Calibrated Clarification: A Computational Maxim of Quantity for Grounding Under-Specified Human Intent

Mehul Menon, Ruixuan Liu, Philip Huang, Jiaoyang Li and Changliu Liu

***Abstract*— A short prompt rarely says exactly what a person wants, yet a generative system commits to a finished result anyway, sometimes at real cost. We close that gap by clarifying interactively: the system shows two candidates, asks which one comes closer to what the person had in mind, and repeats. Done naively this fails dangerously, because the system grows confident while its guess has not moved toward the target. The scoring rule causes it. Comparing candidates by the angle between them leaves each answer nearly uninformative once the space grows large, so the system stops early having learned almost nothing. We introduce CLASP, which replaces the angle with an un-normalized closeness comparison and adds a learned low-dimensional intent basis, uncertainty-directed probing, a self-regulating update, and calibrated commit guards, which is tested on a simulated response model. Together these reach most of what such comparisons can convey in roughly $8\times$ fewer questions. We then certify three properties of the committed result: that it covers what the person meant, that it is specific enough to act on, and that someone could realize it. Knowing when to stop proved harder, since seven candidate rules all failed to commit near a session's best moment, so we cap the budget instead, which suffices because over-asking measurably degrades the result.**

I. INTRODUCTION

A short prompt names a category, not the particular thing a person pictures. The pipeline commits to one anyway and delivers a finished artifact built from an under-specified request: an image, a 3D asset, or a physical object that a robot then assembles [1]. Dialogue closes that gap, all the more so because people often cannot describe what they want and only recognize it when they see it. Language does not always serve either. Intent that is comparative ("warmer than that"), perceptual, or continuous resists description but invites recognition, so showing two candidates and asking which comes closer turns an ill-posed description into a well-posed choice [2, 3]. We study non-verbal clarification [4], where the system asks with candidate images rather than words. Take the prompt "a LEGO dog." It fixes the category but not the pose, so a pipeline would fall back on some default build. Our system shows brick-built candidates two at a time and, after a fixed budget of A/B questions, commits to the intended one: a walking dog rather than a sitting, lying, or jumping one (Fig. 1).

The harder question is not what to ask but when to stop, without wasting the person's effort. Grice's Maxim of Quantity puts it as be as informative as required and no more [5], and current human-robot dialogue systems implement no principled version of it. Throughout we call the system's running estimate of the target its belief and one full dialogue a session. Recent work has a robot ask for help once a calibrated uncertainty threshold falls [6]. Grounding an intent means controlling several risks at once, and we show which of them admit a distribution-free certificate, one assuming nothing about the shape of the data beyond calibration and future requests staying exchangeable.

We make four contributions and validate each on a high-dimensional grounding testbed.

1) **Why naive clarification miscarries (Sec. III-B):** Our concept space carries $d{=}2304$ coordinates, far more than the questions a person will tolerate. There, comparing candidates by angle makes each answer less informative as the space grows, at a rate we characterize, while the belief tightens anyway, so the system turns confident and stops before its guess has moved.
2) **CLASP, a sample-efficient method (Sec. III-C):** CLArification by Self-regulating Probing has five parts: a learned low-dimensional intent basis, an un-normalized closeness comparison replacing cosine, uncertainty-directed probing of the least-known direction, a self-regulating update whose step shrinks with the belief's spread, and calibrated commit guards at a fixed one-pass budget. The first four reach most of what these comparisons can reach in $\sim 8\times$ fewer questions, and the fifth makes the committed result certifiable.
3) **A commit-quality criterion and a limit on stopping (Sec. III-D):** A Learn-then-Test procedure [7] certifies three risks of the committed belief at once: coverage, commit-readiness, and feasibility. Against this we report a negative result: seven mechanisms failed to certify a useful stopping rule, one committing near a session's peak, because those peaks fall at widely varying turns, though we claim no impossibility in general. Since over-asking degrades the result, we commit at a fixed budget of one question per intent dimension, which beats every other budget we tried.
4) **The response modality matters most when the person answers noisily (Sec. IV-A):** A graded strength-of-preference answer carries more than a binary choice and pins down intent in fewer turns. Its edge holds across answer noise and grows largest where noise runs highest, since a graded answer reports magnitude and so discounts unreliable answers on its own.

