ANVIL: SCALING EMBODIED SPATIAL INTELLIGENCE THROUGH KNOWLEDGE RECIRCULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied spatial intelligence requires more than visual recognition: models must localize entities, reason about their relations, and express this knowledge in action-relevant forms. Yet joint training can conflate the knowledge needed at different stages because supervision is fragmented across incompatible geometric and semantic outputs. We introduce ANVIL, a unified framework for data generation, training, and benchmarking for embodied spatial intelligence. ❶ANVIL-GEN is a heterogeneous data flywheel that combines form-specific generation, cross-stage consistency verification, and failure-driven recollection across 11 Ground–Relation–Act forms. ❷ANVIL-BEN provides scene-aligned supervision for learning diverse embodied spatial capabilities and scene-linked assessments that distinguish stage competence from error propagation. ❸ANVIL-MOPD is a 9B student that consolidates heterogeneous spatial knowledge through distillation from specialist teachers. Semantic-first-fault routing targets teacher supervision to the student's earliest decisive failure, while stage normalization controls the loss scale across native outputs of different lengths. We generate and verify 5.8M records spanning 160K images, reserving 580K verified spatial records for held-out evaluation with human review and using the remaining 5.22M records for training. On average across seven public benchmarks, ANVIL-MOPD outperforms Qwen3.5-9B and RynnBrain-8B by 9.29 and 11.03 percentage points, respectively, while reaching 80.67% on VABench-Point. Controlled studies support targeted coverage, selective transfer, and feedback-guided repair as complementary framework components.

1 INTRODUCTION

Embodied spatial intelligence requires more than visual recognition: a model must localize entities, reason about relations, and identify possible interactions (Yuan et al., 2025; Zhou et al., 2025; Li et al., 2026c). Recognizing a coffee machine, for example, does not determine where a cup can be placed beside it. Learning these capabilities requires supervision that connects visual observations to precise, action-relevant spatial judgments. Web images, reconstructed scenes, and simulated environments offer complementary sources, but differ in their coverage and available geometry (Song et al., 2024; Chen et al., 2024; Zhi et al., 2025). *Scaling embodied spatial intelligence therefore requires both broader supervision and a way to consolidate the knowledge expressed in different output forms.*

Recent data-driven approaches have advanced spatial grounding and reasoning through synthetic supervision, geometric reconstruction, and capability-diverse training (Chen et al., 2024; Zhou et al., 2025; Cai et al., 2026). Broader multimodal and embodied training also demonstrated the value of integrating diverse visual experience and tasks (Tong et al., 2024; Liu et al., 2025; 2026c; Zhang et al., 2026a; Tan et al., 2026; Li et al., 2026c). These studies support scaling spatial supervision, although gains vary across capabilities and transfer settings (Cai et al., 2026; Li et al., 2026c). Diagnostic benchmarks also revealed differences that aggregate scores can obscure (Du et al., 2024; Wu et al., 2025; Zhang et al., 2026e). Building on these advances, we investigate a data-centered question:

> *How can heterogeneous spatial expertise be turned into reliable supervision and consolidated in a model that also supports subsequent data verification?*

We introduce ANVIL, which connects a spatial data flywheel and model training through **knowledge recirculation** (Figure 1): ANVIL-GEN operates as a heterogeneous data flywheel, where form-specific

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.

generation produces spatial supervision, cross-stage verification identifies inconsistencies, and failure feedback guides targeted data regeneration and collection. A shared Ground–Relation–Act interface connects 11 native output forms through common entities and explicit dependencies, allowing the same capabilities to be specified across generation, learning, and evaluation. ANVIL-BEN organizes the aligned supervision into ANVIL-TRAIN for learning and ANVIL-EVAL for held-out assessment through native-output and scene-linked tasks. ANVIL-MOPD integrates knowledge from specialists trained on this supervision through multi-teacher distillation, building on on-policy knowledge transfer (Agarwal et al., 2024; Ma et al., 2026a; Gao et al., 2026). Semantic-first-fault routing directs teacher supervision to the earliest decisive student error, while stage normalization controls the loss scale across output lengths. Validated records can refresh specialists, while the distilled generalist returns verification signals to the data flywheel. Our contributions are threefold:

(i) ANVIL-GEN: a scene-aligned, cross-verified data flywheel. Shared scene and entity references align source annotations and specialist outputs, while geometry-aware CoT structures schema-constrained generation. Cross-form adjudication turns dependencies among outputs into explicit verification constraints, and failure feedback guides targeted repair across sources and forms.

(ii) ANVIL-BEN: a scene-aligned benchmark for spatial learning and evaluation. Aligned scene images and 11 native annotation formats support training in ANVIL-TRAIN and held-out assessment in ANVIL-EVAL, covering spatial understanding, reasoning, and state-conditioned plan verification. Native-output tasks measure individual capabilities; linked assessments and controlled upstream substitutions measure how errors propagate across dependent stages.

(iii) ANVIL-MOPD: specialist-to-generalist transfer with multi-teacher distillation. Semantic-first-fault routing jointly selects a specialist and correction span, while stage normalization balances supervision across heterogeneous outputs. The resulting generalist can assist ANVIL-GEN under evidence-backed verification. Benchmark comparisons and controlled studies demonstrate the benefits of structured supervision and selective specialist transfer.

2 RELATED WORK

Spatial intelligence. Spatial intelligence research spans relation recognition, geometric estimation, and reasoning about possible interactions. Early evaluations of relational understanding (Liu et al., 2023; Kamath et al., 2023) have been complemented by image-based, multi-view, and video-based benchmarks (Tong et al., 2024; Du et al., 2024; Fu et al., 2024; Yang et al., 2025b), with further emphasis on transformations, planning, and spatial-functional reasoning (Zhang et al., 2026e; Wu et al., 2025; Zhang et al., 2026b;c). Training has likewise moved beyond image–text pairs toward supervision derived from geometry and robot experience (Chen et al., 2024; Cheng et al., 2024; Chen et al., 2025), including precise, interaction-relevant grounding (Yuan et al., 2025; Zhou et al., 2025). Procedural geometry and executable representations broaden the available targets (Assouel et al., 2026; Liu et al., 2026a), while program execution and trajectory consistency help check automatically generated annotations (Zhi et al., 2025; Hu et al., 2024; Blank et al., 2026). This expanding coverage raises a broader evaluation question: *when does success on a particular spatial task indicate competence that transfers to other tasks?*

Embodied foundation models. Large-scale pretraining has connected visual–linguistic knowledge to embodied reasoning and robot control (Driess et al., 2023; Zitkovich et al., 2023; Kim et al., 2025). Autoregressive and diffusion policies provide alternative action interfaces, with generalist policies supporting adaptation across robot platforms (Octo Model Team et al., 2024). In parallel, spatially focused models develop geometric and spatiotemporal understanding (Cai et al., 2026; Tan et al., 2026; Li et al., 2026c), using structured scene representations, targeted supervision, or progressive curricula (Ju et al., 2026; Ray et al., 2025; Li et al., 2026b). Reasoning–action models increasingly connect these capabilities to action prediction (Liu et al., 2025; 2026c). *The challenge is to turn diverse training signals into generalizable capabilities, rather than a collection of task-specific skills.*

Knowledge distillation. Distillation transfers teacher knowledge through output distributions or generated sequences (Hinton et al., 2015; Kim & Rush, 2016); subsequent language-model methods address student behavior through reverse-KL and on-policy learning (Gu et al., 2024; Agarwal et al., 2024). Multi-teacher extensions introduce choices about both knowledge sources and supervision

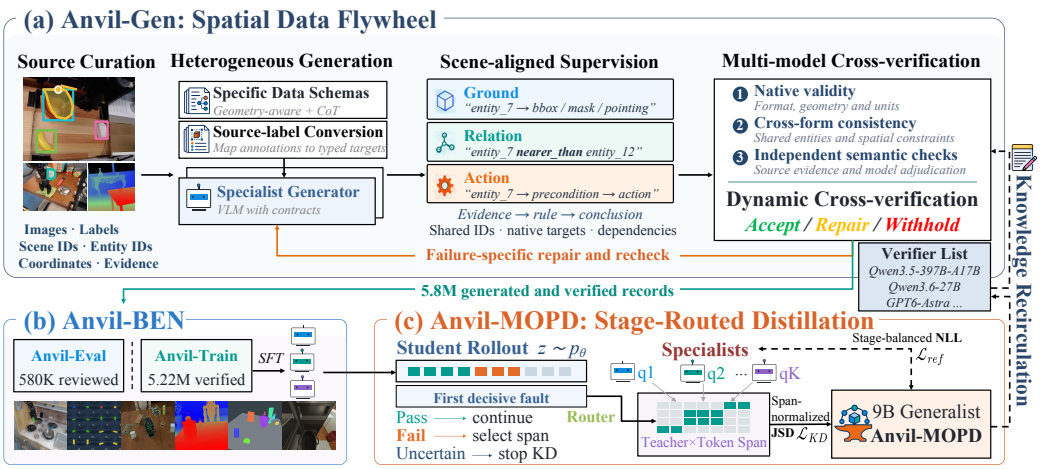


Figure 1: **Overview of the Anvil framework.** Anvil-Gen produces spatial supervision through generation, cross-verification, and repair. Anvil-Ben curates verified records into training and held-out evaluation sets. Anvil-MOPD consolidates specialist knowledge into a generalist through teacher-and-span routing on student rollouts. The generalist then assists subsequent verification, returning learned knowledge to supervision production while evidence-backed checks retain acceptance authority.

budgets (Ma et al., 2026a; Gao et al., 2026). Selection can operate over examples and tasks through rationale consolidation, answer verification, or privileged context (Jin et al., 2026; He et al., 2026; Ma et al., 2026b), or within trajectories through discrepancy-triggered intervention and failed-prefix recovery (Koo et al., 2025; Xu et al., 2026). Related approaches target informative tokens (Liu et al., 2026b; Li et al., 2026a), cross-view representations (Nguyen et al., 2026), and intermediate reasoning steps (Lightman et al., 2024), while iterative bootstrapping reuses answer-checked model outputs for further training (Zelikman et al., 2022). Selective supervision thus raises a question of reliability: *how can useful teacher knowledge be retained without also reinforcing its errors?*

To address these gaps, ANVIL treats supervision reliability and knowledge transfer as coupled problems. Scene-aligned annotations support specialist-to-generalist distillation and held-out evaluation through knowledge recirculation; the resulting student can then assist further data verification. The knowledge recirculation mechanism links data refinement to model learning so that their effects on spatial competence can be studied together.

3 ANVIL: LEARNING THROUGH KNOWLEDGE RECIRCULATION

ANVIL connects spatial data production and model learning through **knowledge recirculation** (Figure 1). Shared entities and task dependencies align heterogeneous targets within a scene. These references support cross-form verification and localized repair in ANVIL-GEN, then identify the specialist and response span requiring correction during distillation. The resulting generalist assists subsequent verification, while evidence-backed checks retain acceptance authority and ANVIL-BEN reserves separate targets for evaluation.

