SAY WHERE: LANGUAGE AND GOAL SETS FOR FROZEN VISUAL PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A task often accepts many outcomes, while a goal image depicts only one. We study an interface that preserves these alternatives for planning with a frozen visual world model. Task specifications define sets of acceptable states; finite clouds of their visual representations supply the planner's targets. This supports joint constraints, alternative destinations, precision, and factors left free by the task. Across four simulated environments, a shared-start comparison with the same nearest-goal cost increases average success from 31.1% with one accepted target to 52.3% with a large cloud. Targeted studies show that jointly valid targets improve simultaneous constraint satisfaction, and retaining alternatives improves attainment over committing to the initially nearest branch. We then learn to construct goal clouds from language through retrieval or generation. On the full endpoint suite, the retrieval interface achieves 46.3% success versus 31.6% for one accepted image, and is especially effective on region and property instructions. Set-agreement measurements distinguish successful specification from remaining grounding errors. These results connect richer task specifications to useful planning behavior without retraining the predictive dynamics, while identifying precise constraints and selection among valid alternatives as continuing challenges.

1 INTRODUCTION

"Put the block on the left, at any angle." This instruction specifies what must change and what may remain free. A goal image can show a successful outcome, but it also fixes an angle, a pusher position, and other details that the task does not require. The same distinction arises when an arm may reach anywhere in a region, a robot may choose either destination, or a cube must be lifted without specifying its horizontal location.

Visual world models make this distinction actionable: they predict the effects of candidate actions and support search toward encoded targets (Maes et al., 2026; Zhou et al., 2025). A single image provides one such target, but does not tell the planner which alternatives also count as success. For a conjunction, the target should satisfy both conditions together. For a disjunction, the planner should retain the option of reaching either outcome. These are properties of the task specification, even when the predictive dynamics remain unchanged.

Our key idea is to construct acceptable outcomes before planning, then let language specify this construction. A task defines a *goal set*; a finite cloud of its visual representations makes that set usable by a frozen planner (Figure 1). We call the resulting interface *Say Where*. GoalOracle constructs clouds using known state membership. GoalBank learns to retrieve accepted states from a bank, and GoalGen learns to generate goal representations directly from text. All reuse LeWorldModel's encoder, dynamics, and low-level action search (Maes et al., 2026).

The central finding is that richer specifications have measurable planning benefits, and language can make those benefits available from task descriptions. In a matched cloud-size comparison, average success increases from 31.1% to 52.3%. Targeted conjunction and disjunction studies connect this improvement to jointly valid targets and retained choices. On the separate full endpoint suite, GoalBank reaches 46.3% versus 31.6% for one accepted image. Its region/property mean matches the known-membership reference at 64.9%, although precise point and angle constraints remain harder to ground.

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.

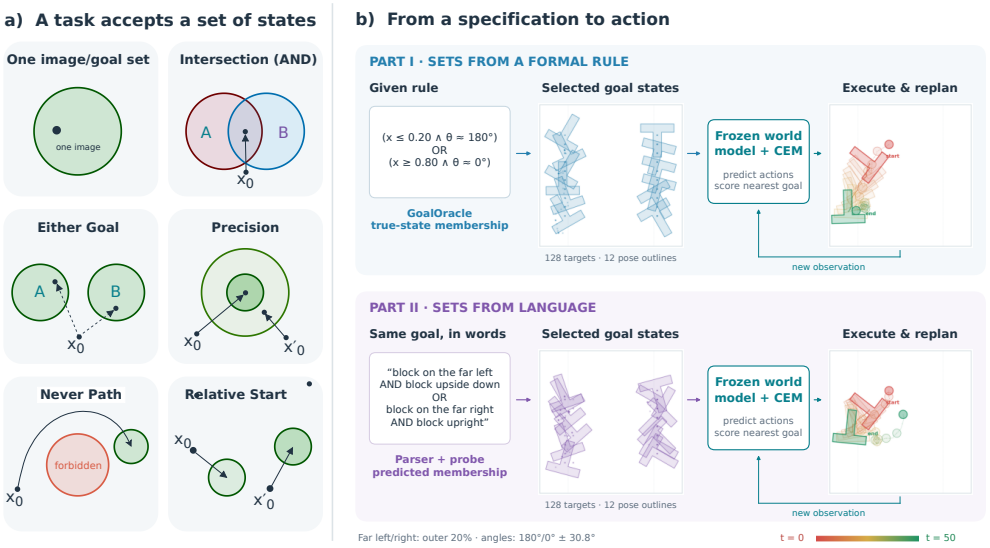

Figure 1: A task description preserves acceptable alternatives for the same frozen planner. Left: joint conditions, alternatives, precision, avoidance, and relative goals. Right: a supplied rule or a sentence constructs a goal cloud for the same frozen world model and cross-entropy action search (CEM). The language example uses a parser and physical-factor probe. Both recorded PushT trajectories succeed from the same start; poses run from red to green. This illustrative bank adds rendered poses to cover both alternatives and is separate from benchmark banks. Avoidance additionally requires a path filter.