Nothing in the method ties to images, to a backbone, or to non-verbal questions, and it carries over to verbal clarification; we use visual intent as a testbed because ground-truth targets make grounding measurable. BRICKGPT [1] is the

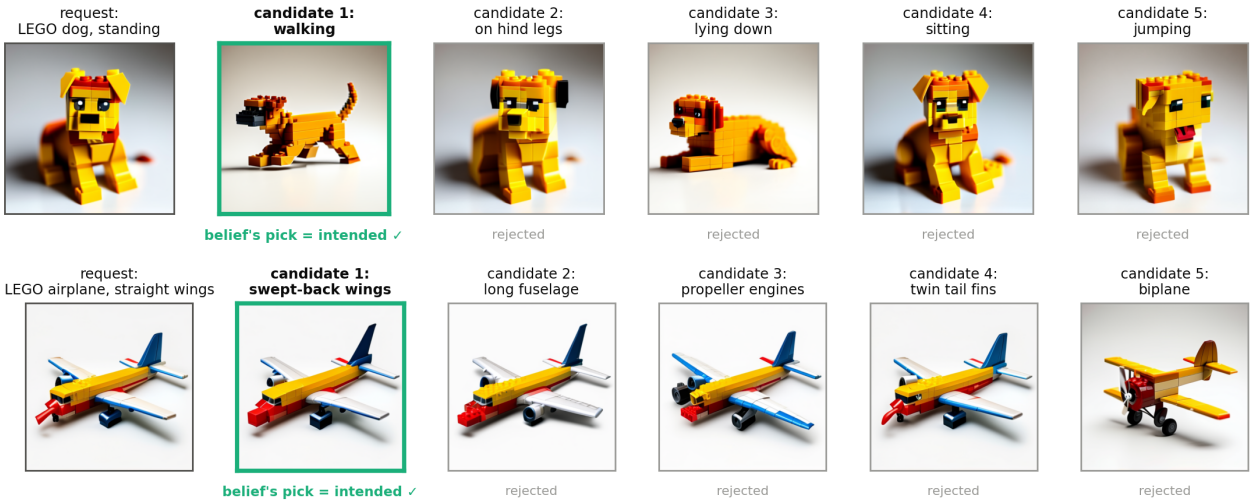


Fig. 1. **A prompt names a category; clarification picks the product.** Each strip shows an under-specified request (left) and five candidate designs of one LEGO model. Our system asks 64 A/B questions, ranks the candidates by its belief, and boxes its top pick in green: "a LEGO dog" resolves to a walking dog, and an airplane to swept wings. We fit the basis only on EMU-EDIT photo edits, so separating these LEGO variants tests out-of-distribution transfer.

application we have in mind: it turns a prompt into a stable LEGO structure a robot assembles brick by brick, its input stays under-specified, its output costs real time to redo, and its buildability constraint is the physical form our feasibility guarantee would take. We leave that front-end to future work and report no robot experiments.

II. Related Work

Our work sits between three lines: dialogue research on clarification, which studies what to ask; calibrated uncertainty for robots, which studies when to defer; and generative pipelines, which do neither.

Clarification and knowing when to stop: Common ground built through interaction anchors a foundational account of communication [2, 8], clarification requests form a well-studied dialogue act [3, 4], and grounding language for manipulation [9, 10] sits at the centre of situated human-robot interaction [11, 12]. Almost all of it studies what to ask; far less asks when the system has learned enough to act, though stopping is classical elsewhere, in sequential testing [13] and active learning [14, 15]. We certify the quality of what the system commits to rather than the stopping time, and report where that certification ran out (Sec. III-D).

Uncertainty and help-seeking: A robot that can tell when it does not know can defer instead of act. Most work estimates that uncertainty with model-based tools such as MC-dropout [16], deep ensembles [17], or a split between aleatoric and epistemic uncertainty [18], then asks for help once a threshold falls, as robots do when requesting assistance after a failure [19, 20]. Such estimates guarantee nothing in finite samples, because a person tunes the threshold and the realized error rate lands wherever it lands. Closest to us, Ren et al. [6] apply conformal prediction [21] to trigger help-seeking with a guaranteed error rate on a single risk at a single decision. Learn-then-Test [7] widens that guarantee from prediction sets to arbitrary risks by recasting risk control as multiple hypothesis testing, which lets us certify several properties of one object together. We extend it from one threshold to a multi-risk criterion on a belief committed at the end of an interactive loop, and map its boundary.

Generative pipelines commit from one prompt: Text-to-image systems [22–24], instruction-based editors [25, 26], text-to-3D [27], and text-to-build pipelines [1] share one interface: one prompt in, one artifact out. When the prompt under-specifies the target the person can only re-prompt or sample a batch and pick, and neither elicits what they meant. None contains a clarification stage or any notion of having asked enough. Our criterion attaches to an intent representation rather than a generator, so it can front-end any of them.

