MULTIPLE-CHOICE SUPERVISION FOR EMBODIED VISION-LANGUAGE MODELS: A DATA-CENTRIC FINE-TUNING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) can improve the perception and reasoning of embodied vision-language models (VLMs), yet how supervision should be formulated to promote transfer across embodied tasks remains unclear. In particular, comparisons between choice and open-ended training conflate the information supplied by candidate answers with the form of the supervised target. To disentangle these factors, we introduce Embodied Choice Tuning (ECT), a data-centric framework that jointly designs capability coverage, candidate inputs, distractors, and output targets. ECT retains candidate sets with confusable distractors in the input and supervises the correct option with a compact letter target. We instantiate ECT with a corpus of 10,000 examples spanning eight embodied capabilities and 102 question types. From the same visual observations, questions, and correct answers, we derive four supervision formats that isolate candidate availability, distractor semantics, and output targets. We evaluate the resulting models through Embodied Arena on nine established benchmarks in their native task formats. Across three model families, ECT yields the highest aggregate performance among the four formats. On Qwen3.5-4B, this advantage amounts to 3.89 points over the base and 2.27 points over open-ended supervision. Controlled comparisons further show that candidate inputs improve transfer when answer targets are fixed. A controlled distractor intervention shows that confusable distractors outperform clearly mismatched alternatives for all three choice targets. With candidates fixed, supervising answer text, either alone or alongside the letter, does not outperform letter-only supervision. Together, these results support ECT as a practical approach for data-limited embodied SFT: retain confusable semantic alternatives in the input and supervise the final decision with a compact letter target.

1 INTRODUCTION

Embodied vision-language models (VLMs) must distinguish object states, reason about spatial relationships, interpret motion, and select actions consistent with a task. Effective adaptation with limited training data requires both adequate coverage of embodied capabilities and an appropriate formulation of the supervision signal. Consider a robot touching a bottle beside a mug. A correct answer names the contact target; alternative answers also specify the nearby objects and parts that the model must distinguish. How should this information be presented during supervised fine-tuning (SFT) to improve transferable embodied capabilities?

Recent datasets organize visual questions around spatial and embodied capabilities (Brown et al., 2025; Rouhi et al., 2026). Studies of instruction formats, symbol binding, and distractor construction also show that supervision design affects learning outcomes (Kumar et al., 2025; Xue et al., 2024; Le Berre et al., 2022). However, a direct comparison between choice and open formats changes both the input information and the training target. For the same visual observations and correct answers, how does choice supervision affect embodied transfer compared with open-ended supervision? Answering this question requires separating the effects of candidate inputs, distractor construction, and supervised output formats.