3.1 SCENE-ALIGNED SUPERVISION

Let $I \in \mathbb{R}^{H\times W\times 3}$ be an RGB image, with height H and width W, x an instruction, and E its entity inventory. Contract d specifies target space $\mathcal{Y}_d$, required fields, coordinates, and a semantic evaluator. Eleven contracts comprise four Ground forms (box, detection, mask, point), four Relation forms (depth order, metric comparison, direction, relation graph), and three Act forms (affordance, gripper configuration, action plan). A task graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ declares active stages $\mathcal{S} \subseteq \{G, S, A\}$ and the upstream fields consumed along edges $\mathcal{E}$.

Stage k has form d_k and candidate target $\tilde{y}_k \in \mathcal{Y}_{d_k}$, giving $\tilde{r} = (I, x, E, \mathcal{G}, \{d_k, \tilde{y}_k\}_{k\in\mathcal{S}})$. Source annotations and specialist outputs are mapped to this interface using available coordinate and calibration metadata. Alignment preserves entity identities and supported units; it does not turn relative

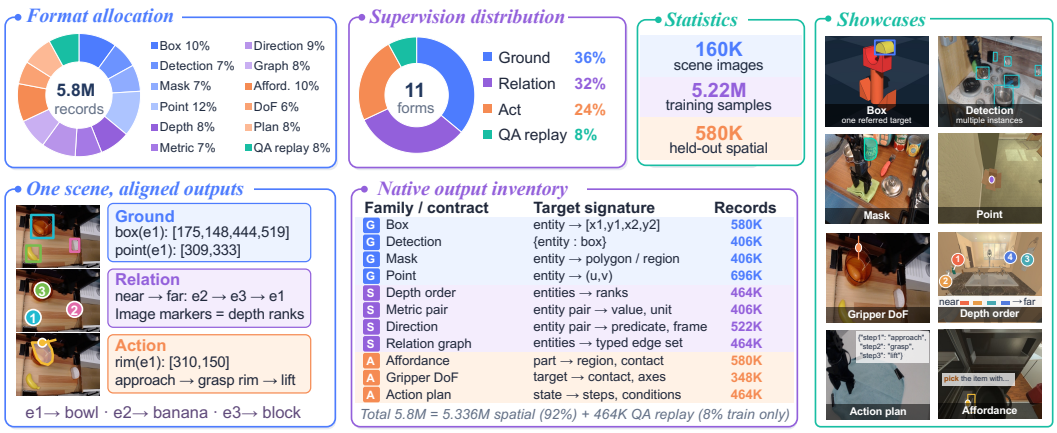


Figure 2: **Scene-aligned supervision and training interfaces.** The distributions summarize the full pool, including training-only QA replay; the inventory lists eleven native spatial formats and their target signatures. Shared entities connect the left example, while the right gallery illustrates eight formats from different scenes. Appendix Table 5 describes source provenance.

estimates into metric evidence. An atomic record contains one stage, whereas linked records expose the fields consumed downstream. Separating native representation from task dependency preserves the distinction between geometry and judgments derived from it, without requiring every scene to support all eleven forms (Appendix A).

3.2 Generation, Verification, and Feedback

Native targets are obtained through source-annotation conversion or schema-constrained specialist generation. Both paths retain scene and entity identities for the shared admission interface. For candidate $\tilde{r}$, binary $C_k(\tilde{r})$ checks native fields and geometry. Let $\mathcal{P}(\tilde{r})$ contain its declared cross-form constraints, including dependency edges and available geometric anchors. Binary $C_c(\tilde{r})$ checks constraint c: entity agreement, point–region containment, or mask–anchor consistency. Let $V(\tilde{r}) \in \{\text{pass}, \text{uncertain}, \text{fail}\}$ be the semantic decision supported by source annotations, calibrated measurements, action rules, or stronger-model adjudication. Admission requires

$$\text{Accept}(\tilde{r}) = \underbrace{\prod_{k \in \mathcal{S}} C_k(\tilde{r})}_{\text{native validity}} \underbrace{\prod_{c \in \mathcal{P}(\tilde{r})} C_c(\tilde{r})}_{\text{cross-form agreement}} \underbrace{\mathbb{1}[V(\tilde{r}) = \text{pass}]}_{\text{semantic support}}. \tag{1}$$

Here $\mathbb{1}$ is the indicator and empty products equal one. Mutually consistent outputs can still be wrong, motivating external evidence and explicit uncertainty rather than acceptance by agreement alone. Accepted targets become references $y_k^\star$ in record r. Appendix A describes the native interfaces and the annotation review procedure. Rejected candidates receive a failure category, not the reference answer. Repair preserves the scene, instruction, and entity IDs; aggregate failures guide source–task sampling. Accepted training records supply supervised targets and distillation references, whereas evaluation targets remain outside this feedback path.

Constructing ANVIL-BEN. Figure 2 summarizes the eleven native spatial formats, their shared scene annotations, and the composition of the full data pool. We stratified the 5.8M generated and verified records by source and native form, assigning scene groups to disjoint partitions: 580K spatial records form ANVIL-EVAL, with sample-based human review, while ANVIL-TRAIN contains the remaining 5.22M records, including all 464K QA replay records.

Records sharing a scene identity remain in one partition; where scene identities are unavailable, grouping uses image identity. The 160K scene images cover the entire pool. Evaluation targets remain outside training and feedback. Native-output tasks assess geometric and relational correctness against reference annotations, while linked tasks vary upstream context with the downstream target held fixed. ANVIL-EVAL assesses both spatial intelligence and the reliability of model outputs as evidence for subsequent reasoning.

3.3 STAGE-ROUTED MULTI-TEACHER DISTILLATION

Student p_θ, with parameters θ, and specialists q_G, q_S, q_A derive from native-supervised QWEN3.5-9B checkpoints. Teachers share the student's vocabulary and remain frozen within each distillation run. Following on-policy distillation (Agarwal et al., 2024; Ma et al., 2026a), they score student-generated prefixes. Rather than routing by prompt domain alone, we use stage references to choose **both teacher and correction span**. For a linked response $\hat{z} \sim p_\theta(\cdot \mid I, x)$, the evaluator assigns $a_k \in \{+1, 0, -1\}$ for pass, uncertain, or fail; unreached stages receive zero. In a declared topological order $\prec$, the first decisive fault

$$\rho(r, \hat{z}) = \min_{\prec} \big\{ k \in \mathcal{S} : a_k = -1,\ a_j = +1\ \forall j \prec k \big\}, \tag{2}$$

with $\rho = \varnothing$ when the set is empty; $j \prec k$ ranges over active stages. The selected teacher q_ρ supervises token positions $\widehat{T}_\rho$ of that stage in $\hat{z}$; uncertainty stops downstream attribution. Atomic records bypass this search and use their declared family. Appendix B defines structural failures and thresholds.

Let ℓ_{ref} be the average target-token negative log-likelihood within each stage, then averaged equally across active stages. Let ℓ_{KD} average the teacher–student generalized Jensen–Shannon divergence over the selected span, with nonnegative token weights summing to one. Stage and span normalization prevent serialization length alone from setting supervision mass. Define $\widehat{T}_\varnothing = \varnothing$ and eligibility $e = \mathbb{1}[|\widehat{T}_\rho| > 0]$. For the accepted native-task training distribution $\mathcal{D}_{\mathrm{train}}$, the objective is

$$\min_\theta\ \mathcal{L}(\theta) = \lambda_{\mathrm{ref}} \mathbb{E}_{r \sim \mathcal{D}_{\mathrm{train}}} [\ell_{\mathrm{ref}}(r)] + \lambda_{\mathrm{KD}} \mathbb{E}_{r, \hat{z}} [\ell_{\mathrm{KD}}(r, \hat{z}) \mid e = 1]. \tag{3}$$

Here $\lambda_{\mathrm{ref}}, \lambda_{\mathrm{KD}} \geq 0$ set the loss scales; the second expectation uses training records and their student-sampled responses. Sampling, teachers, and routing are fixed during each update. With no eligible response, the transfer term is zero, but reference supervision remains active. Appendix B.1 gives the exact losses and shows how the selector induces a normalized teacher–token allocation.

The two objectives serve complementary roles: reference learning maintains supervision across all active stages, whereas routed distillation concentrates specialist guidance on the selected failure. This separation allows correction to be selective without discarding the remaining task supervision. Normalization also keeps a long action serialization from receiving greater loss weight solely because it contains more tokens than a compact spatial answer.

Knowledge recirculation. Verified scene structure guides model correction, while learned competence returns to supervision production:

$$(\mathcal{D}^{(n)}, Q^{(n)}) \xrightarrow{\mathrm{distill}_{\rho, w}} p_{\theta^{(n)}} \xrightarrow{\text{assist verification}} \widetilde{\mathcal{D}}^{(n+1)} \xrightarrow{\text{accept \& merge}} \mathcal{D}^{(n+1)}. \tag{4}$$

Here $\mathcal{D}^{(n)}$ denotes the accepted training set and $Q^{(n)}$ the specialists frozen within round n; ρ selects the correction stage, and w specifies normalized token weights within its response span. The generalist supplies auxiliary signals for candidate verification and repair, producing $\widetilde{\mathcal{D}}^{(n+1)}$. Admission remains governed by Accept rather than student agreement alone; accepted records can support specialist updates before the next round. Shared scene references thereby connect data admission, selective transfer, and knowledge return, while ANVIL-EVAL remains outside training and feedback. Appendix B.3 formalizes this coupling. The following experiments quantify the contributions of verified supervision, selective transfer, and feedback-driven repair within this recirculation framework. At deployment, ANVIL-MOPD performs spatial inference without specialist teachers or reference-aware routing.

4 EXPERIMENTS

We first evaluate ANVIL-MOPD against general-purpose and specialist models on public spatial benchmarks (**RQ1: Performance**). Controlled ablations and retention tests examine how supervision coverage and teacher allocation affect spatial learning and existing capabilities (**RQ2: Learning mechanisms**). We then measure the efficiency of source-based supervision construction and feedback-driven repair (**RQ3: Production efficiency**). Finally, task-structure and input controls test whether predictions depend on relevant scene evidence and task dependencies (**RQ4: Evaluation validity**). Complete comparisons appear in Appendix D.

Table 1: Spatial and embodied-perception evaluation results (%, higher is better).