Active preference-based learning: Learning rewards or intents from pairwise comparisons [28–30] yields question-efficient elicitation, but usually guarantees nothing distribution-free about what it elicits and says little about why elicitation degrades in high dimensions. We supply both.

III. Problem and Method

A. Setup

A concept is one of the atomic, semantically meaningful factors that a concept-aligned generative model learns from paired text and images, such as a colour, a material, or a pose. CONCEPTALIGNER [31] learns these with identifiability guarantees and represents a result as $C{=}64$ concept vectors of dimension $d{=}2304$ each. We represent the person's intent, the specific result they want, as an unknown point ϕ^* in a concept space $\mathbb{R}^d$ where each point stands for one possible result. The prompt fixes only a category default ϕ_0, so the

person wants the displacement $\Delta^* = \phi^* - \phi_0$, and Sec. IV reports progress as the fraction of Δ^* the system recovers.

The system keeps a mean μ_t, its best guess, and per-coordinate variances $\sigma^2_{m,t}$ that we collect in a diagonal Σ_t. We call the pair the belief. It tracks an estimate rather than an exact Bayesian posterior, for reasons Sec. III-C gives, and it does three jobs: choosing the next question, setting the step size, and defining a credible region at commit time. Since it is not an exact posterior, that region guarantees nothing on its own, and the calibration in Sec. III-D supplies the guarantee.

At turn t the system picks a unit direction v_t and shows two candidates either side of its current guess,

$$a = \mu_t + \rho_t\, v_t, \qquad b = \mu_t - \rho_t\, v_t,$$

setting the half-width ρ_t proportional to the belief's spread along v_t, so the candidates stay about as far apart as the system is unsure. The person answers $y_t \in \{+1, -1\}$. To turn "closer" into a number we posit a score $s(x)$ for how well a candidate matches the target, and we model the noisy answer with the Bradley-Terry law for pairwise choice [32],

$$\Pr(a \succ b) = \sigma\big(\beta\,(s_a - s_b)\big), \quad s_a \coloneqq s(a),\ s_b \coloneqq s(b), \quad (1)$$

where σ denotes the logistic function and $\beta > 0$ a temperature, with larger β meaning a more reliable person. The choice of s is no detail. It decides how much one answer can teach, and it is where naive clarification fails.

Notation: Subscript t indexes turns and m indexes coordinates. Three numbers describe the space, and two of them happen to both be 64, so we pin them down here. The backbone encodes a result as 64 concept vectors of dimension d=2304, and Sec. III-B analyses d, the dimension in which the system compares candidates. The belief runs not there but over K learned directions (K=64 by default, Sec. III-C), so "one pass" means K questions. We fix every constant once and list them all in Sec. IV.

B. Why naive clarification miscarries

The conventional score takes the angle between candidate and target, $s_{\cos}(x) = \langle x, \phi^* \rangle / (\|x\|\,\|\phi^*\|)$. It serves as the standard similarity in learned embedding spaces [33] and as the score in the design we start from, a weighted cosine over the concepts of CONCEPTALIGNER [31], a state-of-the-art controllable text-to-image model. The rationale holds up: dividing by each vector's length discards raw magnitude and stops a candidate from winning by sheer size. We adopt it as the naive default, computing both scores in Eq. (1) this way, and ask what it costs.

An answer earns its place only if the person would have chosen differently had the target lain elsewhere, that is, if the preference probability in Eq. (1) shifts as ϕ^* moves. The Fisher information measures this [34], and for one answer probed along v it splits into two readable terms,

$$\mathcal{I}(v) = \underbrace{\beta^2\,\bar\sigma(1-\bar\sigma)}_{\text{does the question run close?}} \underbrace{\big\|\nabla_{\phi^*}(s_a - s_b)\big\|^2}_{\text{do the scores react?}}, \quad (2)$$

where $\bar\sigma$ abbreviates $\sigma(\beta(s_a - s_b))$ at the current pair, a number rather than the logistic function. The first term says an obvious question teaches nothing, so information peaks at a coin flip. The second says the scores have to move when the target does.