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

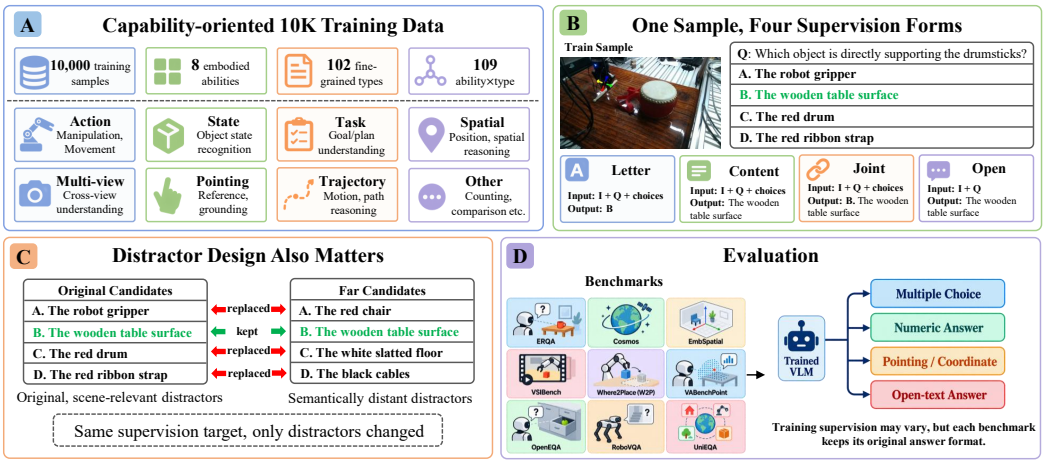


Figure 1: Embodied Choice Tuning and its experimental instantiation. **(A)** ECT organizes training content by embodied capability; our instantiation uses a common 10K corpus. **(B)** Matched formats derived from the same visual observation, question, and correct answer isolate the effects of candidate inputs and output targets. **(C)** ECT pairs confusable Near candidates with a compact Letter target; Far provides a controlled distractor intervention. **(D)** All trained models use each benchmark's native answer format at evaluation.

To address this question, we introduce **Embodied Choice Tuning (ECT)**, a data-centric framework for jointly designing capability coverage, candidate inputs, distractors, and output targets (Figure 1). ECT retains candidate sets with confusable distractors in the input and uses the correct option letter as a compact target. In this study, we instantiate ECT with a 10K corpus organized by embodied capability and question type. From the same visual observations, questions, and correct answers, we derive matched formats that isolate candidate availability, distractor semantics, and output targets. We evaluate the resulting models using Embodied Arena, a comprehensive and unified evaluation platform for embodied AI (Ni et al., 2025). Our evaluation covers nine established benchmarks spanning diverse embodied capabilities, with each benchmark's original task instructions and response requirements. We first assess gains over the common base, then compare choice supervision with open-ended supervision. Paired comparisons test the contribution of candidate text with answer targets fixed, examine distractor construction, and identify the most effective output target within choice supervision.

Across three model families, ECT achieves the highest Embodied Arena aggregate among the four formats. On Qwen3.5-4B, it improves the aggregate by 3.89 points over the base and 2.27 points over open-ended supervision. Candidate text adds training value, and confusable distractors support better transfer than clearly mismatched alternatives. These findings explain ECT's advantage: confusable alternatives enrich the input, while compact letter targets focus supervision on the final decision.

Our contributions are threefold.

- We propose **ECT**, a unified framework that jointly designs capability coverage and choice supervision. ECT combines confusable Near candidates with compact Letter targets.
- We instantiate **ECT with a 10K corpus** spanning eight capabilities and 102 detailed question types. Matched controls isolate the roles of candidate inputs, distractor semantics, and output targets.
- We demonstrate **ECT's effectiveness on Embodied Arena across three model families**. ECT achieves the highest aggregate among the four formats in every tested model and Qwen training run.

2 Related Work

Visual instruction data for embodied capabilities. Visual instruction datasets differ in the capabilities they cover and in how they organize supervision. PRISM structures retail video supervision through a capability ontology, including choice questions with difficult distractors (Rouhi et al., 2026). VLM-AutoDrive combines choice questions, captions, visual QA, and reasoning traces for driving adaptation (Bhat et al., 2026). VisCon-100K places contextual captions and related free-form and choice QA in the same training conversations, while ComVint studies task type and instruction complexity before reformulating synthesized instructions into several response formats (Kumar et al., 2025; Du et al., 2025). Together, these studies motivate attention to both capability coverage and instruction organization when constructing visual training data.

Supervision formats and capability transfer. Supervision formats affect both answer selection and transfer across tasks. Letter targets are used in visual instruction tuning (Liu et al., 2024), while format fine-tuning and stronger binding between option symbols and answer content improve label selection (Bunn et al., 2025; Xue et al., 2024). Studies of text models compare label generation, continuation training on the gold answer alone, and scoring with a classification head (Sood & Savelka, 2026), or train equivalent questions in multiple formats (Liu et al., 2026a). SIMS-V generates open-ended and choice versions of spatial video questions and evaluates transfer across question types (Brown et al., 2025). ReVeL reformulates visual choice questions for reinforcement fine-tuning with verifiable open-ended answers (Liu et al., 2026b). Building on these format comparisons, ECT separates the contributions of candidate inputs and output targets to embodied transfer under shared training content.