Our contributions are:

1. A goal-set interface for joint constraints, alternatives, precision, and task freedoms in frozen visual planning (Section 2).
2. Four-environment evidence that preserving accepted alternatives improves attainment, with matched starts and planning cost in the cloud-size comparison (Section 4.1).
3. An analysis of composition's planning properties: intersection targets improve simultaneous satisfaction, and union targets improve attainment over nearest-branch commitment in targeted PushT studies (Section 4.2).
4. Two learned language interfaces that construct useful goal clouds, evaluated through physical success and accepted-set agreement (Section 5).

Goal-set formulations have established precedents in planning and offline goal-conditioned learning (Pavlasek et al., 2023; Vélez García et al., 2026). Our contribution is an interface and empirical analysis within frozen visual prediction: what its specifications enable, how they affect planning, and how successfully language constructs them.

2 From Task Descriptions to Visual Goal Sets

The interface connects a specification to the existing planner through acceptable target states. We first define those targets, then explain how known rules and learned language models construct them.

2.1 Specify What Counts as Success

An instruction defines a membership decision over physical states. Let x denote state, o its image, $z = E(o)$ its frozen encoding, and x_0 the initial state. An acceptance predicate A defines

$$\mathcal{G}(x_0) = \{x : A(x; x_0) = 1\}. \tag{1}$$

A region predicate admits many positions; a precision predicate sets a tolerance around a requested value. A factor absent from A is unconstrained. A relative command, such as "lift higher," evaluates its target against x_0.

Composition determines which outcomes the planner may pursue. For primitive goal sets $\mathcal{G}_A$ and $\mathcal{G}_B$,

$$\mathcal{G}_{A\cap B} = \mathcal{G}_A \cap \mathcal{G}_B, \qquad \mathcal{G}_{A\cup B} = \mathcal{G}_A \cup \mathcal{G}_B. \tag{2}$$

Intersection therefore requires both predicates in one state; union preserves either destination. The interface implements this logic explicitly. A single image may depict a member of either set, but does not communicate the membership rule or the factors left free.

2.2 Give the frozen planner a cloud of targets

We approximate a goal set with a finite *goal cloud* $\mathcal{C}$ of visual latents. An offline bank $\mathcal{D} = \{(z_i, f_i)\}_{i=1}^{N}$ stores image encodings and physical factors. Given membership scores, we retain accepted bank states and select up to K spread-out latents by farthest-point sampling. GoalOracle evaluates A on true factors f_i; GoalOracle($K = 1$) supplies one randomly selected accepted image. True membership is privileged information shared by this entire reference family. Exact bank membership does not establish reachability from the current start.

Planning searches for an action sequence whose predicted endpoint approaches the cloud. For candidate actions a over prediction horizon H, let $\hat{z}_{t+H}$ be the terminal encoding predicted from time t. The cost is

$$c(a_{t:t+H-1}) = d_{(k)}(\hat{z}_{t+H}, \mathcal{C}), \qquad d_{(k)}(z, \mathcal{C}) = \text{kth}\min_{v\in\mathcal{C}} \|z - v\|_2^2. \tag{3}$$

For $k = 1$, a candidate can approach any cloud member. A larger rank requires nearby support from more members. The cloud cap K and distance rank k have different roles. The size study holds $k = 1$ fixed; the full endpoint suite retains $k = 1$ for one/ten images and $k = 4$ for large-cloud interfaces. These settings followed exploratory sweeps on reused task families; Appendix F reports the cost and selection controls.

We retain the released cross-entropy method (CEM) action search and frozen encoder and predictor. In the quantitative control studies, each plan spans five action blocks of five simulator steps. The controller executes 25 simulator steps before replanning, with a total budget of 50. Goal construction changes the objective passed to this search; it does not retrain predictive dynamics.

2.3 Learn to construct clouds from language

Language lets the user specify acceptable outcomes without supplying goal images or writing a predicate. For "block on the left, any orientation," the desired output is a collection of left-side states spanning orientation. We implement two direct language routes to this collection.

GoalBank: predict membership and retrieve. A text Transformer conditions a scorer over bank latents. We train it on simulator-derived membership labels and an auxiliary distance target, using binary cross-entropy plus weighted Huber loss. At inference, text and stored latents suffice: no parsed rule or physical-factor decoder is needed. States with membership probability at least 0.5 are retained, with a top-16 fallback for sparse predictions; farthest-point sampling then forms the cloud. Only the membership output selects goals.