Only the second term tracks the size of the space. Since the candidates straddle the current guess at a half-width matching the belief's own spread, every question already sits near a coin flip, which pins the first term at $\Theta(1)$ in d. Cosine suppresses the second, because it divides by the candidate norm: with coordinates of comparable scale $\|x\|^2$ is a sum of d comparable terms, so $\|x\| = \Theta(\sqrt{d})$ and the score gap is rescaled toward zero as coordinates are added. Writing $\mathcal{I}_{\cos}$ for Eq. (2) under cosine scoring,

$$\mathcal{I}_{\cos} = \Theta(1/d), \qquad \text{per-answer signal } S(d) \propto d^{-1/2}, \quad (3)$$

Two consequences follow, both from $\mathcal{I}_{\cos} = \Theta(1/d)$. Information scales as the square of the score gap, so the gap itself, the per-answer signal $S(d) = \mathbb{E}|s_a - s_b|$ that Fig. 3 plots, falls as $\sqrt{\mathcal{I}} \propto d^{-1/2}$. And since the error after n answers scales as $1/\sqrt{n\mathcal{I}}$ [34], holding that error fixed needs $n \propto 1/\mathcal{I} \propto d$ answers. At d=2304 that is a dialogue no person would sit through.

We repair this by dropping the normalization without giving up on "closer." We score a candidate by un-normalized closeness, $s(x) = -\frac{1}{2}\|x - \phi^*\|^2$, which states what we ask the person: prefer whichever candidate sits nearer the target. For the symmetric pair above this yields

$$s_a - s_b \;=\; 2\rho_t \,\big\langle v_t,\ \phi^* - \mu_t \big\rangle, \quad (4)$$

an inner product with the residual $\phi^* - \mu_t$. It informs when the probed direction overlaps the error that remains and stops informing once that error runs orthogonal to it, which is how a clarification question should behave, and its gradient in ϕ^* has norm $2\rho_t$ independent of d, so $\mathcal{I} = \Theta(1)$ and the collapse disappears. A raw inner product $\langle x, \phi^* \rangle$ carries the same information, and RLHF/DPO reward models score that way [28], but its gap never involves μ_t and so cannot express "closer than my current guess"; the $-\frac{1}{2}\|x\|^2$ term supplies that dependence.

Cosine does worse damage than slow learning, because the system cannot tell it is learning nothing. The update contracts Σ_t on a schedule that counts every answer as evidence, while μ_t moves only as far as the answers informed it. Under cosine the two come apart: the belief tightens, the system concludes it has understood, and it commits with μ_t still far from the target. That cosine grows uninformative in high dimensions is classical geometry; the consequence for interactive elicitation, where it corrupts the system's read on its own progress, is what we add. Note that our fix addresses the score, not this contraction schedule, which CLASP inherits.

C. CLASP: a sample-efficient method

Fig. 2 shows the loop. Four changes follow from Sec. III-B, and together with the commit guards of Sec. III-D they make up CLASP.

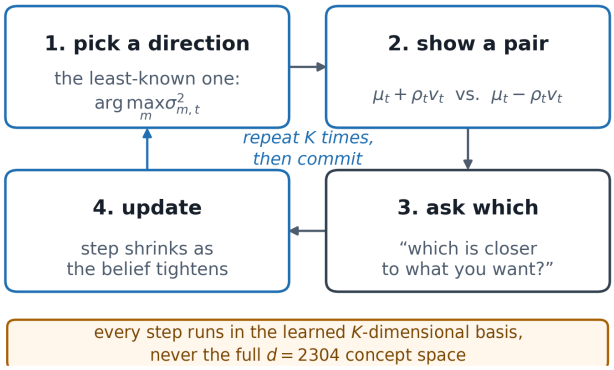


Fig. 2. **One turn of CLASP.** Step 2 places a candidate a half-width ρ_t either side of the current guess μ_t, so the pair sits as far apart as the system is unsure, and step 4 moves μ_t by an amount that shrinks as the belief tightens. After K questions, one per intent dimension, the system commits under the guards of Sec. III-D.

1) **Compare by un-normalized closeness:** Eq. (4), rather than by angle. This restores $\Theta(1)$ information per answer and pulls the most weight of the four.
2) **Probe the most uncertain coordinate:** $m^\star = \arg\max_m \sigma^2_{m,t}$, which follows standard uncertainty sampling [14]. A real intent stays diffuse and touches most coordinates, so spreading questions this way matches an oracle told in advance which coordinates matter.
3) **Take a self-regulating step:** Write $y_t = +1$ when the person prefers a and -1 for b, and let v_t be the unit vector along the probed coordinate $m^\star$. We move the mean along that direction,

$$\mu_{t+1} = \mu_t + \eta_t, y_t, v_t, \qquad \eta_t = \frac{\sigma_{m^\star!,t}}{1 + c, n_{m^\star}},$$