Model \ Benchmark	CV-Bench (2024)	Emb-Spatial (2024)	Ref-Spatial (2025)	W2P (2025)	BLINK Depth (2024)	ERQA (2025)	VABench Point (2026)	**Avg.**	Anvil -Eval
Spatial and embodied models									
Cosmos-Reason1-7B (2025)	77.10	68.87	0.54	3.21	69.35	40.07	0.00	37.02	2.80
Magma-8B (2025a)	60.98†	64.59†	4.50†	10.89†	65.32†	25.73†	8.20	34.32	15.18
RoboPoint-13B (2025)	79.68	49.31†	22.87$^{\dagger\mathrm{L}}$	46.80†	61.29†	29.30	19.09	44.05	18.80
RoboBrain2.0-7B (2025)	85.75†	76.32†	32.50†	63.59†	84.68	30.31†	26.67†	57.12	45.94
MiMo-Embodied-7B (2025)	88.82†	76.24†	48.00†	63.60†	62.10	46.75†	46.93†	61.78	4.37
RynnBrain-8B (2026)	87.87	79.78	59.20†	67.00	90.32	46.80†	11.67	63.23	55.98
RoboBrain2.5-8B-NV (2026)	94.58†	75.58†	60.50†	68.00	84.68	45.99	42.00	67.33	61.17
RoboRefer-2B-RFT (2025)	92.20	75.34	52.00$^{\dagger\mathrm{L}}$	71.00†	84.32	46.52	33.34	64.96	39.25
RoboRefer-8B-SFT^{D} (2025)	96.24$^{\dagger\mathrm{S}}$	72.53†	52.00$^{\dagger\mathrm{L}}$	70.00†	92.74†	52.33	27.12	66.14	52.21
General-purpose models									
GPT-4o (2025)	78.63†	71.92†	8.78†	20.41†	78.23†	32.48†	13.67†	43.45	29.67
Gemini-2.5-Pro (2025)	84.59†	78.74†	38.16†	42.38†	87.90†	51.02†	27.92†	58.67	55.99
Gemini-3-Pro-Preview (2026)	92.00†	76.62†	65.50†	61.52	90.30	66.00	32.34	69.18	57.94
Qwen2.5-VL-7B (2025b)	75.40†	70.25†	38.00†	42.00†	79.00	38.80†	24.50†	52.56	3.21
InternVL3.5-8B (2025)	81.46†	70.26†	16.80†	16.49†	79.84	41.00†	11.93†	45.40	48.46
Qwen3-VL-8B-Instruct (2025a)	86.69	78.87	53.50	34.32	87.90	44.25	46.33	61.69	55.16
QWEN3.5-9B (2026)	87.15	75.58	61.67	47.97	90.32	49.83	42.33	64.98	56.87
ANVIL-MOPD (ours)	**88.21**	**82.34**	**59.00**	**68.00**	**91.13**	**50.52**	**80.67**	**74.27**	**65.35**

Notes. Avg. is the mean of the seven public benchmarks, excluding our evaluation benchmark. †: published source-protocol results; ^{D}: supports RGB-D input; ^{L}: RefSpatial Location only; ^{S}: CV-Bench subset aggregate excluding counting.

4.1 EXPERIMENTAL SETUP

Eight public benchmarks cover spatial understanding (CV-Bench (Tong et al., 2024), EmbSpatial (Du et al., 2024), BLINK (Fu et al., 2024)), referring and placement (RefSpatial (Zhou et al., 2025), Where2Place (Yuan et al., 2025), RoboSpatial (Song et al., 2024)), and embodied reasoning and pointing (ERQA (Google DeepMind, 2025), VABench (Yuan et al., 2026)). The baselines span general-purpose VLMs, spatial specialists, and embodied models. We train ANVIL-MOPD from QWEN3.5-9B using native spatial supervision and multi-teacher distillation, and report the final checkpoint in the main table. Local comparisons share task-specific prompts, decoding settings, and scorers; published scores retain their source protocols. Ablations control training updates and record exposures for coverage, initialization, teachers, and maximum rollout budgets for routing, and candidate pools and per-candidate call counts for repair. Coverage and routing comparisons each use two seeds; full settings appear in Appendix C. The main-table mean covers seven public tasks; RoboSpatial supplies an additional coverage control. Earlier source-filtered checkpoints support the intervention analyses, while linked-source models support the routing study. These studies test individual mechanisms rather than provide additional scores for the final model. Published cross-protocol averages are descriptive; within-protocol comparisons provide the controlled evidence.

4.2 SPATIAL INTELLIGENCE ASSESSMENT AND COVERAGE

Observation 1: heterogeneous supervision transfers to spatial reasoning and pointing. We first ask whether ANVIL-MOPD improves both spatial understanding and interaction-relevant prediction (**RQ1**). Across the seven public benchmarks in Table 1, the final checkpoint achieves a 74.27% mean, exceeding QWEN3.5-9B by 9.29 points. On EmbSpatial, it outperforms Qwen3-VL-8B and RynnBrain-8B by 3.47 and 2.56 points, respectively, showing gains over both general-purpose and embodied baselines. These comparisons show that the gains extend from spatial judgments to explicit location prediction, rather than remaining confined to a single task or output format.

The aggregate gain is not confined to pointing. Even excluding VABench-Point, the mean improvement over the base remains 4.45 points. The overall gain therefore does not depend entirely on the

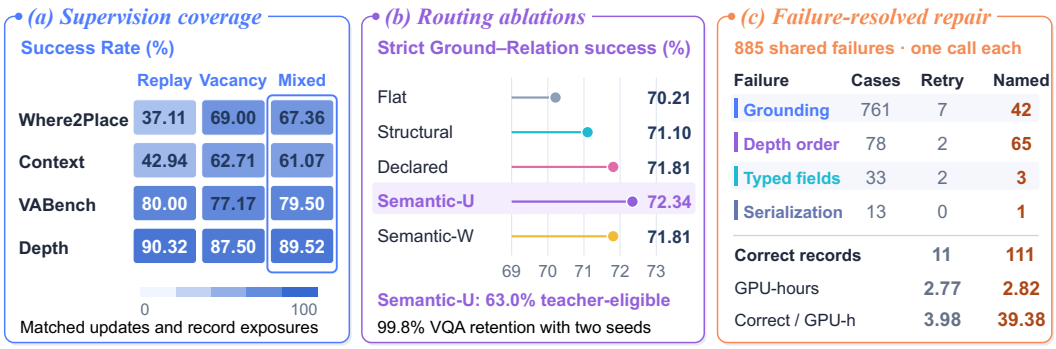


Figure 3: **Controlled tests of coverage, routing, and repair.** (a) Coverage matrix at matched updates and record exposures (Context: RoboSpatial; Depth: BLINK). (b) Routing controls at matched maximum rollout budgets (U: uniform; W: weighted). (c) Correct repairs by failure type, using the same 885 candidates and one call each.

largest task-specific increase or its any-point scoring rule. Transfer remains uneven, with smaller regressions on RefSpatial. In a preliminary control study with a source-filtered model, we test whether pointing gains rely on multiple predictions. Scoring only the first point on the common 286-item VABench subset retains a 34.62-point advantage over the base (Table 7), supporting a localization benefit beyond additional guesses.

4.3 Mechanism Ablations

Observation 2: targeted coverage improves adaptation, while replay mitigates forgetting. We use preliminary controlled studies to isolate three decisions: which supervision to add, where to apply teacher correction, and how to repair rejected outputs (**RQ2**). Mixed training improves Where2Place over replay by 30.25 points across two seeds and also benefits RoboSpatial context (Figure 3a). Fixed updates and exposures separate these gains from simply training longer; increasing the decoding budget alone does not recover placement performance. Vacancy-only training favors placement, while replay better preserves pointing and depth. This trade-off motivates targeting ANVIL-GEN at missing capabilities while retaining replay. Mixed training does not improve every task, and unique images and realized tokens remain unmatched (Appendix C).

Observation 3: selective teacher correction improves transfer with limited retention loss. We test where teachers should intervene. With common initialization and a matched maximum rollout budget, semantic-uniform routing improves strict Ground–Relation success over flat distillation by 2.13 points (Figure 3b). Semantic references expose incorrect but well-formed geometry that structural checks miss. Teacher supervision reaches 63.00% of sampled responses, while reference learning covers all active stages, with 99.83% general-VQA retention. The contrast supports selective correction over flat distillation.

Observation 4: failure-specific feedback improves repair yield at comparable cost. We test whether failure-specific information improves repair. Using Qwen3.6-27B as the repair model, both arms receive the same 885 rejected candidates and one call per candidate. Named feedback recovers 10.09× as many source-correct records as generic retry at comparable GPU cost (Figure 3c). It identifies the failed category without reference coordinates or ranks, preserving the task while localizing correction. Recovery concentrates on depth ordering; grounding remains harder to repair.

Together, these controls support organizing supervision around capability gaps, localizing teacher correction, and using explicit failure information to recover usable training targets.

4.4 Qualitative Analysis and Production Efficiency

Figure 4 shows how the preceding mechanisms change individual outputs. In panel (a), adaptation moves placement predictions into the reference region; for the chair example, all three original predictions miss that region. In panel (b), feedback corrects the depth order of the coffee maker

and sink while preserving their boxes. The latter case illustrates how an error can be localized to Relation while retaining the Ground output. These selected cases explain the form of the correction; the ablations measure its aggregate benefit, and Appendix E includes a placement regression.

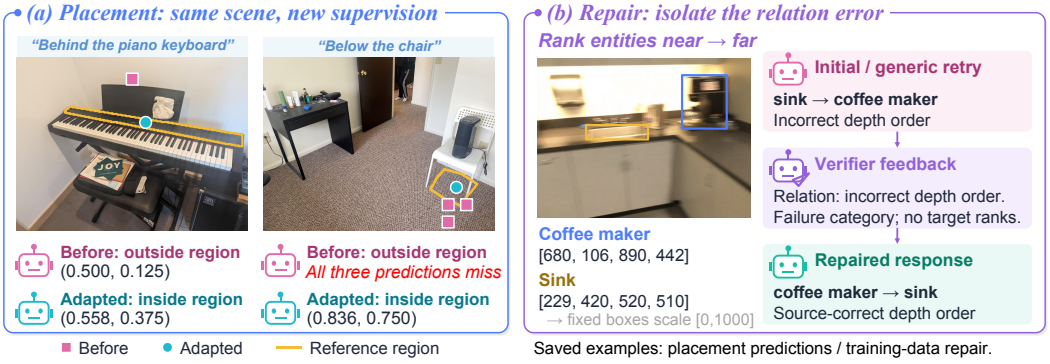


Figure 4: **Qualitative examples of coverage adaptation and targeted repair.** (a) Saved placement predictions before and after adaptation (square: before; circle: adapted; yellow: reference region). (b) Verifier feedback guides a depth-order correction while object boxes remain unchanged. Selected successes illustrate the mechanisms, not their frequency.

Observation 5: source evidence amortizes supervision construction. We next examine where supervision production incurs computational cost (**RQ3**). From 9,608 prepared RGB-D frames, eight CPU workers construct 2,463 linked records in 325.25 seconds without generative annotation. Registered depth and instance geometry provide the shared references needed to link Ground and Relation targets. A separate pass rechecks 598,891 stored records in 13.26 seconds, covering reading, parsing, hashing, and rule evaluation across four shards (Appendix E). These measurements show that available source evidence supports both target construction and large-scale structural checking with CPU computation. They are component timings: source preparation, image loading during rechecking, and semantic review remain outside the measured scope.