GoalGen: sample goal representations. A separate text encoder conditions a flow-matching model trained on accepted latents (Lipman et al., 2023). Sampling integrates a learned velocity field from noise to goal representations, using 40 steps. This route requires no retrieval bank at inference, although generated latents need not correspond to real states. Both interfaces use simulator supervision during training and feature-wise conditioning (Perez et al., 2018); architectures and optimisation are detailed in Appendix D.

Rules and parsers provide intermediate comparisons. Given rule + probe applies the correct predicate to physical factors predicted by a fitted linear decoder. Parser + probe first recovers a rule from text using a template grammar, a scratch Transformer, Contrastive Language–Image Pretraining

(CLIP) features with a trained decoder (Radford et al., 2021), or prompted Llama (Grattafiori et al., 2024). These comparisons help distinguish set construction from execution. They do not form an additive error decomposition.

Relative and avoidance extensions. The same interface supports start-relative tasks and forbidden regions. Relative learned interfaces receive the initial latent; symbolic interfaces receive true initial factors. For Never, a separate forbidden cloud and candidate-path filter check avoidance through attainment. This filter is used by the rule-based extension, but not by the evaluated GoalBank controller. Appendices G and H define these extensions and report their separate evaluations.

3 Evaluation setup

We separate the usefulness of accepted targets from the difficulty of constructing them from language. We examine whether alternatives improve attainment, how intersection and union affect planning, and how much of the benefit language recovers.

Environments and populations. We used released LeWorldModel checkpoints for PushT, Reacher, TwoRoom, and OGBench-Cube (Maes et al., 2026; Park et al., 2025), covering pushing, reaching, navigation, and manipulation. The endpoint suite contains 20 instructions per environment and 100 starts per instruction. The size study uses a fixed 20-start subset of every instruction. The targeted intersection and union, paraphrase, relative, and avoidance studies have separate populations; their rates are not pooled (Appendix A).

Physical grade and information. Success means the intended predicate holds at any post-action simulator step; intersection requires simultaneous satisfaction. Grades use true physical factors independently of predicted membership. Tight, medium, and loose tolerances are 1.5, 3, and 6 times held-out probe error. Image references receive true accepted members, GoalOracle receives the rule and true factors, and language interfaces receive text. The supplied-rule + probe comparison measures set fidelity but has no endpoint control run. Thus language versus image comparisons evaluate complete interfaces, including their recorded cost ranks.

Data and uncertainty. Each bank contains 100,000 frames from the same source dataset as evaluation starts. The current episode is excluded from retrieval, but may have appeared in interface training. A separate PushT configuration-disjoint check tests interface transfer; backbone pretraining exposure remains unresolved (Appendices B and B.1). Each direct reader has one training seed. We weight instruction cells equally and count missing outputs as failures. Principal control intervals use paired physical-cluster resampling: configurations in PushT, episodes elsewhere, with the instruction suite fixed. Superscript margins enclose 95% percentile intervals and are not standard deviations. They do not measure variation across new tasks or training seeds. Appendix P gives task-resampling sensitivity and the distinct intersection and union procedures.

4 Goal sets improve attainment and change planning choices

4.1 Preserving acceptable alternatives improves control

We first examine whether more acceptable targets help planning. With known membership, increasing the cloud cap from one to 1,024 improves average success from 31.1% to 52.3% (Figure 2). Every size uses the same 1600 starts, predicates, interaction budget, and nearest-member cost. This gain requires no language learning or world-model retraining.

The benefit depends on the environment and construction. Reacher gains most of its improvement by 32 members, while PushT benefits more at larger caps. The nested 32/128 comparison gains 2.0 percentage points on average, with a positive paired interval. Across the full curve, historical selection and planning randomness differ; larger clouds do not improve every task monotonically. Appendix R gives the matched subset and its controls, and Appendix E separates task families.

These results support preserving acceptable alternatives as a useful task interface. They establish a comparison of implemented goal constructions, not a universal cardinality law. A separate start-aware