where $\sigma^2_{m^\star!,t}$ is the probed coordinate's entry of Σ_t, $n_{m^\star}$ counts earlier probes of it, and $c > 0$ sets the decay. The $1/(1 + c, n)$ factor is the classical Robbins–Monro form [35]; the $\sigma_{m^\star!,t}$ factor is what makes the step self-regulating, shrinking it as the belief tightens on the coordinate being asked about. Since that spread contracts geometrically, the steps are summable and we claim no asymptotic convergence, only a mean that keeps moving without overshooting across a short budget. An exact posterior step instead moves in proportion to the variance, which contracts faster than the mean can travel under noisy binary comparisons, so the estimate freezes short of the target.
4) **Work in a learned basis:** Edit displacements do not fill the space, so we fit their principal components [36] on a corpus and run the belief over the top K modes, letting one question settle a correlated bundle of coordinates. Reducing to a low-dimensional subspace repeats the move that makes high-dimensional Bayesian optimization tractable [37].

The probe width earns its keep here too: holding ρ_t proportional to the belief's spread keeps each question near a coin flip, where the first factor of Eq. (2) peaks. CLASP therefore never operates in the raw d=2304 regime, so the $1/\sqrt{d}$ collapse describes the space it escapes rather than the one it runs in.

D. Commit-quality, and the limit on stopping

The belief tracks a heuristic estimate, so its credible region cannot certify itself. We repair this at commit time with Learn-then-Test [7], a distribution-free procedure that turns a chosen set of risks into a guarantee, and we use it unchanged as a black box. We supply a calibration set of completed sessions, the three risks below, and a target level α_j for each. It returns a configuration λ, for us the credible region's radius, the spread below which the system will act, and how far off that manifold we tolerate, carrying

$$\Pr\left(R_j(\lambda) \le \alpha_j \quad \text{for all } j\right) \;\ge\; 1 - \delta$$

over the draw of the calibration set, assuming only that calibration and future sessions stay exchangeable. We then run with that λ, and Table II reports what it achieves on held-out sessions. We claim no credit for the procedure, only for what we ask it to certify.

The three risks are miscoverage R_{cov}, whether the region contains the intent; too-diffuse-to-commit R_{help}, whether the remaining uncertainty runs tight enough to act on; and infeasibility R_{mfld}, whether the committed point could exist as a real result at all. Our image proxy tests membership of the manifold the training edits trace out. A build pipeline such as BRICKGPT [1], where the result has to stand up, would make it a stability check.

None of this settles when to stop. Grounding peaks after about one pass and then degrades, and that peak falls at session-dependent turns, so none of the stopping rules we built commits near it (Sec. IV). Coverage and feasibility instead guard the committed point, ruling out a commit that looks confident but lands wrong, or that no one could realize, and we bound the timing with a fixed one-pass budget.

IV. EXPERIMENTS

Setup: A session is one grounding episode: the person has an edit in mind, the system sees only the starting request and its category, and after a fixed budget of A/B questions it commits. We score by recovery, the fraction of the way the committed edit travels from the start to the true target, so 1.0 lands on target and 0 never moved. Requests come from EMU-Edit [26]: 1608 before/after caption pairs in 8 balanced categories, encoded through the backbone's text-to-concept map [31]. The categories partition the pairs rather than setting a sample size, and enter only through leave-one-category-out splits. The person is a simulated Bradley-Terry answerer with tunable noise; since it uses the likelihood the belief update assumes, every number below is a best case.

Implementation: Four constants set the loop's behaviour. The answerer's temperature is $\beta = 4.0$. The probe half-width is $\rho_t = 2\sigma_{m^\star,t}$, which keeps each question near the coin flip where Eq. (2) is largest. The step decays at $c = 0.3$, and

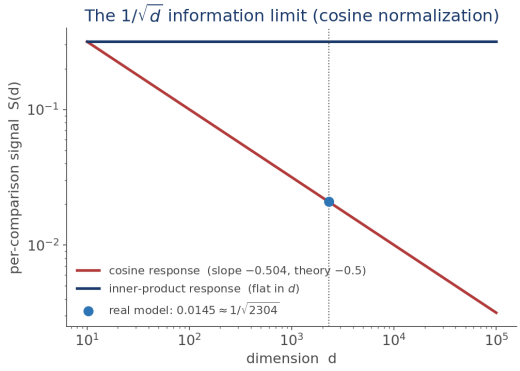


Fig. 3. Why naive clarification stalls. Per-answer signal against dimension, log-log. Compared by angle the signal falls off like $1/\sqrt{d}$; compared without normalization it stays flat. The clean $-1/2$ slope is the isotropic case, and real spaces flatten to a plateau instead (see text).