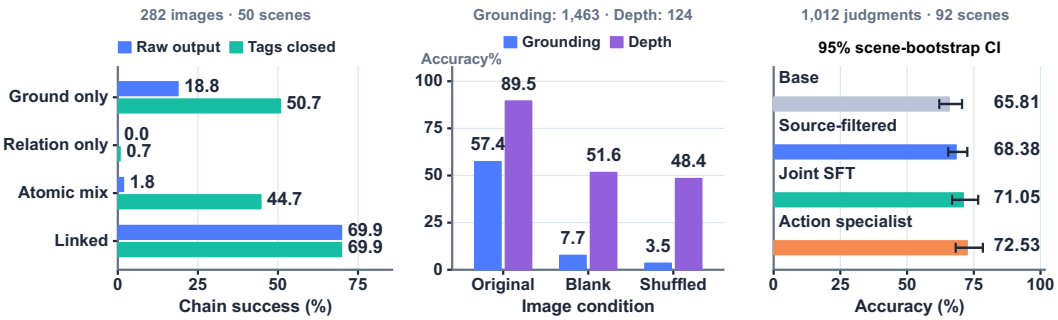
For candidates requiring model-based repair, the relevant efficiency measure is source-correct output per unit of computation. Normalizing the repair comparison by allocated GPU time gives 39.38 correct records per GPU-hour for named feedback and 3.98 for generic retry, a 9.90× advantage. The images, questions, and candidate pool are fixed, so this gain reflects the value of localized error information under the same repair workload. Thus, we reuse source geometry, apply inexpensive structural checks, and direct model calls to unresolved semantic errors.

The practical implication is to allocate computation according to the evidence available for each record. Source-supported targets can be constructed and checked directly; unresolved candidates receive model-based adjudication or repair. This division makes heterogeneous sources useful for different parts of the production process, rather than requiring every annotation to pass through the same generative workflow.

4.5 Task Structure and Evidence Dependence

We ask whether linked-task gains survive format correction, whether predictions depend on visual and upstream evidence, and whether action judgments distinguish plans containing the same actions (**RQ4**). Figure 5 addresses the question through composition, visual-input, and action-order controls.

Observation 6: scene-linked supervision substantially improves spatial composition. We first test whether independently learned Ground and Relation outputs compose under a shared interface. On 282 images from 50 held-out physical scenes, linked supervision reaches 69.86% strict Ground–Relation success, substantially exceeding the atomic mixture at matched updates (Figure 5a). To separate missing delimiters from errors in the predicted content, we close stage tags in the saved responses without changing their JSON outputs. This raises the atomic mixture to 44.68% while leaving linked performance unchanged. The remaining gap indicates that output syntax is only part



(a) Composing spatial outputs *(b) Using visual evidence* *(c) Verifying action plans*

Figure 5: **Task structure and evidence dependence.** (a) Linked-task success before and after closing missing stage tags, with JSON outputs unchanged. (b) Grounding and depth under original, blank, and shuffled images. (c) Plan-validity accuracy on 1,012 constructed judgments with 95% scene-bootstrap intervals. Panels use the study-specific checkpoints described in Appendix E.

of the composition problem. A downstream stage must also receive the correct entities and geometry from its predecessor. Scene-linked supervision trains this dependency explicitly, making intermediate outputs useful inputs to subsequent reasoning rather than merely separately correct answers.

Observation 7: spatial predictions depend on images and upstream geometry. We next examine dependence on the evidence supplied to these stages. With questions and targets fixed, blank images reduce the source-filtered model's accuracy on the 1,463-example grounding panel from 57.35% to 7.66%. Cross-scene shuffling also degrades grounding, while BLINK-depth approaches chance (Figure 5b). Keeping images and depth targets fixed but shifting upstream boxes reduces the Relation specialist's accuracy by 10.64 points (Table 18). In particular, errors in Ground can alter the context available to Relation even when the scene image remains fixed. These interventions establish sensitivity to both visual content and consumed geometry, motivating joint evaluation of stage accuracy and linked success.

Observation 8: action rules support state-conditioned plan verification. We extend the evaluation to action order under state constraints. From 506 ALFRED trajectories (Shridhar et al., 2020) across 92 scenes, we construct 1,012 balanced judgments, pairing each valid plan with an invalid permutation of the same actions. This fixes vocabulary and length, so action occurrence alone cannot distinguish the labels. Given the initial image, complete Boolean state, goal, and action rules, the Action specialist improves over the base by 6.72 points (Figure 5c). Class-wise results remain uneven: accuracy is 89.53% on invalid plans but only 55.53% on valid ones (Table 19). The improvement mainly comes from recognizing more valid plans, although many feasible sequences are still rejected.

Summary and discussion. The experiments suggest that spatial learning depends on how supervision is organized, not only how much is available. Coverage controls favor task-relevant additions with replay; routing controls favor localized specialist correction; and repair controls show the value of actionable failure feedback. Source-based construction complements these mechanisms by reusing available geometric evidence. Beyond aggregate benchmark gains, the composition, input, and action-order controls test whether predictions remain useful when later decisions depend on their content. These findings motivate scaling spatial supervision through capability coverage, selective correction, and explicit task dependencies.

5 Conclusion

We presented ANVIL, which connects spatial data generation, evaluation, and multi-teacher learning through knowledge recirculation. Shared scene references support cross-form verification, guide selective specialist transfer, and govern the acceptance of returned supervision. The final model improves the seven-benchmark mean by 9.29 points over its base; controlled studies support targeted coverage, selective transfer, and feedback-driven repair. Together, these findings support jointly designing reliable supervision and knowledge consolidation for embodied spatial intelligence.

AI USE STATEMENT

Generative AI tools were used for literature retrieval and discovery and for improving the clarity and presentation of author-written text. Generative models were also used, as part of the research method, to produce candidate synthetic spatial-supervision records. They were subjected to the verification, filtering, and auditing procedures described in the paper. The authors verified the cited sources, experiments, reported results, and all AI-assisted content, and remain fully responsible for the manuscript's claims and compliance with applicable data licenses and usage terms.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019. doi: 10.1609/aaai.v33i01.33018076.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Rim Assouel, Amir Bar, Michal Drozdzal, and Adriana Romero-Soriano. PGT: Procedurally generated tasks for improving visual grounding in MLLMs. *arXiv preprint arXiv:2605.23883*, 2026.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoxu Lyu, Yijie Xu, Jiayu Shi, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Shanyu Rong, Zhengliang Cai, Bolun Zhang, Shuyi Zhang, Huaihai Lyu, Mengfei Du, Lingfeng Zhang, Xi Feng, Xiaodan Liu, Yance Jiao, Chenrui He, Mengsi Lyu, Zhuo Chen, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. Spatial result locators use version 1, Tables 2 and 3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Nils Blank, Paul Mattes, Maximilian Xiling Li, Jakub Suliga, Thomas Roth, Moritz Reuss, Pankhuri Vanjani, and Rudolf Lioutikov. SPARC: Reliable spatial annotations from robot demonstrations at scale. *arXiv preprint arXiv:2606.13497*, 2026.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7879–7890, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024.

Kaiyuan Eric Chen, Shuangyu Xie, Zehan Ma, Pannag Sanketi, and Ken Goldberg. Robo2VLM: Improving visual question answering using large-scale robot manipulation data. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-0733.

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*, volume 37, 2024. doi: 10.52202/079017-4293.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. RynnBrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026. Result locators use version 1, Table 3.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, June 2025.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 2024.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-MOPD: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation. *arXiv preprint arXiv:2608.19098*, 2026.

Google DeepMind. Embodied reasoning question answering benchmark. Official ERQA dataset and evaluation repository, 2025. URL `https://github.com/embodiedreasoning/ERQA`.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, Yuchen Zhang, Jing Wu, Jinghui Lu, Chenxu Dang, Jiayi Guan, Jianhua Wu, Zhiyi Hou, Hanbing Li, Shumeng Xia, Mingliang Zhou, Yinan Zheng, Zihao Yue, Shuhao Gu, Hao Tian, Yuannan Shen, Jianwei Cui, Wen Zhang, Shaoqing Xu, Bing Wang, Haiyang Sun, Zeyu Zhu, Yuncheng Jiang, Zibin Guo, Chuhong Gong, Chaofan Zhang, Wenbo Ding, Kun Ma, Guang Chen, Rui Cai, Diyun Xiang, Heng Qu, Fuli Luo, Hangjun Ye, and Long Chen. MiMo-Embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025. Result locators use version 1, Tables 2 and 3.

Xixiang He, Xingming Li, Baiqi Wu, Qiyao Sun, Xuanyu Ji, Ao Cheng, and Qingyong Hu. Learn from whoever is right: Answer-verified multi-teacher distillation for multi-domain LLMs. *arXiv preprint arXiv:2609.02548*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ruihan Jin, Pengpeng Shao, Zhengqi Wen, Jinyang Wu, Mingkuan Feng, Shuo Yang, Chu Yuan Zhang, and Jianhua Tao. Exploring knowledge purification in multi-teacher knowledge distillation for LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yuanchen Ju, Yongyuan Liang, Yen-Jen Wang, Gireesh Nandiraju, Yuanliang Ju, Seungjae Lee, Qiao Gu, Elvis Hsieh, Furong Huang, and Koushil Sreenath. MomaGraph: State-aware unified scene graphs with vision-language models for embodied task planning. In *International Conference on Learning Representations (ICLR)*, 2026. Oral.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.568.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139.

Jahyun Koo, Yerin Hwang, Yongil Kim, Taegwan Kang, Hyunkyung Bae, and Kyomin Jung. SWITCH: Studying with teacher for knowledge distillation of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3733–3746, 2025. doi: 10.18653/v1/2025.findings-naacl.206.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: Multi-Modal In-Context Instruction Tuning. *arXiv preprint arXiv:2306.05425*, 2023.

Enhan Li, Junhao He, and Hongyang Du. CROP: Task relevance via counterfactuals for selective on-policy distillation. *arXiv preprint arXiv:2608.13387*, 2026a.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026b.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, Minghua Zeng, Jiangpin Liu, Ronghao Dang, Jiayan Guo, Siteng Huang, Haoyu Zhao, Heng Ping, Yaxi Zhao, Tong Zhao, Kexiang Wang, Tong Lu, Shengke Xue, Jiahao Tang, Yulei Wang, Zejing Wang, Jianwei Gao, Shijian Lu, Chengju Liu, Jianfei Yang, Mingxiu Chen, and Deli Zhao. RynnBrain 1.1: Towards more capable and generalizable embodied foundation model. *arXiv preprint arXiv:2607.17977*, 2026c.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

Junze Liu, Kun Qian, Florian Dubost, Kai Zhong, Arvind Srinivasan, Nan Chen, Anping Wang, Sam Zhang, Alejandro Mottini, Qingjun Cui, and Tian Wang. 3D primitives are a spatial language for VLMs. *arXiv preprint arXiv:2605.12586*, 2026a.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026b.

Yi Liu, Sukai Wang, Dafeng Wei, Xiaowei Cai, Linqing Zhong, Jiange Yang, Guanghui Ren, Jinyu Zhang, Maoqing Yao, Chuankang Li, Xindong He, Liliang Chen, and Jianlan Luo. Unified embodied VLM reasoning with robotic action via autoregressive discretized pre-training. *arXiv preprint arXiv:2512.24125*, 2025.

Yicheng Liu, Zibin Dong, Baijun Ye, Tianyuan Yuan, Tao Jiang, Anqi Yang, Shicheng Cao, Haonan Liu, Yue Sun, Zihan Guo, Xiao Liu, Dong Ke, Changxun Pan, Chenru Wu, Tailai Cheng, Xiaoshu Ren, Xinlei Zhang, Jianning Cui, Zijie Zhao, Haoyu Zhang, Kaiming Xu, Haodong Yang, Bowen Zhang, Jiahui Niu, Shaoting Zhu, Shiduo Zhang, and Hang Zhao. G0.5: One autoregressive stream for robot reasoning and action. *arXiv preprint arXiv:2608.11739*, 2026c.

Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, Mingfei Han, Meng Cao, Bokui Chen, Ivan Laptev, and Xiaodan Liang. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly, 2025.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026a.

Yingzi Ma, Zichen Zhu, Ming Jiang, and Chaowei Xiao. One student, many teachers: Multi-task on-policy distillation via soft-prompt privileged context. *arXiv preprint arXiv:2607.18293*, 2026b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Kiet T. Nguyen, Hanbo Shim, Jinwoo Kim, and Seunghoon Hong. Multi-view relational distillation for spatial reasoning with vision-language models. *arXiv preprint arXiv:2608.10864*, 2026.

NVIDIA. Cosmos-Reason1-7B: Physical AI common sense and embodied reasoning models. Official model card; August 2025 7B release, 2025. URL `https://huggingface.co/nvidia/Cosmos-Reason1-7B`.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting Vision and Language with Localized Narratives. In *ECCV*, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official release and Qwen3.5-9B model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-9B`.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling*, 2025.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems*, 2015.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *arXiv preprint arXiv:2411.16537*, 2024.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026.

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 (NeurIPS)*, volume 37, pp. 87310–87356, 2024.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Ziming Wei, Bingqian Lin, Zijian Jiao, Yunshuang Nie, Liang Ma, Yuecheng Liu, Yuzheng Zhuang, and Xiaodan Liang. MineAnyBuild: Benchmarking Spatial Planning for Open-world AI Agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. VSP: Diagnosing the dual challenges of perception and reasoning in spatial planning tasks for MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2270–2280, 2025.

Haolei Xu, Xiaowen Xu, Haiwen Hong, Zixuan Ni, Hongxing Li, Yiwen Qiu, Weiming Lu, and Yongliang Shen. Pass the baton: Trajectory-relayed on-policy distillation. *arXiv preprint arXiv:2607.26057*, 2026.

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*, pp. 14203–14214, 2025a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025b.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video. *arXiv preprint arXiv:2511.04670*, 2025c.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction in robotics. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4005–4020. PMLR, 2025.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *International Conference on Learning Representations*, 2026.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

He Zhang, Lingzhu Xiang, Haitao Lin, Zeyu Huang, Minghui Wang, Dingyan Zhong, Yubo Dong, Yihao Wu, Yongming Rao, Dongsheng Zhang, Wanjia He, Ling Chen, Kai Huang, Jiahao Chen, Sichang Su, Xumin Yu, Ziyi Wang, Chengwei Zhu, Xiao Teng, Yuchun Guo, Yufeng Zhang, Yuandong Liu, Rui Wang, Zisheng Lu, Han Hu, and Zhengyou Zhang. Hy-Embodied-0.5-VLA: From vision-language-action models to a real-world robot learning stack. *arXiv preprint arXiv:2606.14409*, 2026a.

Jiyao Zhang, Mingxu Zhang, Yitong Peng, Haoxuan Liu, Chenshuo Wang, Yuxing Long, Haoyang Huang, Dongjiang Li, Nan Duan, Hui Shen, and Hao Dong. Embodied3DBench: Benchmarking low-level embodied spatial intelligence of vision language models. *arXiv preprint arXiv:2605.29074*, 2026b.

Le Zhang, Jihan Yang, Soundarya Krishnan, Jimit Majmudar, Xiou Ge, Prasoon Puri, Prathamesh Nandkishor Saraf, Shruti Bhargava, Dhivya Piraviperumal, Yinan Ling, Cindy Pan, Hong Yu, Aishwarya Agrawal, and Bo-Hsiang Tseng. From where things are to what they are for: Benchmarking spatial-functional intelligence in multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026c.

Wenqi Zhang, Mengna Wang, Gangao Liu, Huixin Xu, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, Jiajun Liu, Weiming Lu, Peng Li, and Yueting Zhuang. Embodied-Reasoner: Synergizing Visual Search, Reasoning, and Action for Embodied Interactive Tasks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41178–41207. Association for Computational Linguistics, 2026d. doi: 10.18653/v1/2026.acl-long.1910.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs. In *International Conference on Learning Representations (ICLR)*, 2026e.

Helu Zhi, Jingjing Huang, Wang Xu, Yangbin Xu, Yibin Huang, Wanyue Zhang, Baoyang Jiang, Shirui Deng, Liang Zhu, Fangfang Li, Tiejun Zhao, Yankai Lin, and Yuan Yao. Scaling spatial reasoning in MLLMs through programmatic data synthesis. *arXiv preprint arXiv:2512.16237*, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023.

A Spatial Data Flywheel and Native Supervision

This section details the supervision interfaces introduced in Section 3.1, including native representations, scene alignment, verification, and source provenance. Routing, training, and evaluation protocols are provided in Appendices B–E.

A.1 Native interfaces and scene alignment

Scene alignment. Each record retains the image, instruction, entity inventory, dependencies, and native references from Section 3.1. Conversion $\mathcal{A}_d(\xi; E, \kappa) \in \mathcal{Y}_d$ maps annotation ξ to form d using entity identities E and calibration metadata κ. Image coordinates (u, v) become $(1000u/W, 1000v/H)$.

Metric targets require supported units and calibration; relative estimates remain relative. Shared identities connect available annotations without requiring every scene to support all eleven forms or a universal sequence. The reported $G \to S$ score measures composition between the grounding and spatial relationship.

Table 2: **Eleven native spatial interfaces.** The symbols summarize output structure; colors and family letters identify Ground (G), Relation (S), and Act (A). Each form retains its own validity conditions, while shared entity references connect available forms.

Family	Form / structure	Native target and consistency requirement
G Ground	$\square$ Box	Referred entity and bounded, ordered corners.
	$\{\square_i\}$ Detection	Entity-indexed boxes with non-duplicate instances.
	$\{(x_i, y_i)\}$ Mask	Instance or part region consistent with its anchor.
	• Point	Object, contact, or free-space location in a declared frame.
S Relation	$i \prec j$ Depth order	Shared entity set and consistent near–far ordering.
	$d(i, j)$ Metric comparison	Distances or comparisons with supported units.
	$i \to j$ Direction	Relative direction with an explicit reference frame.
	$(E, \mathcal{R})$ Relation graph	Typed edges over the shared entity inventory.
A Act	(a, p) Affordance	Action, functional part, and compatible contact region.
	$(p, \mathbf{v})$ Gripper	Contact, approach, and jaw geometry.
	$a_1 \to a_2 \to \cdots$ Plan	Goal, ordered steps, entity references, and state constraints.

Native supervision examples Figure 6 expands the compact examples in Figure 2. The eight stored forms contain object boxes, multiple detections, instance polygons, surface points, depth ordering, contact regions, gripper configurations, and action plans. These examples come from an earlier reviewed pilot and illustrate the supervision interfaces; they are not predictions of ANVIL-MOPD. Three additional panels illustrate relations derived from stored image-plane geometry. Their distances use normalized image coordinates, whereas physical distances require the calibration described in Appendix A.1.

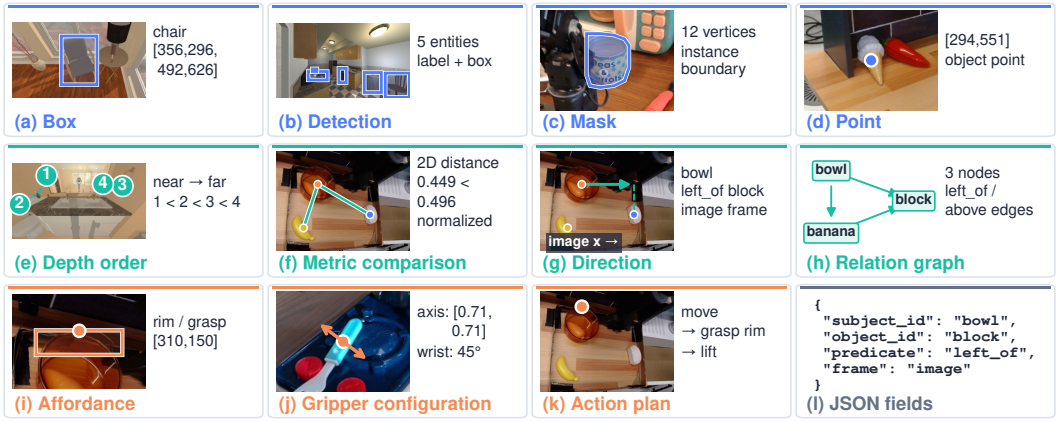


Figure 6: **Examples of native spatial supervision.** Panels (a–e) and (i–k) show stored annotations from an earlier eight-form pilot. Panels (f–h) illustrate image-plane distance, direction, and graph relations derived from saved entity boxes; panel (l) shows fields rather than an additional task form. Coordinates refer to the original full image on a 0–1000 scale, including when a panel displays a crop. The examples illustrate recorded supervision and geometry-derived interfaces, rather than model predictions or executed robot trajectories.

A shared contact across native forms. The orange-bowl example links a functional region to a plan through the same contact point. The affordance record identifies the rim, while the plan names

that part in its grasp step and retains the contact coordinates as its target. The following excerpts preserve the stored field values; plan preconditions and failure modes are omitted for brevity.

```
Affordance:
{"label": "rim", "affordance": "grasp",
 "bbox_2d": [175,148,444,250],
 "contact_point_2d": [310,150]}

Action plan:
{"goal": "pick up the orange bowl",
 "steps": ["move to the orange bowl",
           "grasp the rim of the bowl",
           "lift the bowl upwards"],
 "target_point_2d": [310,150]}
```

Matching coordinates make the dependency inspectable. Semantic review must additionally establish that the point lies on a visible contact surface and that the stated action is supported by the available evidence.

Generation constraints. The archived generation templates make geometric requirements explicit before requesting a native output. For example, the point template asks for visible object material and warns that a box center may lie in a hole or between chair legs. The mask template requires a silhouette of the same grounded entity and permits an empty output when a single polygon would include substantial background. The depth template requires visible ordering evidence and permits abstention for ambiguous pairs. These constraints provide concrete examples of how the interfaces in Table 2 enter generation.

Table 3: **Representative generation constraints.** Concise summaries of the archived templates; the requirements below are not additional experimental measurements.

Form	Constraint supplied to the generator
Visible point	Preserve the grounded entity label; select visible material rather than a box center by default; return an empty array if no unambiguous point exists.
Mask	Segment the same grounded entity with 8–16 nested coordinate pairs; follow its silhouette rather than returning box corners; abstain when one polygon cannot represent it adequately.
Depth order	Preserve the entity set and labels; rank near to far using visible evidence; image height or apparent size alone is insufficient.
Contact region	Specify a visible functional sub-part and compatible action; place the contact on material rather than an empty container interior.
Gripper	Use the contact proposal, an image-plane jaw axis, an approach class, and a consistent wrist angle; abstain if secure contact is unsupported.