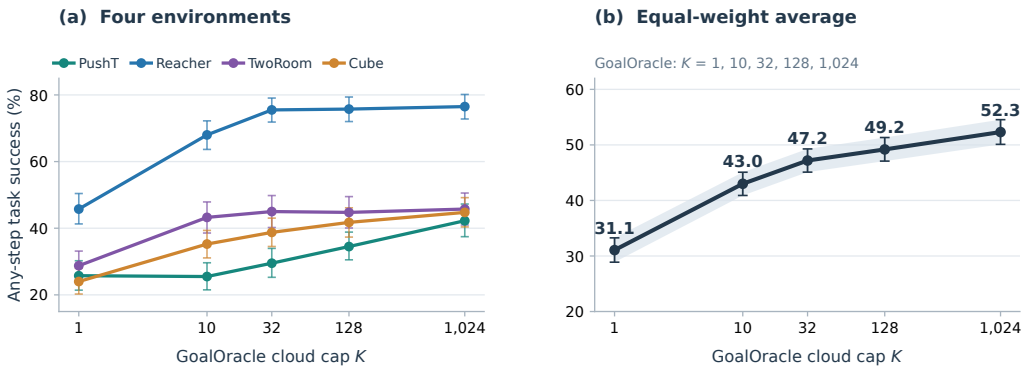

Figure 2: Preserving more accepted alternatives improves average attainment. Left: GoalOracle with different cloud caps K in each environment. Right: the equal-weight average. Each size uses the same 1600 cases and nearest-member cost; bars show fixed-task 95% intervals. The one-member target is random, larger clouds use spread selection, and only the 32/128 runs share planning draws and nested clouds. This compares goal constructions, with the nested pair providing the stricter size contrast.

nearest-image control does not recover the set benefit (Appendix B.1), leaving stronger single-goal selectors as meaningful alternatives.

4.2 Composition gives the planner joint targets and alternative destinations

We next examine how composition changes the planner's targets. Intersection restricts targets to jointly valid outcomes; union lets the search approach either branch. We test these properties against simpler objectives under matched action-search and interaction budgets.

Intersection: aim at a state satisfying both conditions. On deliberately easier positional conjunctions, an intersection cloud succeeds in 523/1024 runs (51.1%), compared with 37.3% for summed operand costs and 37.5% for their maximum (Table 1). We select eight of sixteen candidate formulas using historical average success across all three objectives, then evaluate on 512 fresh task–start pairs with two planning seeds each. Selection is fixed before evaluation and does not use the intersection advantage. Both paired gains have positive configuration-cluster intervals. The separate costs may refer to different valid examples for each operand; intersection directly supplies jointly valid targets. These results concern the selected positional suite, with the original four-formula pilot retained separately (Appendices I and K). They measure simultaneous attainment, not sequential execution or shorter paths.

Union: retain choices instead of committing by proximity. Union planning reaches 73.8% success versus 64.8% for initially nearest-branch commitment. The paired gain is 8.92 percentage points, with interval $[5.57, 12.30]$. Keeping alternatives therefore helps relative to a geometric commitment heuristic. A planning-based selector that compares predicted branch costs reaches 79.2%, improving on union. Union versus random commitment remains unresolved in success; full controls are in Appendix J.

Independently calibrated branch difficulty clarifies the remaining choice problem. Union targets the easier branch 96.4% of the time when it is nearer, but 44.1% when it is farther. Retained alternatives improve attainment over nearest commitment without establishing reliable difficulty-aware selection. Together, the intersection and union studies show useful planning consequences of specification structure within their tested regimes; matching candidate budgets does not imply equal runtime or path length.

Figure 3 connects these set constructions to recorded images and executed paths in three other environments. It shows the progression from one satisfied condition to joint satisfaction, the alternatives available to union, and attainment of a permitted goal under Never.

Table 1: Compositional targets improve attainment over specific simpler objectives. Left: eight calibration-selected easier PushT conjunctions, 64 fresh starts and two planning seeds each (1024 runs per objective), graded on simultaneous satisfaction. Right: four disjunctions, 400 starts and three evaluation seeds per start. All methods evaluate 18,000 action candidates within 50 simulator steps. Margins are 95% intervals under the study-specific resampling schemes. Nearest and random commitment are derived from matched fixed-branch runs. The planning-based selector is included as a stronger union control.

Intersection objective	Success (%) ↑
Sum of costs	$37.3^{\pm 6.4}$
Maximum of costs	$37.5^{\pm 6.2}$
Intersection cloud	$51.1^{\pm 5.9}$

Union objective	Success (%) ↑
Nearest commitment	$64.8^{\pm 4.3}$
Random commitment	$70.1^{\pm 3.7}$
Union cloud	$73.8^{\pm 4.1}$
Planning-based selector	$79.2^{\pm 3.8}$