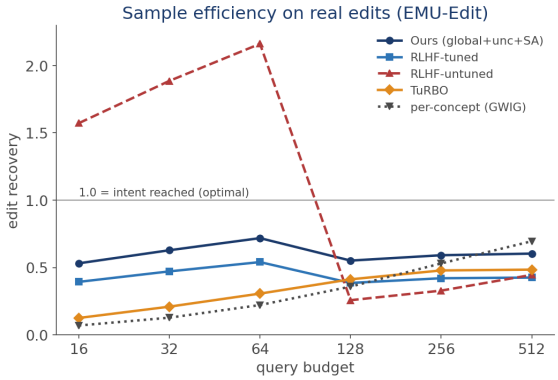


Fig. 4. Recovery against number of questions on held-out EMU-Edit requests, where 1.0 is the true target and 0.61 is the most this K=64 basis can represent. CLASP reaches that level in ∼64 questions where the factored baseline needs ∼512, though the two differ in dimension as well as in rule (see text).

each answer scales the probed coordinate's variance by 0.7, which is what makes the schedule sweep the $K = 64$ modes in turn. Calibration runs at $\delta = 0.10$. Sample sizes differ by experiment: efficiency uses 120 held-out edits, calibration 400 each for the calibration and test splits, the stopping study 260 sessions split 130/130 over a 512-question horizon, the basis comparisons 200 edits, selection 1250 trials per point, and the scaling simulations 4000 per point. Every number is on held-out requests.

Baselines: All methods see the same sessions and the same simulated answerer. Preference-tuning follows the reward gradient at a hand-tuned rate (swept on calibration sessions) or a default one, which is our update without the self-regulating schedule; this is the RLHF/DPO recipe and the closest comparison. Trust-region search probes a random unit direction and shrinks its radius after unproductive rounds, in the spirit of trust-region Bayesian optimization [38] but without a surrogate. Per-concept is the published factored belief, treating the 64 concept coordinates independently at 8 modes each.

Cosine loses information at the rate the analysis predicts: We measure the mean per-answer signal at each dimension on a grid from d=4 to d=2304 and fit a least-squares line to $\log S$ against $\log d$. When every direction carries equal uncertainty its gradient is -0.504 against a theoretical -0.5, and the signal falls to 0.016 at d=2304 while the un-normalized signal holds near 1.1 (Fig. 3). Real spaces are kinder, because they vary along fewer directions: the slope shallows to -0.319 for CLIP text embeddings [33] and -0.239 for our concept space, where cosine settles at a plateau of 0.286 rather than vanishing. So the regime is starved rather than collapsed, and it is effective rather than ambient dimension that sets the floor. A 4× drop still leaves a short dialogue uninformative, which is what the next result has to overcome.

CLASP reaches the ceiling 8× sooner: A K=64 basis can represent at most 0.61 recovery, so that is the number to judge the rest against. CLASP reaches 0.718 at 64 questions, overshooting it, so we score by distance from perfect recovery rather than recovery itself: overshoot should be penalised, not rewarded. On that measure CLASP sits 0.28 from perfect, ahead of tuned preference-tuning (0.46), trust-region search (0.69), and the factored baseline (0.78), while untuned preference-tuning is worst at 1.16, racing past the target before wandering back. The factored baseline needs ∼512 questions to match, hence 8× (Fig. 4). Since CLASP is representation-agnostic, its learned basis is one of the five parts we propose rather than a nuisance variable, so this compares two complete systems and not two update rules.

What the system commits to is certified: Learn-then-Test holds all three risks under their targets at once (Table II): miscoverage 0.068, so coverage 0.932, against a certified bound of 0.20 from the union over the coverage and feasibility levels; infeasibility 0.095; and too-diffuse-to-commit 0.000, a zero-violation count rather than a risk of exactly zero. These beliefs come from the loop run at a synthetic offset rather than from full CLASP sessions, so what they establish is that the certification machinery holds on real edits. Held-out categories are a deliberate shift rather than a certified transfer, and recovery across them averages 0.52.

Grounding does not improve monotonically with questions: It climbs to a peak near 64 and then falls away if the system keeps asking (Fig. 5). The probe schedule puts the peak there. Probing $\arg\max_m \sigma_m^2$ spends its first K=64 turns visiting each intent dimension exactly once, since probing a coordinate shrinks that coordinate's variance and reduces its uncertainty (since it has been asked about). Once no fresh coordinate is left, every later question re-probes a dimension whose variance has already collapsed, and with it the step size and the information the answer carries, so those answers move the estimate without informing it. That makes the peak a kink rather than a smooth optimum, and it predicts the kink moves with K.