Review interface and decision rules. The annotation tools pair the source image and spatial overlays with the candidate structured output. The review view exposes the relevant entity context and records a support decision, an error category, and an optional explanation. Unsupported decisions require an error category. Stable sample identifiers and saved label snapshots bind each decision to the image and candidate that were reviewed.

The archived review protocol specifies separate initial judgments followed by adjudication of disagreements. It also distinguishes operational inspection, which may display automatic verification status, from the initial review view, which omits model confidence and another reviewer's decision. Updating an image, target, or task definition requires reviewing the affected candidate again rather than transferring its previous decision.

Review decisions and automatic admission results have different denominators. Stored pilot judgments describe the candidates in that pilot; a held-out audit must use the sample identities and protocol of that held-out partition.

Table 4: **Decision rules used to inspect spatial candidates.** The checks concern the visible evidence and the declared task. A valid serialization alone does not establish semantic support.

Target	Evidence to inspect	Reason to withhold or reject
Box / detection	Identity and visible extent of the requested entity.	Wrong instance, omitted material parts, or duplicated detections.
Mask / point	Visible object material and its boundary.	Background inside a polygon; a point in a hole or behind an occluder.
Depth order	Occlusion, contact layering, or a supported perspective cue.	A reversed pair or insufficient evidence to order it.
Affordance / gripper	Named contact part, contact point, and jaw geometry.	A grasp in empty container space or on an unsupported contact surface.
Action plan	Referenced entities, stated preconditions, and ordered steps.	Inapplicable actions or a target inconsistent with the named contact.

Concrete rejection cases. The stored judgments include cases in which individually well-formed outputs fail a physical or semantic check. In one pot scene, an affordance names the rim while its contact point lies in the dark opening. The gripper record and action plan reuse that location. Agreement across the three outputs therefore preserves the same unsupported contact. In another scene, a proposed grasp targets the center of a sticker adhered to a board, without a pinchable edge. A grounding example identifies a lamp but omits its visible base. These cases motivate inspecting the referenced entity and contact surface in addition to checking the output fields.

Source provenance and benchmark scope. ANVIL-GEN generates and verifies 5.8M records over 160K images: 5.336M spatial records and 464K QA replay records. We reserve 580K spatial records for held-out evaluation and use 5.22M records for training, including all QA replay. Evaluation targets remain outside training and feedback. Table 5 summarizes the source families. The controlled studies use task-specific evaluation panels: 1,463 box/point examples, 1,024 masks, 282 linked images from 50 scenes, and 1,012 plan judgments from 92 scenes. Each reported score uses its panel's denominator. We manually sampled and audited 1,463 evaluation records from the held-out 580K evaluation partition. The audit reported 88.79% annotation correctness.

Table 5: **Source families and supervision roles.** MIMIC-CGD denotes the CGD subset of MIMIC-IT; VSI-590K is the instruction-tuning corpus released with Cambrian-S.

Source family	Supervision role	Illustrative provenance
Robot video	Grounding, interaction	BridgeData V2 (Walke et al., 2023)
Embodied scenes	Spatial targets	RefSpatial (Zhou et al., 2025)
Simulation	Objects, actions, physical relations	Embodied-Reasoner (Zhang et al., 2026d); PhyBlock (Ma et al., 2025); MineAnyBuild (Wei et al., 2025)
Spatial imagery	Placement, spatial QA, relations	FSD (Yuan et al., 2026); SAT (Ray et al., 2025); VSI-590K (Yang et al., 2025c); SpatialSense (Yang et al., 2019)
Object states	State-conditioned grounding and actions	ChangeIt frames (Souček et al., 2022)
Pointing	Points and caption QA	PixMo (Deitke et al., 2025)
QA replay	Paired-image and compositional QA	MIMIC-CGD (Li et al., 2023); NLVR2 (Suhr et al., 2019); CLEVR (Johnson et al., 2017); TallyQA (Acharya et al., 2019); VSR (Liu et al., 2023)
	General visual QA	Localized Narratives (Pont-Tuset et al., 2020); VQAv2 (Goyal et al., 2017); COCO-QA (Ren et al., 2015); A-OKVQA (Schwenk et al., 2022); OK-VQA (Marino et al., 2019); Visual7W (Zhu et al., 2016)

Provenance, grouping, and record accounting Source conversion, model-generated candidates, and review decisions carry distinct provenance. A converted reference is supported by its source annotation or measurement; a generated candidate is associated with the generator and template version; a review decision is tied to a specific saved candidate. Geometry-derived examples additionally

declare their coordinate frame and units. This distinction prevents an image-plane relation from being interpreted as a calibrated physical measurement.

Image identity, scene grouping, and target identity address different forms of duplication. Multiple native targets from the same scene remain associated with their shared group when a scene identity is available; otherwise, the grouping unit is the image, as specified in Section 3.2. A count of generated records can therefore exceed the number of images and does not count independent scenes. The figures illustrating formats are selected examples rather than a representative sample from which corpus proportions are estimated.

B Routing, Distillation, and Knowledge Return

Stage decisions. A reached stage with malformed output or a violated native constraint has $a_k = -1$. Otherwise, its native score s_k fails below $\tau_k - \delta_k$, passes at or above $\tau_k + \delta_k$, and is uncertain between these bounds. The source-linked study uses $(\tau_G, \delta_G) = (0.75, 0.125)$ and $(\tau_S, \delta_S) = (0.5, 0.25)$, giving boundaries $0.625/0.875$ and $0.25/0.75$. Equality at the lower boundary is uncertain; equality at the upper boundary passes. Unreached stages are uncertain. The scan advances through passed stages. A definite failure selects that stage; uncertainty stops attribution without selecting a later stage. Malformed outputs use the nonempty suffix after the last valid boundary as the correction span. Reference learning remains active for every record. For linked records, distillation is omitted when the scan selects no stage or the selected correction span is empty. Atomic records use their declared family and sampled response span.

For linked tasks, a downstream stage consumes upstream fields, so an upstream error can invalidate its input context. First-fault routing therefore selects a downstream specialist only after preceding stages pass their checks. Once an upstream fault is selected, changing only later predictions cannot redirect correction.

B.1 Normalized teacher–token allocation

Let $\widehat{T}_k$ be the sampled positions of stage k, and let $w_{k,t}$ be nonnegative weights summing to one on a nonempty span and zero outside it. The selector jointly allocates a specialist and correction tokens:

$$\Omega_{k,t} = \mathbb{1}[\rho(r, \hat{z}) = k] w_{k,t}, \qquad \ell_{\mathrm{KD}} = \sum_{k \in \mathcal{S}} \sum_{t \in \widehat{T}_k} \Omega_{k,t} D^{(k)}_{\beta,t}. \tag{5}$$

Here $D^{(k)}_{\beta,t}$ compares student and teacher q_k on the same student prefix. The matrix $\Omega \in \mathbb{R}^{|\mathcal{S}| \times |\hat{z}|}_{\geq 0}$ has at most one nonzero row and total mass $e \in \{0, 1\}$. This is the implemented allocation, not an additional trainable module.

Student and teacher distributions share the union U_t of teacher top-K vocabulary entries and structural tokens realized in the selected span. A residual category collects omitted probability. The resulting vectors $\tilde{p}_{\theta,t}, \tilde{q}_{k,t} \in \mathbb{R}^{|U_t|+1}$ retain that mass. With stop-gradient sg and $m_t = \beta \,\mathrm{sg}[\tilde{q}_{k,t}] + (1-\beta)\tilde{p}_{\theta,t}$,

$$D^{(k)}_{\beta,t} = \beta \,\mathrm{KL}(\mathrm{sg}[\tilde{q}_{k,t}] \| m_t) + (1-\beta)\,\mathrm{KL}(\tilde{p}_{\theta,t} \| m_t), \qquad K = 128, \quad \beta = 0.5. \tag{6}$$

Let $\mathcal{K}_k$ contain structural positions (fields, coordinates, relation labels, action keywords, and delimiters). For $t \in \widehat{T}_k$,

$$w_{k,t} = \frac{1 + \alpha \mathbb{1}[t \in \mathcal{K}_k]}{\sum_{j \in \widehat{T}_k} (1 + \alpha \mathbb{1}[j \in \mathcal{K}_k])}. \tag{7}$$

Uniform and weighted variants use $\alpha = 0$ and $\alpha = 2$. If $z^\star$ serializes references and $T^\star_k$ is its nonempty stage span,

$$\ell_{\mathrm{ref}}(r) = -\frac{1}{|\mathcal{S}|} \sum_{k \in \mathcal{S}} \frac{1}{|T^\star_k|} \sum_{t \in T^\star_k} \log p_\theta(z^\star_t \mid I, x, z^\star_{<t}). \tag{8}$$

The minibatch teacher loss is averaged over eligible examples and is zero when none qualify. Sampling, teachers, support selection, and routes are fixed within an update; no policy gradient passes through the sampled response. The residual probability $1 - \sum_{v \in U_t} p(v)$ retains omitted mass. Teacher positions count scored logits, including distributed padding, separately from eligible responses, useful tokens, and wall time.

B.2 WHY NORMALIZE THE CORRECTION SPAN?

For eligible responses, Ω has unit mass on the selected span. Reference learning assigns equal aggregate weight to active stages. Both normalizations control loss coefficients independently of serialization length.

The transfer term also has a length-independent bound. For $0 < \beta < 1$, define $h(\beta) = -\beta \log \beta - (1-\beta)\log(1-\beta)$. Since $m_t(v) \geq \beta \tilde{q}_{k,t}(v)$ and $m_t(v) \geq (1-\beta)\tilde{p}_{\theta,t}(v)$,

$$0 \leq D^{(k)}_{\beta,t} \leq h(\beta), \qquad 0 \leq \ell_{\mathrm{KD}} \leq e\, h(\beta). \tag{9}$$

The upper bound follows by substituting the two mixture inequalities into the KL terms and using $\sum_{k,t} \Omega_{k,t} = e$; nonnegativity follows from KL nonnegativity. At $\beta = 0.5$, it becomes $e \log 2$ in natural-log units. This bounds loss magnitude, not gradient norms or downstream error.

B.3 ROUND-LEVEL COUPLING

Let $\mathcal{U}_{\mathrm{tr}}$ denote training sources and $\mathcal{V}^{(n)}$ the verification ensemble used in round n. The ensemble includes the strong-model verification stage and the generalist models used for cross-checking. Let $\mathcal{C}^{(n+1)}$ be the candidates generated or repaired with auxiliary student signals:

$$\begin{aligned}
\mathcal{C}^{(n+1)} &= \mathsf{GenRepair}(\mathcal{U}_{\mathrm{tr}}; p_{\theta^{(n)}}), \\
\Delta\mathcal{D}^{(n+1)} &= \{\tilde{r} \in \mathcal{C}^{(n+1)} : \mathrm{Accept}(\tilde{r}; \mathcal{V}^{(n)}) = 1\}, \\
\mathcal{D}^{(n+1)} &= \mathcal{D}^{(n)} \cup \Delta\mathcal{D}^{(n+1)}, \\
\theta^{(n+1)} &= \mathsf{Train}_{\Omega}(\theta^{(n)}; \mathcal{D}^{(n+1)}, Q^{(n)}).
\end{aligned} \tag{10}$$

Here $\mathrm{Accept}(\cdot; \mathcal{V}^{(n)})$ applies Eq. 1 using the round's verification ensemble, and the union counts unique records. Student signals assist candidate generation or repair, while admission remains governed by Eq. 1. The symbols distinguish the verification role of $\mathcal{V}^{(n)}$ from the distillation role of $Q^{(n)}$. Distillation specialists remain frozen within each training run. Accepted records may support specialist refresh between runs. Evaluation targets remain outside this process; the experiments measure its component effects.