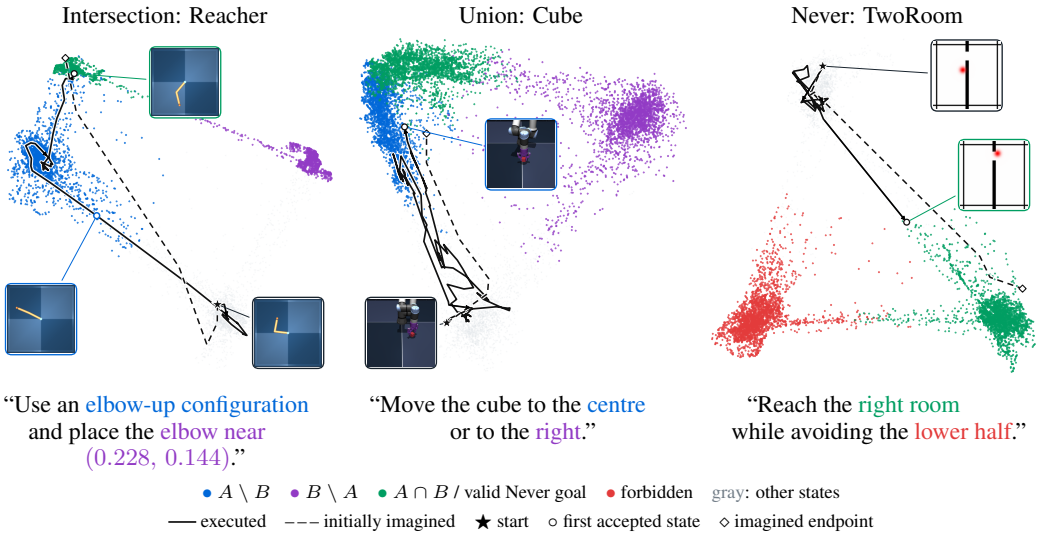

Figure 3: Composed targets and recorded planning trajectories. Instructions summarize the supplied rules. Bubbles show recorded frames along the executed path, from each start to its first accepted state. Reacher progresses from the blue-only condition to the green joint target. Cube can target blue, violet, or green states; the endpoint shows its first accepted state. TwoRoom reaches the green goal while avoiding the red region. Solid black lines show execution and dashed black lines the initial forecast. All points share an oracle-supervised, task-conditioned projection. These selected examples use separate budgets (Cube first succeeds at step 83); they are excluded from the 50-step success rates. Details are in Appendix L.

5 Language makes goal-set planning usable from task descriptions

5.1 Descriptions recover a substantial control benefit

We next ask whether language can recover this benefit. GoalBank reaches 46.3% endpoint success and GoalGen reaches 43.1%, compared with 31.6% for one accepted image (Table 2). GoalBank improves by 14.7 percentage points, with a positive paired interval. These learned interfaces turn a sentence into visual targets for the same frozen dynamics.

The practical capability is specifying acceptable outcomes directly in words. GoalBank's mean lies near that of ten accepted images and the strongest parser pipelines, while requiring the user to supply neither those images nor an explicit predicate. Their pairwise success differences remain unresolved, so the evidence establishes neither superiority nor equivalence. GoalGen supplies a complementary bank-free inference route, with lower paired success than GoalBank in this suite. Reduced user effort is a motivation for the interface, not a measured usability result.

Table 2: Language-specified goal clouds recover much of the endpoint control benefit. Success (%) over 80 instructions and 8000 starts per interface. One/ten-image references use true accepted states and nearest-member cost; large-cloud interfaces use fourth-nearest cost. GoalOracle is a true-factor reference. Average margins are fixed-task 95% intervals; no numerical winners are bolded because paired comparisons, information access, and cost settings govern interpretation.

Goal interface	PushT	Reacher	TwoRoom	Cube	Avg. ↑
One accepted image	26.8	43.4	30.9	25.3	$31.6^{\pm 1.1}$
Ten accepted images	33.9	69.5	41.7	37.3	$45.6^{\pm 1.1}$
GoalBank	41.8	70.3	30.9	42.3	$46.3^{\pm 1.0}$
GoalGen	32.8	70.9	34.7	34.1	$43.1^{\pm 1.1}$
Template + probe	33.2	58.0	35.7	36.9	$40.9^{\pm 1.0}$
Scratch parser + probe	37.7	46.7	33.4	31.6	$37.3^{\pm 1.0}$
CLIP parser + probe	38.9	59.0	41.3	43.4	$45.6^{\pm 1.1}$
Llama parser + probe	37.9	63.6	40.7	43.9	$46.5^{\pm 1.1}$
GoalOracle (true factors)	44.4	73.7	44.0	45.7	$51.9^{\pm 1.1}$

5.2 Language preserves task freedoms and recovers accepted regions

Language can identify an entire family of valid outcomes, including factors that need not be controlled. For "put the block on the left half, any orientation," GoalBank retrieves states spanning orientation, with 99.9% intersection-over-union (IoU) against the true accepted set in that example. Figure 4 also shows alternative regions in TwoRoom and lifted Cube states with horizontal position free. These are selected demonstrations of membership; they are not estimates of average control success. Appendix Q includes physical block silhouettes and full-bank reconstruction details.

Region, property, and proximity instructions form a clear strength in the aggregate evaluation. GoalBank attains 64.9% success, the same reported mean as GoalOracle, and 91.7% accepted-set IoU on this group (Table 3). This agreement in mean control does not imply equality on every task. It shows that language can preserve useful task freedoms with high set fidelity in an identifiable regime.