Yet no rule we built commits near it: Sessions peak at a median of 130 turns with a middle-half range of $[64, 321]$, so no single turn is right for all of them. We tried seven rules across four families, among them Learn-then-Test over the

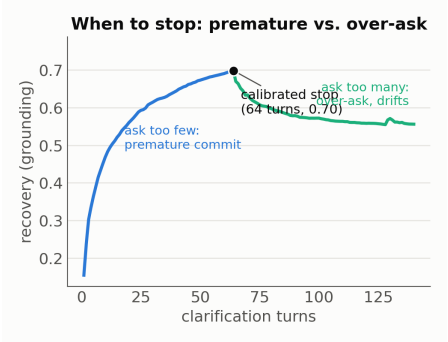


Fig. 5. **Over-asking degrades grounding.** Recovery measures how far the committed result travels from the prompt's default toward the person's true intent, where 1.0 lands on target and 0 never moved. Grounding climbs to a peak near K=64 questions, one pass over the K learned intent dimensions the belief runs in, then falls as the system keeps asking and settled dimensions drift. The calibrated stop we tried (green, "sessions committing") lands at the budget end rather than the peak, because per-session peaks scatter too wide for a distribution-free threshold to track. CLASP therefore commits at a fixed one-pass budget of K questions.

TABLE I

QUESTIONS NEEDED TO PIN THE COMMITTED DIRECTION TO cos =0.85 AGAINST GRADING NOISE. LOWER IS BETTER; THE LAST COLUMN IS THE BEST GRADED SPEED-UP OVER BINARY.

grading noise	binary	Likert-5	continuous	best gain
0.0	104	81	73	1.43×
0.3	105	96	91	1.16×
0.6	199	136	129	1.55×
1.0	> 256	> 256	> 256	—

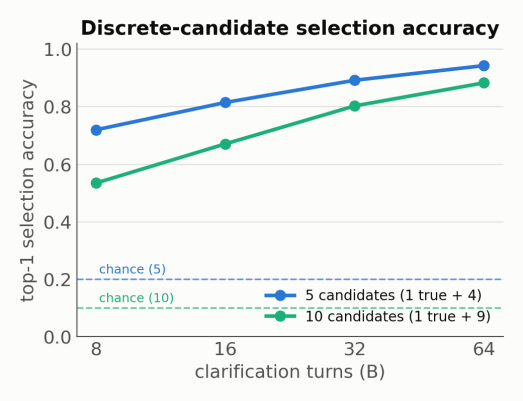


Fig. 6. Picking the intended edit from a candidate set on held-out EMU-Edit requests (1250 trials per point). Asked to rank the true target against 4 or 9 decoys, the top choice lands correct 0.94 and 0.88 of the time by 64 questions, against chance at 0.20 and 0.10 (dashed).

TABLE II

THREE-RISK CALIBRATION OF THE COMMITTED BELIEF ON HELD-OUT REAL EDITS, A GUARANTEE ON WHAT THE SYSTEM COMMITS TO RATHER THAN ON WHEN IT STOPS. WE CHOOSE EACH TARGET IN ADVANCE AS A TOLERANCE; LEARN-THEN-TEST CERTIFIES EACH RISK STAYS BELOW IT WITH CONFIDENCE $1-\delta$ ($\delta = 0.10$), AND ACHIEVED GIVES THE HELD-OUT ESTIMATE.

	miscoverage (R_{cov})	too-diffuse (R_{help})	infeasibility (R_{mfld})
achieved	0.068	0.000	0.095
target ($\leq$)	0.200	0.200	0.100

belief's uncertainty, iterate-averaging, coordinate-freezing, and a change-point statistic. None commits near the per-session peak, and the calibrated stop exhausts the budget instead, which is a negative result about the mechanisms we evaluated rather than an impossibility proof. Committing after one pass is the robust fallback: 0.28 from perfect recovery at K, against 0.37 to 0.47 at every other budget from 16 to 512.

Choosing among candidates works where rendering does not: Grounding in concept space barely reaches pixels, with the committed render only +0.005 CLIP-I [33] and +0.019 DINO [39] closer to the target than the request-only render. That is a ceiling of the representation rather than a failure of inference, and it is specific to rendering a partial edit: it vanishes once the belief ranks discrete candidates instead. We hand the elicited belief a pool holding the person's true target edit together with four or nine decoy edits drawn from other requests, ask it to rank the pool, and report how often the true target comes out first. After 64 questions it does so 0.94 of the time in a pool of five and 0.88 in a pool of ten, against the 0.20 and 0.10 a random ranking would give (Fig. 6).