Distractor construction and learning. Distractor research examines how incorrect alternatives influence answer selection and learning. ADS-ICT targets matching and distractor similarity biases in visual QA through adversarial data synthesis and counterfactual training; data synthesis alone also improves performance (Wang et al., 2023). Related work in driving samples distractors from other answers and removes candidates during training to reduce text bias (Kulgod et al., 2026). Text QA studies refine distractors for supervised transfer (Le Berre et al., 2022) or select them iteratively for reinforcement learning with verifiable rewards (Guo et al., 2026). ECT examines how distractor construction affects transfer across embodied capabilities when the correct supervision is unchanged.

3 Method

ECT jointly specifies capability coverage, candidate inputs, distractor semantics, and output targets (Figure 1). ECT pairs candidate sets containing confusable distractors, denoted Near, with the correct option letter as a compact target, denoted Letter. To identify the contribution of each component, we construct paired training records that vary one factor while preserving the underlying training content. Each question has four candidates and exactly one correct answer. A record is $r_i = (v_i, q_i, O_i, k_i, a_i)$, where v_i is the visual observation, q_i the common question, O_i the candidates, k_i the correct letter, and $a_i = O_i[k_i]$ its text.

3.1 Paired Data Organized by Embodied Capability

Our experimental instantiation contains 10,000 training examples drawn from 19 data sources. It spans eight capability families: action, state, task, spatial, multi-view, pointing, trajectory, and other visual reasoning. The taxonomy comprises 102 detailed question types and 109 combinations of capabilities and question types. Table 1 summarizes the training allocation across capabilities. Appendix A details the source distribution and construction rules.

The original 10K corpus and its **Near** candidates are constructed by adapting released multiple-choice questions, deriving questions from structured visual annotations, or generating questions from images and task specifications. Each route keeps alternatives within the question's answer domain; model-authored examples also undergo a separate answer-hidden review. Appendix A.2 details the construction and screening procedures.

Capability	Action	State	Task	Spatial	Multi-view	Pointing	Trajectory	Other	Total
Examples	1,054	1,000	1,200	1,300	1,400	1,357	1,500	1,189	10,000
Type entries	16	8	13	22	16	15	7	12	109

Table 1: Distribution of the 10,000 training examples across eight embodied capabilities. The corpus contains 102 unique question types and 109 assignments across capabilities, as some types are associated with more than one capability.

Across formats, each example retains its visual observations, source identity, question, correct answer, and candidate ordering. The shared question preserves answer constraints, such as angle discretization, when candidates are omitted. Correct positions are balanced: A, B, C, and D each occur 2,500 times. We deduplicated the constructed training data against the test sets of all nine evaluation benchmarks.

3.2 Four Supervision Formats

Format	Training input	Target example
Letter	v, q, O	`B`
Content	v, q, O	`wooden table`
Joint	v, q, O	`B. wooden table`
Open	v, q	`wooden table`

Table 2: Paired input and output representations. Letter with Near candidates is ECT's supervision design; the remaining rows are matched controls. Each input also specifies its required answer format.

Letter predicts the correct option letter; **Content** predicts its text; **Joint** predicts a single sequence containing the letter followed by its text. **Open** predicts that same answer text without textual candidates (Table 2). Content and Open have identical target strings and token sequences for all 10,000 questions. Letter, Content, and Joint share candidate sets, with response instructions matched to their targets. All formats contain the same 10,000 examples in the same dataset order; training epochs are matched within each model and seed.

ECT combines Near candidates with Letter targets. Content, Joint, and Open are matched controls. **Letter versus Open** tests whether training to select an option improves embodied transfer more than training to generate its content without candidates. **Content versus Open** holds the correct target fixed and tests whether candidate text provides additional training information. Finally, comparing **Letter, Content, and Joint** with candidates fixed tests whether supervising the letter, answer text, or both yields the strongest transfer.