Point and angle constraints leave a larger grounding gap. GoalBank reaches 35.1% success against 44.1% for GoalOracle, with 50.9% set IoU. The group includes qualitative angles as well as numerical requests. Its contrast with region/property tasks localizes a weakness, but cannot attribute it solely to numerical wording rather than precision, geometry, composition, or environment mix.

5.3 Set fidelity identifies where specification needs improvement

Measuring membership before control distinguishes an inaccurate target set from difficulty executing a valid target. On all endpoint instructions, GoalBank reaches 66.2% IoU and Llama + probe reaches 68.8%. Even the supplied rule with the physical-factor probe reaches only 77.6%, showing that state estimation can alter membership before parsing. These comparisons diagnose different constructions, not additive causal losses (Table S23). GoalGen has no directly comparable bank-membership mask.

Language variation and held-out configurations provide additional checks on scope. On a separate matched-meaning protocol, GoalBank retains 70.7% IoU after rephrasing, compared with 90.6% for the original wording (Table S18). On the configuration-disjoint PushT check, its success is 35.0%, versus 27.2% for one image and 36.6% for ten. The paired advantage over one image is unresolved on that split. Precise hindsight-derived goals provide a further stress test (Appendix C); relative and avoidance comparisons retain distinct information and mechanisms.

Together, these results show that language constructs useful goal sets and recovers much of their control benefit, with particularly strong region/property results. Its evaluated vocabulary and supervision remain grounded in simulator predicates. This supports task-description interfaces in the tested domains, while motivating better precision, independently reviewed instructions, and broader transfer tests.

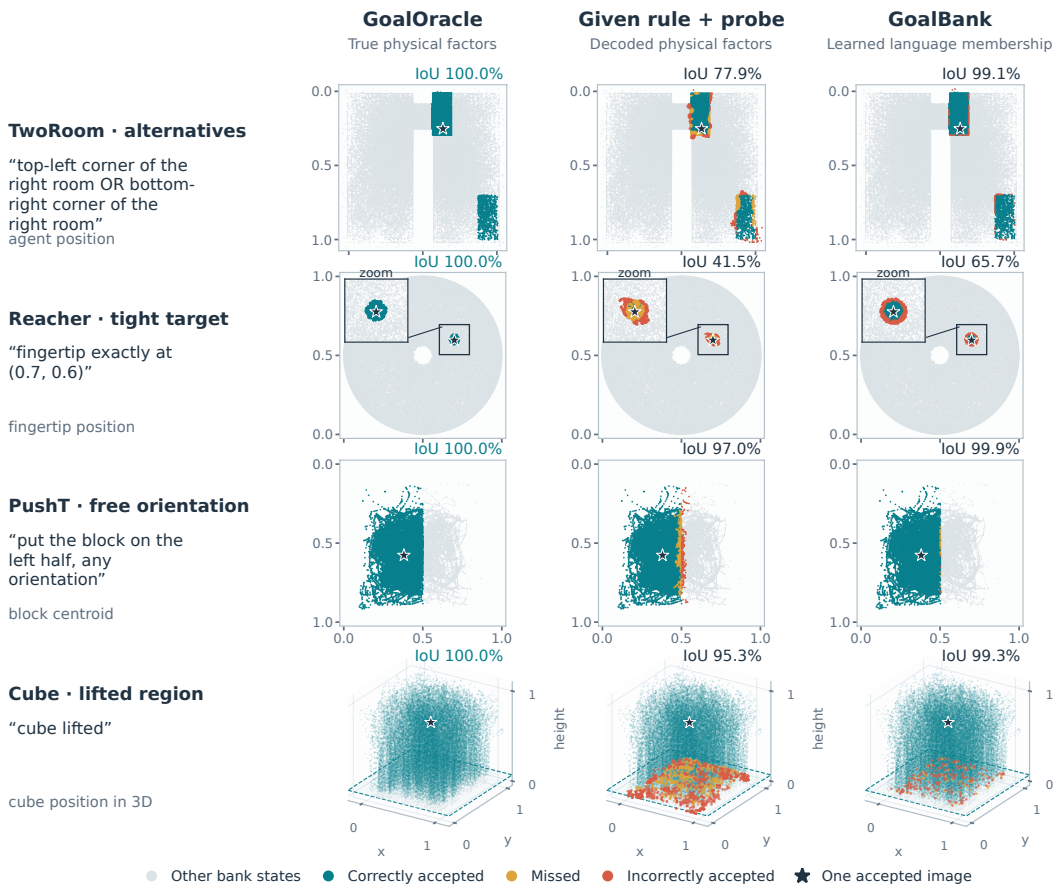

Figure 4: Language recovers alternative regions and task freedoms. All bank states are shown at true physical coordinates, colored by acceptance relative to the intended rule. Columns compare true-factor membership, the supplied rule on probe-decoded factors, and learned GoalBank membership. The examples retain either destination, a tight fingertip target, free block orientation, and free horizontal position while lifted. Stars mark valid image targets. IoU uses full-bank masks; panels are selected examples, not average performance. Cube uses true 3D coordinates.

