint already included in the sequence. Reusing this observation does not supply an additional scene measurement. Neither inequality establishes that smoothing preserves every task-relevant cue or that a particular ordering improves predictions.

For a fixed model representation map ϕ_θ, write

$$H_t = \phi_\theta(I^{(t)}), \qquad H_{\text{raw}} = \phi_\theta(\bar{I}). \tag{39}$$

The data processing inequality alone does not order $\mathcal{I}(Z; H_t)$ and $\mathcal{I}(Z; H_{\text{raw}})$. Different presentations of the same source image may induce different representations in a fixed model, without creating scene information. We do not estimate these internal mutual informations or use them as evidence that C2F improves inference. The C2F/F2C comparison instead evaluates the presentation order of the same transformed observations, while Raw evaluates the protocol without that schedule.

C Complete Inference Procedure

Algorithm 1 summarizes initialization, state refinement, and terminal readout. The ordered observations follow Appendix A.1: C2F traverses the grayscale levels from coarse to fine, Raw repeats the unblurred endpoint, and F2C reverses the refinement order. The terminal readout image remains $I^{(T)} = \bar{I}$ in each case. The procedure uses $T + 2$ logical calls, so the six-call conditions use $T = 4$.

All roles use the same frozen VLM with fixed role-specific instructions. The image transforms and normalizer introduce neither an external detector nor a geometric optimizer or task-specific learning step. Intermediate cuboids are absent from the recurrent loop. In all schedules, the final call receives the unblurred grayscale image and the refined state. Logical-call counts and realized token usage are distinguished in Sec. 4.1.

D Additional Qualitative Results

Figures 3 and 4 show selected unannotated indoor scenes, complementing Fig. 2. The first figure presents Gemini outputs; the second shows GPT-5.6-Sol and Qwen3.8-Flash. All model inputs follow the grayscale convention in Appendix A.1; registered RGB images are visualization backgrounds rather than extra inputs. The panels compare final hypothesis sets, including their multiplicity, overlap, projected extent, and depth-like structure.

The visible differences are not limited to displacements of corresponding corners: protocols may emit different object sets, overlapping cuboids, or different projected depth extents. A compact or

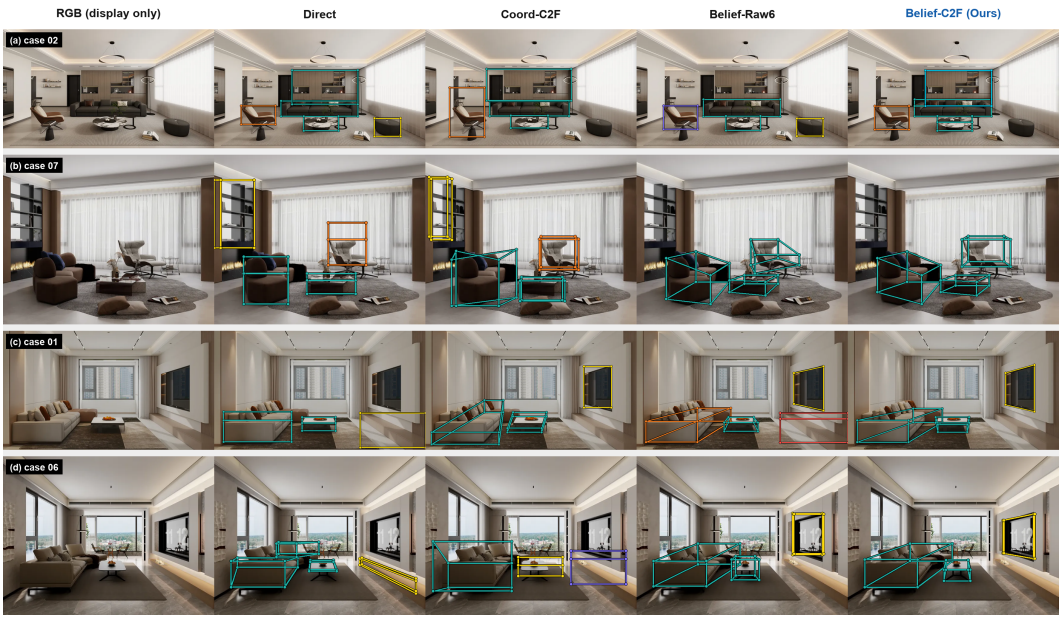


Figure 3: **Additional qualitative results with Gemini 3.1 Pro.** Each row is a selected example from the separate unannotated qualitative set. Columns show the registered RGB visualization and final cuboids from DIRECT-1, COORD-C2F, BELIEF-RAW$_6$, and BELIEF-C2F; the figure uses the shortened labels DIRECT and BELIEF-RAW. Inference uses grayscale observations, including the terminal readout; color backgrounds are for display only. The examples illustrate differences in overlapping hypotheses, nearly planar projections, and cuboids spanning large image regions. Examples were selected for visual clarity and contain no GT overlays; they illustrate prediction structure rather than provide an independent accuracy estimate.

visually coherent set can nevertheless omit real objects, and apparent projected depth need not be correct. Final-output panels also do not reveal when hypotheses changed during refinement. Accordingly, accuracy conclusions rest on the held-out metrics in Sec. 4.2, the targeted controls in Sec. 4.3, and the cross-backbone comparison in Sec. 4.4, rather than on these unannotated visualizations.



Figure 4: **Qualitative results across additional VLM backbones.** Rows (a–b) show GPT-5.6-Sol outputs and rows (c–d) show Qwen3.8-Flash outputs. The column order matches Fig. 3: RGB visualization, DIRECT-1, COORD-C2F, BELIEF-RAW$_6$, and BELIEF-C2F. The backbones receive grayscale inputs; RGB backgrounds are used only for visualization. No GT cuboids are shown. These examples display variation in final object sets and projected geometry across protocols and backbones. Quantitative cross-backbone comparisons are reported separately in Table 3.