3.3 Paired Distractor Construction

To examine the role of confusable distractors, we construct a paired **Far** condition from Near. Far replaces incorrect candidates with alternatives that answer the same type of question but are clearly incompatible with the visual evidence. Natural phrasing and comparable answer granularity are preserved. Using superscripts N and F for Near and Far, every paired example satisfies

$$v_i^F = v_i^N, \quad q_i^F = q_i^N, \quad k_i^F = k_i^N, \quad O_i^F[k_i] = O_i^N[k_i] = a_i. \tag{1}$$

Only distractors change; each pair retains exactly the same supervision target. Both conditions retain all 10,000 examples, including those whose answer domains constrain replacement. Appendix A.4 details generation, validation, rules by question type, and replacement coverage.

3.4 Analysis of Supervision Objectives

We analyze three aspects of supervision: the complete target sequences, candidate selection through a letter target, and the weighting of examples by target length.

Sequence supervision. For format $s \in \{L, C, J, O\}$ (Letter, Content, Joint, or Open), the model input x_s contains the visual observation, question, and response instruction. All formats except Open also include candidates (Table 2). Let $p_\theta(y \mid x_s)$ be the probability of output sequence y under model parameters θ. For correct letter k, answer text a, separator d, and response terminator e, the sequence losses ℓ_s are

$$\ell_L = -\log p_\theta(k, e \mid x_L), \qquad \ell_C = -\log p_\theta(a, e \mid x_C),$$
$$\ell_J = -\log p_\theta(k, d, a, e \mid x_J), \qquad \ell_O = -\log p_\theta(a, e \mid x_O). \tag{2}$$

Candidate selection. The output vocabulary extends beyond the valid labels $\mathcal{A} = \{A, B, C, D\}$. Let $m(x_L) = \sum_{j \in \mathcal{A}} p_\theta(j \mid x_L)$ be the total probability of a valid label. Normalizing within this set gives $\pi(j \mid x_L) = p_\theta(j \mid x_L)/m(x_L)$. The loss for the first Letter token decomposes as

$$-\log p_\theta(k \mid x_L) = -\log m(x_L) - \log \pi(k \mid x_L). \tag{3}$$

Minimizing the first term increases the probability of producing a valid label. Minimizing the second favors the correct candidate within that set. Both condition on the complete input, so a single letter can supervise selection among semantically complete alternatives.

Weighting by target length. For a fixed format, let W contain n training examples whose supervised token losses are averaged together. Example i has input x_i and a target of T_i supervised tokens. Let z_{it} be token t, $z_{i,<t}$ its preceding tokens, and $\bar{\ell}_i$ the example's mean token loss. The objective can then be written as

$$\mathcal{L}(W) = \frac{\sum_{i \in W} \sum_{t=1}^{T_i} -\log p_\theta(z_{it} \mid x_i, z_{i,<t})}{\sum_{i \in W} T_i} = \sum_{i \in W} \frac{T_i}{\sum_{j \in W} T_j} \bar{\ell}_i. \tag{4}$$

Thus, each example's mean token loss has coefficient $\alpha_i = T_i / \sum_{j \in W} T_j$, giving $\alpha_i/\alpha_j = T_i/T_j$. A target twice as long has twice the coefficient. When all n examples have equal target lengths, $\alpha_i = 1/n$. Equal coefficients on mean token losses do not imply equal gradient magnitudes. Section 5.5 relates this analysis to measured target lengths and transfer performance.

4 Experiments

Models and training. Our primary model is Qwen3.5-4B (Qwen Team, 2026). Each format starts from the same base and trains on the same 10K examples for two epochs, with seeds 42, 43, and 44. LoRA (Hu et al., 2022) adapts the language module with rank 32, scaling 64, dropout 0.05, learning rate 5×10^{-5}, and effective batch size 64. Visual modules remain frozen. Training uses BF16, gradient checkpointing, and a sequence limit of 32,768 tokens. We evaluate the final checkpoint at step 314. Far experiments use seed 42. An internal 8B model (seed 42) and InternVL3.5-4B-Instruct (seed 43) (OpenGVLab, 2025) each supply a base and a comparison of the four formats. The training settings held fixed within each model family are detailed in Appendix B.