A. Graded versus binary answers

A clarification question need not stay a bare either/or: the person can say *how much* closer on an L-point scale [40], which carries more than one bit per turn. Grading helps, and more levels help more, with direction agreement at a 64-question budget rising from 0.785 for binary to 0.804 for a 5-point scale and 0.811 for continuous, though those gaps are too small to read without seed-averaged intervals. Table I shows the effect on question counts: graded feedback saves 1.1 to 1.55×, and its edge is widest at noise 0.6, where a binary answer is a bare sign that a noisy flip sends the wrong way while a graded one reports magnitude and so discounts itself. The gain sits well below the $\log_2 L \approx 2.3\times$ a bit count would suggest, capped by the same $1/\sqrt{d}$ limit, and at noise 1.0 no modality reaches the threshold at all. The comparison also stays inside the non-verbal channel, so it shows only that grading beats a binary answer, not that language is the wrong modality.

V. CONCLUSION AND LIMITATIONS

We recast interactive intent grounding as calibrated clarification: why naive clarification turns confident but wrong, CLASP as a sample-efficient answer, and a certification of commit quality paired with a stopping-time negative. Four runs would sharpen the evidence. A human study would replace an answerer simulated with the belief's own likelihood. Calibrating on full CLASP sessions rather than a synthetic offset would turn a validated mechanism into an earned guarantee. Near-miss decoys would test selection at its hardest, and a component ablation would show how the 8× divides between basis and update.

References

[1] A. Pun, K. Deng, R. Liu, D. Ramanan, C. Liu, and J.-Y. Zhu, "Generating physically stable and buildable brick structures from text," *arXiv preprint arXiv:2505.05469*, 2025.

[2] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on Socially Shared Cognition*, pp. 127–149, 1991.

[3] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.

[4] M. Purver, "The theory and use of clarification requests in dialogue," Ph.D. dissertation, King's College London, 2004.

[5] H. P. Grice, "Logic and conversation," *Syntax and Semantics: Speech Acts*, vol. 3, pp. 41–58, 1975.

[6] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," in *Conference on Robot Learning (CoRL)*, 2023.

[7] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei, "Learn then test: Calibrating predictive algorithms to achieve risk control," *arXiv preprint arXiv:2110.01052*, 2021.

[8] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[9] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 1507–1514, 2011.

[10] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone, and R. J. Mooney, "Jointly improving parsing and perception for natural language commands through human–robot dialog," *Journal of Artificial Intelligence Research (JAIR)*, vol. 67, pp. 327–374, 2020.

[11] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.

[12] C. Matuszek, "Grounded language learning: Where robotics and nlp meet," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[13] A. Wald, *Sequential Analysis*. John Wiley & Sons, 1947.

[14] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2012.

[15] A. Vlachos, "A stopping criterion for active learning," *Computer Speech & Language*, vol. 22, no. 3, pp. 295–312, 2008.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, 2016.

[17] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[18] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[19] R. A. Knepper, S. Tellex, A. Thomas, and N. Roy, "Recovering from failure by asking for help," *Autonomous Robots*, vol. 39, no. 3, pp. 347–362, 2015.

[20] S. Tellex, R. A. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics," in *Robotics: Science and Systems (RSS)*, 2014.

[21] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv preprint arXiv:2204.06125*, 2022.

[24] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[25] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[26] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman, "Emu edit: Precise image editing via recognition and generation tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[27] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," in *International Conference on Learning Representations (ICLR)*, 2023.

[28] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[29] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions," in *Robotics: Science and Systems (RSS)*, 2017.

[30] E. Bıyık and D. Sadigh, "Batch active preference-based learning of reward functions," in *Conference on Robot Learning (CoRL)*, 2018.

[31] S. Xie, L. Kong, Y. Zheng, Z. Tang, E. P. Xing,

G. Chen, and K. Zhang, "Learning vision and language concepts for controllable image generation," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 68 687–68 709.

[32] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[34] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Springer, 1998.

[35] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[36] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[37] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Freitas, "Bayesian optimization in a billion dimensions via random embeddings," *Journal of Artificial Intelligence Research*, vol. 55, pp. 361–387, 2016.

[38] D. Eriksson, M. Pearce, J. R. Gardner, R. D. Turner, and M. Poloczek, "Scalable global optimization via local bayesian optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[39] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[40] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.