Table 3: Region/property instructions are a strength of learned goal construction. Endpoint success (%) on 30 region/property/proximity cells and 50 containing point/angle predicates. Superscripts are fixed-task 95% margins. The groups contain different tasks and environment proportions; this is not a matched wording or tolerance ablation. Full interface and IoU breakdowns appear in Appendix Q.

Goal interface	Region/property ↑	Point/angle ↑
One accepted image	$39.6^{\pm 1.7}$	$26.8^{\pm 1.3}$
Ten accepted images	$56.4^{\pm 1.8}$	$39.1^{\pm 1.4}$
GoalBank	$64.9^{\pm 1.7}$	$35.1^{\pm 1.3}$
GoalGen	$59.6^{\pm 1.7}$	$33.2^{\pm 1.3}$
GoalOracle	$64.9^{\pm 1.6}$	$44.1^{\pm 1.4}$

6 Related Work

Planning to acceptable outcomes. Ready, Set, Plan! plans directly to sampled goal configurations using generalized Bayesian inference (Pavlasek et al., 2023). GS-HER uses queries to select which state variables define a goal in offline goal-conditioned learning (Vélez García et al., 2026). VICE learns control objectives from successful events (Fu et al., 2018). These precedents establish the

value of acceptable sets. We study their construction and composition inside a frozen visual predictor, linking target membership, planning behavior, and language specification in the same evaluation.

Visual world models and latent goals. LeWorldModel and DINO-WM support planning in learned visual spaces (Maes et al., 2026; Zhou et al., 2025). PiJEPA uses policy guidance for action search (Chahe & Zhou, 2026), while FF-JEPA predicts latent subgoals (Masip et al., 2026). Our evaluated system reuses LeWorldModel's predictive dynamics and action search, changing the acceptable targets supplied to that search. The evidence covers this model family rather than a comparison of world-model architectures.

Language-conditioned latent goals. LAGO (Barbeau et al., 2026) predicts intermediate latent goals from the current observation, a language instruction, and task-completion values. It refreshes these goals during planning and scores candidate trajectories by their alignment with the predicted subgoals. Its shared predictor supports both action-conditioned transitions and language-conditioned subgoals. Say Where also maps language into visual latent targets, but preserves the existing frozen dynamics and learns separate goal-construction interfaces. Our cloud represents an unordered collection of acceptable outcomes, rather than intermediate stages of task completion. Intersection, union, and unconstrained factors determine which outcomes belong to that collection. Thus LAGO provides guidance toward completing a task, while our interface preserves the choice among outcomes that satisfy it. Combining intermediate guidance with accepted-outcome sets is a possible extension.

Language as a control interface. SayCan grounds instructions through skill affordances (Ahn et al., 2023); Code as Policies and Language to Rewards translate language into programs or objectives (Liang et al., 2023; Yu et al., 2023). VoxPoser constructs composable spatial value maps (Huang et al., 2023), while natural-language-to-temporal-logic translation provides structured task specifications (Pan et al., 2023). VIP learns visual representations and rewards (Ma et al., 2023b), while LIV learns representations and rewards for language and image goals (Ma et al., 2023a). We evaluate language through the accepted outcomes it constructs, measuring set agreement separately from downstream control and comparing direct latent construction with rule/probe routes.

7 Limitations and Conclusion

The evidence is specific to one world-model family, four simulated environments, and single reader-training seeds. Historical banks and evaluation data overlap; the PushT split tests interface transfer with pretraining exposure unresolved. GoalOracle uses true factors, and supplied-rule + probe endpoint control is missing. Full-suite interfaces differ in cost rank; the shared-cost size curve retains selection and randomness differences. Intersection findings condition on deliberately easier positional formulas; union uses four fixed formulas. Relative interfaces have unequal initial-state information, and Never adds a filter. Template-derived instructions and unreviewed paraphrases limit claims about human language. The paired uncertainty estimates quantify the tested comparisons, not these broader sources of variation.

Structured goal sets expand the task interface of a frozen visual planner and improve attainment in the tested settings. Their capabilities have practical consequences: joint targets help satisfy conditions together, retained alternatives improve over nearest-branch commitment, and language constructs useful clouds while preserving task freedoms. The next step is to improve both sides of this interface: recover precise membership from independently authored instructions, and use predicted dynamics to choose well among valid outcomes. Keeping these questions separate makes the interface easier to evaluate and improve.