Evaluation. We use the Embodied Arena evaluation pipeline (Ni et al., 2025) on nine benchmarks containing 18,915 questions. ERQA, Cosmos, EmbSpatial, and VSIBench assess embodied, spatial, or temporal understanding; Where2Place and VABenchPoint assess pointing; OpenEQA, RoboVQA, and UniEQA assess open-ended reasoning. Every training condition receives the same questions in each benchmark's native format, media, response instructions, and scoring rules. Models use their native chat templates, greedy decoding, and a maximum of 1,024 new tokens. Preprocessing and video sampling are shared within each model family. Six benchmarks use deterministic scoring. The three open-ended benchmarks use Qwen3.8-Max with thinking disabled and a common rubric; UniEQA averages three judgments per answer. Embodied Arena provides a common evaluation scale across these tasks, so we use the aggregate defined below as our primary measure of overall embodied capability. Appendix C specifies the benchmark aggregations and response parsing.

Primary metric and uncertainty. Let c denote a training condition and $S_b(c)$ its score on benchmark b, expressed on a 0–100 scale. Our primary summary metric is the arithmetic mean of the nine benchmark scores:

$$M(c) = \frac{1}{9} \sum_{b=1}^{9} S_b(c), \qquad \Delta_{c_1, c_2} = M(c_1) - M(c_2). \tag{5}$$

Each benchmark therefore contributes equally, regardless of its number of questions, and a positive Δ_{c_1,c_2} favors c_1. We report all nine scores for every trained model. For Qwen3.5-4B, we report each of the three training seeds and their arithmetic mean; the internal 8B and InternVL models each use one seed. We calculate averages and differences before rounding scores for display. For Letter–Open and Near–Far, we obtain individual 95% confidence intervals from 20,000 paired bootstrap resamples of scene or video groups. Each resample keeps the two conditions aligned and recomputes every benchmark score before calculating M and Δ. These intervals reflect variation across evaluation scenes and videos for fixed checkpoints. Variation across the three Qwen3.5-4B training runs is summarized separately.

5 RESULTS

5.1 ECT IMPROVES AGGREGATE PERFORMANCE ON EMBODIED ARENA

Training	Seed	ERQA	Cosmos	EmbSp.	VSI	W2P	VAB-P	OpenEQA	RoboVQA	UniEQA	Mean
Base	–	39.78	61.96	74.97	47.04	48.10	30.67	66.20	61.77	50.12	53.40
Letter	42	43.01	60.78	80.19	55.39	57.62	36.33	65.92	64.60	51.18	**57.23**
Content	42	45.11	57.45	80.44	54.96	54.52	38.00	65.01	64.45	50.54	56.72
Joint	42	43.66	60.98	80.52	55.82	52.14	35.33	65.36	65.22	51.33	56.71
Open	42	39.91	60.20	75.08	55.52	60.95	36.00	65.40	61.14	47.89	55.79
Letter	43	45.13	60.20	79.40	53.67	52.86	37.67	67.05	65.75	50.58	**56.92**
Content	43	43.98	55.10	76.18	53.28	50.71	35.67	65.30	65.00	51.34	55.17
Joint	43	44.51	60.00	78.98	54.20	55.00	35.00	65.68	65.51	50.89	56.64
Open	43	40.91	60.78	75.36	54.36	52.38	37.33	64.91	61.63	47.44	55.01
Letter	44	42.12	60.98	80.19	55.34	60.48	38.00	65.27	66.39	50.84	**57.73**
Content	44	44.08	58.82	76.37	54.69	53.57	37.67	65.71	66.58	50.98	56.50
Joint	44	43.90	60.59	81.24	56.90	54.52	36.33	64.90	65.59	49.43	57.04
Open	44	38.15	61.76	75.22	55.10	47.14	36.33	64.64	62.41	47.59	54.26
Letter	Avg.	43.42	60.65	79.93	54.80	56.98	37.33	66.08	65.58	50.87	**57.29**
Content	Avg.	44.39	57.12	77.66	54.31	52.94	37.11	65.34	65.34	50.95	56.13
Joint	Avg.	44.02	60.52	80.25	55.64	53.89	35.56	65.31	65.44	50.55	56.80
Open	Avg.	39.66	60.92	75.22	54.99	53.49	36.56	64.99	61.72	47.64	55.02