C TRAINING AND EVALUATION PROTOCOLS

Checkpoints and study roles. The main table reports the final ANVIL-MOPD on seven public tasks; RoboSpatial provides an additional coverage control. The coverage study continues an earlier source-filtered checkpoint with replay, vacancy, or mixed supervision, while routing continues a common linked-SFT checkpoint. Composition and data-quality comparisons start from the base model. The image interventions compare the base and the earlier source-filtered ANVIL checkpoint; box interventions compare the Relation specialist and joint SFT. Plan verification additionally evaluates the Action specialist. Throughout the appendix, ANVIL † denotes the *earlier source-filtered checkpoint, not the final main-table model*. The Qwen3.6-27B repair model has a separate production role: both repair arms process the same 885 failures with one call per candidate.

Evaluation and uncertainty. Local comparisons fix prompts, decoding, and scorers; published values retain the modalities and subsets marked in the main table. Strict linked success requires valid serialization and correct dependent outputs; plan validity uses exact VALID/INVALID classification. Malformed or incomplete responses remain in the denominator. Paired scene-bootstrap intervals use 5,000 resamples, keeping all records from a sampled physical scene together and using identical scene draws for the compared models. For routing, per-image outcomes are first averaged across the two recorded seeds. The resulting intervals condition on those seeds and do not estimate variability over unobserved initializations.

Record-level aggregation on ANVIL-EVAL. The ANVIL-EVAL column in Table 1 reports the record-weighted mean over the held-out evaluation partition, excluding training-only QA replay. Let $\mathcal{D}_{\mathrm{eval}} = \{r_i\}_{i=1}^{N_{\mathrm{eval}}}$ contain $N_{\mathrm{eval}} = 580{,}000$ records. For record r_i, let d_i denote its native form, $\hat{y}_i$

the model prediction, and $y_i^\star$ the reference. Its score $s_i = g_{d_i}(\hat{y}_i, y_i^\star) \in [0, 1]$ is computed by the corresponding native-task evaluator. Malformed, missing, or incomplete predictions receive zero and remain in the denominator. The aggregate score is

$$S_{\text{eval}} = \frac{100}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} s_i. \tag{11}$$

Each record contributes equally, regardless of its output length. Consequently, forms contribute in proportion to their evaluated record counts; the score is not an unweighted average across the eleven forms or across scene images. The reported 65.35 for ANVIL-MOPD is this record-level aggregate, expressed on a 0–100 scale.

Controlled-panel metrics. The mechanism studies use separate task-specific panels, whose scores are reported with their own denominators rather than substituted for the aggregate ANVIL-EVAL score. For a linked evaluation of N images, strict success can be written as

$$S_{G\to S} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}[\text{valid serialization}_i]\mathbb{1}[\text{Ground correct}_i]\mathbb{1}[\text{Relation correct}_i]. \tag{12}$$

Missing or incorrect stages contribute zero. Tag-closure analysis changes only recognition of the saved stage boundaries; it does not replace predicted geometry or relation labels. For pointing, any-point scoring succeeds when at least one returned point satisfies the task criterion, whereas first-point scoring evaluates only the first point in the same saved response. Plan accuracy uses all $1{,}012$ judgments, with 506 examples in each class.

Training configurations. Table 6 collects the optimization settings for the controlled studies. These local continuation and fresh-base runs are distinct from the final model reported in the main table.

Table 6: **Optimization settings for the controlled training studies.** Batch denotes the configured effective batch size in records. Coverage exposure counts and realized teacher positions are reported separately.

Study	Updates	Batch	Learning rate	Seeds
Coverage continuation	303	24	2×10^{-5}	43, 44
Atomic / linked source SFT	512	4	10^{-4}	17
Routing continuation	256	4	5×10^{-6}	17, 23
Matched-size / nested-pool SFT	256	4	10^{-4}	17

Coverage controls. Coverage arms use LoRA rank/scaling 16/32 and share a 360-record development split. Single-source arms repeat 3,656 replay or 3,608 vacancy records to match the reported 7,264 mixed-arm row exposures. Terminal checkpoints are fixed before evaluation. Unique images and realized tokens are not matched. A representative mixed run takes 2.37 L20 GPU-hours. Repetition changes the source composition at the stated exposure budget.

Linked SFT and routing. Ground-only, Relation-only, atomic-mixture, and linked SFT use LoRA rank/scaling 64/128, dropout 0.05, a frozen vision encoder, a maximum sequence length of 3,072, and at most 1,536 image tokens. Cosine scheduling uses 3% warmup; checkpoint 512 is fixed for evaluation. Routing continues the common linked model with frozen Ground/Relation teachers, temperature 0.9, and a maximum response length of 256 tokens. Both loss coefficients equal one, $\lambda_{\text{ref}} = \lambda_{\text{KD}} = 1$. Flat and semantic-uniform allocation use $\alpha = 0$; structural, declared, and semantic-weighted variants use $\alpha = 2$. Teacher-scored positions are measured separately (Table 10); the response-length limit is common across selectors.

Controlled repair and data-quality configurations. The local Qwen3.6-27B repair study uses greedy decoding and a 512-token limit. Both repair arms receive the original answer and one call per failed candidate. Named feedback additionally gives the first failed category, checked in the order serialization, typed fields, localization, and depth order, without reference coordinates or ranks. In

the matched-size comparison, each training arm contains 1,137 records matched by physical-scene counts, with the fresh-base recipe in Table 6. Individual images and tokens can differ. The nested-pool subsets and their record counts are reported separately in Table 14.

Failure-localized repair prompt. In the controlled repair study, the named-feedback prompt appends the previous answer and the failed category to the original task. For a stored grounding failure, the added instruction is:

> *The verifier reports: G: localization does not match the requested entities. Correct this failure and return a complete response.*

The generic-retry arm instead requests review and correction of the previous answer without naming the failed stage. Both retain the original task, and the feedback does not supply the reference boxes or target depth ranks. Aggregate repair results are reported in Table 12.

Table 7: **Point-cardinality sensitivity on the common 286-item VABench subset with identical image exclusions and saved responses.** †: earlier source-filtered ANVIL checkpoint.

Model	Any-point (%)	First-point (%)
Base	42.31	36.71
ANVIL †	**81.12**	**71.33**

The subset excludes the union of 14 exact/near-image candidates identified against earlier corpora, identically for every model and without using correctness. This historical image-exclusion check is distinct from the scene/image partitioning used for the held-out corpus. On the full 300 cases, source-filtered and base models return 8.92 and 1.90 points per response on average, motivating the first-point comparison.

D COMPLETE CONTROLLED COMPARISONS

D.1 COVERAGE AND RETENTION

Table 8: **Source-composition controls (%), seed 43 / seed 44.** Updates and row exposures are matched; realized tokens are not. Bold identifies an arm that leads in both seeds.

Benchmark	Replay	Vacancy	Mixed
Where2Place	38.00 / 36.22	**69.00 / 69.00**	68.00 / 66.71
RoboSpatial context	44.14 / 41.74	**63.11 / 62.30**	62.30 / 59.84
VABench-Point	79.67 / 80.33	77.67 / 76.67	80.67 / 78.33
BLINK depth	**90.32 / 90.32**	86.29 / 88.71	89.52 / 89.52
ERQA	**50.17 / 49.83**	46.69 / 46.34	48.08 / 48.08
EmbSpatial	**78.43 / 78.24**	76.95 / 77.42	77.91 / 77.77

Mixed supervision improves placement over replay in both runs; vacancy alone maximizes placement while weakening several retained capabilities. Mixed training does not dominate all tasks. Increasing only the source-filtered model's Where2Place response budget from 256 to 512 changes success from 35.16% to 32.96% on the same 100 cases.

D.2 CORRECTION-STAGE SELECTION

VQA retention compares post- and pre-distillation performance on AI2D, ScienceQA, and TextVQA panels of 309, 202, and 500 examples. A ratio can exceed 100% when a panel improves.

Semantic-uniform exceeds flat in both seeds, with a paired gain of 2.13 percentage points (95% CI: [0.39, 3.90]). Both use uniform token weights, supporting selective teacher–span allocation under common training settings. Table 9 additionally compares routing configurations and token-weighting choices; stage normalization is shared across configurations.

Table 9: **Observed paired routing contrasts on strict Ground–Relation success.** Differences are percentage points; 95% scene-bootstrap intervals condition on the two recorded training seeds. Per-seed values and teacher-position counts appear in Table 10.

Comparison	Δ (pp)	95% CI
Semantic uniform − flat	**+2.13**	[0.39, 3.90]
Semantic uniform − declared	+0.53	[-0.61, 1.63]
Semantic weighted − uniform	-0.53	[-1.57, 0.42]

Table 10: **Per-seed routing observations.** Strict G→S success on 282 images / 50 scenes and sampled VQA retention, reported as seed 17 / seed 23. Teacher positions count scored logits including distributed padding; realized budgets are measured, not matched.

Method	G→S (%)	VQA ratio (%)	Teacher positions
Flat declared-teacher	69.50 / 70.92	100.32 / 99.96	131,334 / 131,704
Structural route	71.28 / 70.92	100.12 / 99.84	221 / 74
Declared route	71.63 / 71.99	100.43 / 99.70	56,267 / 56,449
Semantic (uniform)	**72.70 / 71.99**	99.70 / 99.97	47,648 / 48,213
Semantic (weighted)	71.99 / 71.63	99.77 / 99.85	48,066 / 48,130

D.3 Repair yield and downstream learning

Table 11: **Observed filtering counts in the shared 2,011-candidate pool.** Each percentage uses all initial candidates as its denominator. Source correctness is evaluated against source references, not human precision.

Condition	Candidates retained	Fraction of initial pool (%)
Initial candidates	2,011	100.00
Parseable	1,998	99.35
Pass typed checks	1,965	97.71
Source-correct	1,126	55.99

The filtering and repair denominators differ. Structural retention uses all $2{,}011$ initial candidates; conditional source correctness after typed checks is $1{,}126/1{,}965 = 57.30\%$. The 885 initial failures form the repair pool. Accepting only source-correct outputs makes conditional acceptance precision 100% by construction, so that number cannot independently calibrate a verifier.

Table 12: **Observed repair productivity.** Both arms receive the same 885 rejected candidates and make one call per candidate. GPU-hours measure allocated accelerator time; source-correct recoveries supply the yield numerator.

Repair arm	Correct	Yield (%)	GPU-hours	Correct / GPU-hour
Generic retry	11	1.24	2.7664	3.98
Named feedback	**111**	**12.54**	2.8186	**39.38**

Named feedback recovers 111 correct records versus 11 for generic retry at similar GPU cost. Matched-size feedback training reaches 59.22% strict success versus 56.38% for raw data, a difference of 2.84 points [-1.41, 6.88]. Repair yield improves, but the downstream learning interval includes zero and nested-pool performance is nonmonotonic (Table 14).