AI USE STATEMENT

Generative AI assisted the development and refinement of experimental protocols, implementation and checking of code, interpretation and inspection of experimental artifacts, literature and citation checks, and manuscript and figure preparation. Language models are also experimental components: Llama-3-8B-Instruct supplies a few-shot parser and Qwen2.5-7B-Instruct supplies synthetic paraphrases. Those paraphrases were automatically accepted and are not human annotations. Reported numerical results derive from stored experiment records; generated numerical outcomes do not substitute for simulator measurements. Illustrative environment images are dataset images, simulator renders, or recorded trajectories. Human authors remain responsible for the analysis, claims, citations, and final submission.

REPRODUCIBILITY AND ARTIFACT AVAILABILITY STATEMENT

The supplementary package supplied for review contains a separate appendix PDF, the frozen numerical evidence needed to inspect the reported results, selected figure records, a claim-to-evidence guide, and a CPU-only consistency checker. All appendix references in this paper refer to that PDF. Appendix A defines populations, denominators, and missing-output treatment; Appendices B and B.1 describe overlap and configuration separation; Appendix D gives model and optimisation details; Appendix P specifies the uncertainty procedures. The accompanying records support checking reported calculations and provenance; they do not constitute a complete rerunnable experiment distribution.

Upon acceptance, we will make the full research release available with the final submission, or provide an anonymous link during the rebuttal period. The release will include manuscript and figure sources, table and numerical-asset generators, analysis and evaluation code, task and start manifests, instruction and parser records, model and planner configurations, training and checkpoint-loading instructions, experiment ledgers and trajectory records, provenance hashes, and instructions for obtaining the required datasets and pretrained checkpoints. These broader artifacts are a future release commitment and are not included in the present supplementary package. Original experimental reproduction requires those research artifacts, datasets, and model checkpoints; the included consistency checks require no simulator or GPU.

ETHICS STATEMENT

The experiments use simulated environments and do not constitute a human-subject study. Synthetic language is labelled by its provenance. The avoidance mechanism is a heuristic controller component rather than a verified safety system, and the results do not justify safety-critical deployment.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Samuel Barbeau, Simon Roy, Giovanni Beltrame, Christian Desrosiers, and Nicolas Thome. Latent Goal Prediction from Language for Model-Based Planning. *arXiv preprint arXiv:2606.20627*, 2026. URL https://arxiv.org/abs/2606.20627.

Amirhosein Chahe and Lifeng Zhou. Policy-Guided World Model Planning for Language-Conditioned Visual Navigation. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR) Workshops, pp. 834–841, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/WDFM-EAI/html/Chahe_Policy-Guided_World_Model_Planning_for_Language-Conditioned_Visual_Navigation_CVPRW_2026_paper.html.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational Inverse Control with Events: A General Framework for Data-Driven Reward Definition. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c9319967c038f9b923068dabdf60cfe3-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562, 2023. URL https://proceedings.mlr.press/v229/huang23b.html.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9493–9500, 2023. doi: 10.1109/ICRA48891.2023.10160591. URL https://doi.org/10.1109/ICRA48891.2023.10160591.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. LIV: Language-Image Representations and Rewards for Robotic Control. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23301–23320, 2023a. URL https://proceedings.mlr.press/v202/ma23b.html.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=YJ7o2wetJ2.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels. *arXiv preprint arXiv:2603.19312*, 2026. URL https://arxiv.org/abs/2603.19312.

Sergi Masip, Jonathan Swinnen, Yutong Hu, Renaud Detry, and Tinne Tuytelaars. FF-JEPA: Long-horizon planning in world models with latent planners. *arXiv preprint arXiv:2606.09311*, 2026. URL https://arxiv.org/abs/2606.09311.

Jiayi Pan, Glen Chou, and Dmitry Berenson. Data-Efficient Learning of Natural Language to Linear Temporal Logic Translators for Robot Task Specification. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11554–11561, 2023. doi: 10.1109/ICRA48891.2023.10161125. URL https://doi.org/10.1109/ICRA48891.2023.10161125.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking Offline Goal-Conditioned RL. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ecd92623ac899357312aaa8915853699-Abstract-Conference.html.

Jana Pavlasek, Stanley Robert Lewis, Balakumar Sundaralingam, Fabio Ramos, and Tucker Hermans. Ready, Set, Plan! Planning to Goal Sets Using Generalized Bayesian Inference. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3672–3686, 2023. URL https://proceedings.mlr.press/v229/pavlasek23a.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Carlos Vélez García, Miguel Cazorla, and Jorge Pomares. Goal Sets, Not Goal States: Queryable Robot Goals through Goal-Set Hindsight Relabeling. *arXiv preprint arXiv:2606.09476*, 2026. URL https://arxiv.org/abs/2606.09476.

Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to Rewards for Robotic Skill Synthesis. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 374–404, 2023. URL https://proceedings.mlr.press/v229/yu23a.html.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.