Table 13 makes this concentration explicit: named feedback recovers $65/78$ depth-order failures but only $42/761$ grounding failures. The overall multiplier therefore depends on the composition and information content of the rejected pool.

A retained grounding failure. The repair logs also preserve an unsuccessful case with two picture instances. The initial response passes syntax and typed checks but fails the source-based grounding

Table 13: **Source-correct repairs by initial failure category.** The categories partition the rejected pool according to the first failed check; each row reports recovery counts and within-category percentages in both arms.

Initial failure	Candidates	Generic retry	Named feedback
Serialization	13	0 (0.00%)	1 (7.69%)
Typed fields	33	2 (6.06%)	3 (9.09%)
Grounding	761	7 (0.92%)	42 (5.52%)
Depth order	78	2 (2.56%)	**65 (83.33%)**
Total	885	11 (1.24%)	**111 (12.54%)**

criterion, while its depth order passes. Both generic retry and named grounding feedback return the same boxes as the initial response. This case complements the successful relation repair in Figure 4: localizing a failure can guide another attempt without guaranteeing that the model will correct it. The selected cases illustrate failure modes; their frequency is assessed by the corresponding quantitative studies.

Cost accounting. With 885 calls per arm, the per-call yield ratio is $111/11 = 10.09$. Allocated GPU time gives $(111/2.8186)/(11/2.7664) = 9.90$. These production costs exclude subsequent student training.

Table 14: **Nested source coverage at fixed updates.** Scene-stratified subsets share fresh-base initialization, 256 updates, 1,024 row exposures, and seed 17. Strict G→S success (%) uses 282 images / 50 scenes; brackets give 95% scene-bootstrap intervals. Tokens are measured.

Pool fraction	Unique records	Input tokens	Success (%)
25%	284	1,524,912	59.93 [49.47, 69.73]
50%	568	1,532,961	58.87 [48.28, 68.77]
100%	1,137	1,535,783	59.22 [50.00, 68.42]

At fixed updates and exposures, pool size changes the available unique records; realized tokens remain unmatched. The seed-17 results establish neither monotonic data scaling nor a reliable matched-size learning gain.

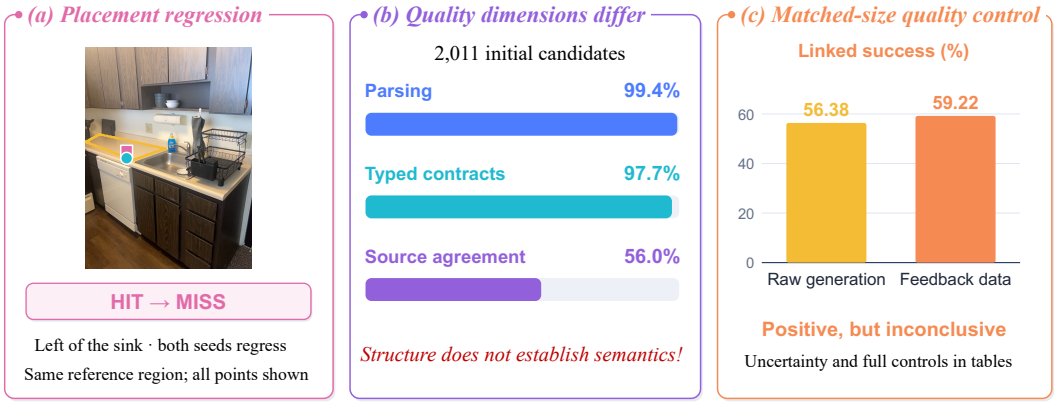


Figure 7: **Failure cases and limits of quality-based learning.** (a) A saved placement example regresses after adaptation in both seeds; selected cases do not estimate failure frequency. (b) Parsing, typed validity, and source correctness all use the same 2,011 initial candidates as the denominator. Conditional source correctness among the 1,965 typed-valid records is reported separately in the text. (c) Matched-size feedback training improves the point estimate by 2.84 points, but the paired interval $[-1.41, 6.88]$ includes zero.

D.4 Construction and rechecking cost

A single CPU pass reads 598,892 lines: 598,891 parse and one is malformed. Parsed records have no violations of the covered structural rules. The pass includes reading, parsing, hashing, and rule evaluation, but excludes image loading, generation, and semantic review. Conformance of retained records does not estimate raw-generation yield or corpus-wide semantic precision. Aggregate throughput is approximately 45.2K records/s. This timing study covers eight historical storage types and measures structural rechecking only. The current data flywheel generates and verifies eleven formats; model-based semantic verification is outside this CPU timing.

Table 15: **Stored-record rechecking: single-run CPU wall time, no model calls.**

Shard	Parsed records	Malformed lines	Time (s)
0	167,459	0	3.81
1	163,286	0	3.61
2	167,682	0	3.65
3	100,464	1	2.18
Total	598,891	1	13.26

E Task Structure and Evidence Dependence

Linked task construction. Eight frames from each of 1,201 ScanNet training scans give 9,608 candidates. After RGB-D extraction, eight CPU workers construct 2,463 linked records from 447 scenes in 325.25 seconds using registered depth and instance geometry. Other frames lack qualifying depth-separated pairs. Visible names identify 856 pairs; interior clicks disambiguate 1,607. Evaluation uses 282 images from 50 held-out scenes. Linked and atomic-mixture training use 3.07M and 2.79M tokens at equal updates. Closing saved stage delimiters preserves JSON contents, geometry, labels, ordering, and finish reasons. Atomic-mixture success rises from 1.77% to 44.68%; linked success remains 69.86%, leaving a 25.18-point gap. This linked-interface advantage coexists with weaker atomic-task performance (Table 16).

Table 16: **Atomic and linked source training.** All four arms use 512 updates and 2,048 row exposures; evaluation uses 282 images from 50 scenes. Detection and Depth use atomic-task prompts, whereas the linked columns use Ground–Relation prompts. Scores are percentages. *Tags closed* changes only saved closing delimiters; training tokens are recorded rather than matched. Bold marks the highest score in each column.

Training view	Atomic tasks		Linked tasks	
	Detection	Depth	Strict G→S	Tags closed
Ground only	**84.75**	0.00	18.79	50.71
Relation only	20.60	**94.68**	0.00	0.71
Atomic mix	84.04	94.33	1.77	44.68
Linked	25.71	0.00	**69.86**	**69.86**

Visual and upstream interventions. Blank inputs are constant-gray images; shuffled images come from another scene, with instructions and targets fixed. Upstream interventions replace reference boxes with predicted or shifted boxes while preserving images and depth targets. These controls establish dependence on supplied evidence, not an internal reasoning mechanism. Reference boxes provide clean context; predicted boxes expose upstream model errors; shifted boxes introduce a controlled perturbation. All three specialist–joint intervals include zero, so their point estimates do not establish a specialist advantage.

Action-order verification. Each of 506 ALFRED trajectories supplies a valid plan and an invalid permutation with identical actions and length. Models receive the initial image, complete Boolean

Table 17: **Visual-input interventions.** Full-split scores (%) with questions and targets fixed. Parentheses give sample counts; failures remain in the denominator. †: earlier source-filtered ANVIL checkpoint, not the final model.

Model	Benchmark	Original	Blank	Shuffled
Base	ANVIL-BENCH-G (1,463)	56.87	3.69	3.69
Base	BLINK (124)	90.32	48.39	47.58
Base	RoboSpatial (122)	44.12	2.79	1.80
ANVIL †	ANVIL-BENCH-G (1,463)	57.35	7.66	3.49
ANVIL †	BLINK (124)	89.52	51.61	48.39
ANVIL †	RoboSpatial (122)	33.60	0.30	4.50

Table 18: **Upstream localization interventions.** Depth accuracy (%) on 282 images from 50 scenes with fixed depth targets. Differences are specialist minus joint, with 95% paired scene-bootstrap intervals (5,000 draws). Both models use 512 source-SFT updates.

Upstream boxes	Relation specialist	Joint	Specialist − joint
Reference	94.68	94.33	+0.35 [-1.79, +2.43]
Student-produced	93.62	92.20	+1.42 [-1.54, +4.38]
Shifted	84.04	82.62	+1.42 [-2.13, +4.91]

Table 19: **State-conditioned plan verification.** Accuracy (%) on 1,012 balanced judgments from 506 ALFRED trajectories and 92 scenes. Each valid plan is paired with an invalid ordering of the same actions; 95% intervals resample scenes. State and rules are supplied. †: earlier source-filtered ANVIL checkpoint, not the final model. Bold marks the highest point estimate in each accuracy column.

Model	Valid	Invalid	Overall ↑	95% CI
Qwen3.5-9B	40.51	**91.11**	65.81	[62.07, 70.60]
ANVIL †	**59.09**	77.67	68.38	[65.42, 72.57]
Action specialist	55.53	89.53	**72.53**	[68.22, 78.41]
Joint SFT	53.56	88.54	71.05	[66.86, 76.65]

state, goal, and source action rules. This pairing removes action occurrence as a sufficient label cue; class-wise scores expose the remaining invalid-answer bias. The Action specialist is separate from the final MOPD model. Because state and rules are supplied, the task measures state-conditioned validity rather than visual-only planning or standard ALFRED execution.

Qualitative failures and quality controls. In Figure 7, the saved "left of the sink" prediction regresses after adaptation in both seeds. The remaining panels separately measure structural validity, source correctness, and downstream learning.

F LIMITATIONS AND FUTURE WORK

From predefined interfaces to open-world spatial supervision. The current framework organizes supervision through eleven Ground–Relation–Act forms with explicit entities and dependencies. These interfaces provide a common basis for generation, verification, and learning, while open-world environments introduce spatial concepts and interaction requirements beyond a predefined inventory. A broader direction is to develop extensible spatial interfaces that incorporate new relations, affordances, and temporal dependencies as agents encounter unfamiliar situations. The central challenge is to expand what the data flywheel can express while preserving the connection between each target and the scene evidence that supports it.

From specialist consolidation to compositional spatial competence. Anvil studies knowledge transfer within specified spatial tasks and stage dependencies. General embodied intelligence also calls for reusing that knowledge when familiar capabilities appear in unfamiliar combinations.

Future work could extend specialist-to-generalist learning toward flexible composition across goals, environments, and embodiments, so that grounding, relational reasoning, and action knowledge remain useful both individually and together. This direction connects knowledge consolidation with a broader question: how can spatial expertise become reusable knowledge that supports new tasks with less task-specific supervision?

From knowledge recirculation to learning through interaction. The present framework draws on curated sources and explicit verification signals. Extending knowledge recirculation to embodied experience raises the question of how an agent should decide what to observe, attempt, and learn next. A promising direction is to use spatial uncertainty and encountered failures to guide active data acquisition, while grounding returned supervision in the outcomes of physical interaction. Such a loop could connect data generation with experience collection and specialist refinement with continued generalist learning. The long-term goal is an agent whose growing spatial competence improves both its decisions and its ability to acquire useful knowledge about the world.