Table 3: Qwen3.5-4B results on nine Embodied Arena benchmarks. ECT results appear in the Letter rows, which use Near candidates. Avg. averages seeds 42, 43, and 44; Mean weights the nine benchmarks equally. Bold marks the highest aggregate among the four supervision formats in each block. EmbSp., VSI, W2P, and VAB-P denote EmbSpatial, VSIBench, Where2Place, and VABenchPoint. Scores use a 0–100 scale.

The Embodied Arena aggregate places nine established benchmarks on a common evaluation scale and is therefore the primary outcome throughout our analysis. ECT, instantiated with the 10K corpus, Near candidates, and Letter targets, raised Qwen3.5-4B's aggregate from **53.40 to 57.29**, a **gain of 3.89 points** averaged over three training seeds (Table 3). Every format improved over the base in every seed, establishing the utility of the common 10K corpus. ECT achieved the highest aggregate in each seed, with scores of 57.23, 56.92, and 57.73. This repeated ordering across three training runs supports ECT's compact choice supervision as the strongest tested format for overall embodied performance. Appendix D reports variability across seeds.

ECT also achieved the highest aggregate for both additional model families (Table 4). It raised the internal 8B score from 62.58 to 63.71 and the InternVL3.5-4B score from 41.81 to 44.53. Across the three Qwen seeds and two additional models, ECT ranked first among the four supervision formats in all five comparisons. This consistent ordering supports its combination of Near candidates and Letter targets across the three tested model families and different initial performance levels.

Training	ERQA	Cosmos	EmbSp.	VSI	W2P	VAB-P	OpenEQA	RoboVQA	UniEQA	Mean
Internal 8B (seed 42)										
Base	39.72	64.51	83.10	62.54	72.14	60.00	61.95	73.59	45.70	62.58
Letter	44.29	64.71	81.54	62.88	70.71	64.33	62.21	74.28	48.44	**63.71**
Content	37.98	65.10	82.34	61.05	69.76	61.33	62.97	72.86	47.12	62.28
Joint	43.20	65.49	81.54	62.08	70.48	63.67	61.87	74.21	47.21	63.31
Open	40.14	66.47	79.95	59.84	70.48	64.00	62.35	71.50	44.98	62.19
InternVL3.5-4B (seed 43)										
Base	36.13	52.35	68.93	52.73	15.71	13.67	49.40	43.89	43.47	41.81
Letter	37.52	51.96	71.68	51.48	29.76	19.00	50.58	45.25	43.55	**44.53**
Content	35.16	50.39	70.22	52.20	24.05	20.00	49.59	44.03	43.34	43.22
Joint	38.19	52.16	70.99	52.20	25.71	15.33	50.92	44.88	43.71	43.79
Open	35.75	50.39	69.75	52.26	26.90	14.67	50.64	45.03	41.98	43.04

Table 4: Results for the additional model families. Benchmark abbreviations, aggregation, and bolding follow Table 3.

Model / seed	Δ_{L-B}	Δ_{L-O}	95% CI for L−O	Δ_{C-O}	Δ_{L-C}	Δ_{L-J}
Qwen3.5-4B / 42	+3.83	+1.44	[+0.05, +2.78]	+0.93	+0.51	+0.52
Qwen3.5-4B / 43	+3.52	+1.91	[+0.53, +3.32]	+0.16	+1.75	+0.28
Qwen3.5-4B / 44	+4.33	+3.47	[+2.01, +4.96]	+2.23	+1.24	+0.69
Internal 8B / 42	+1.13	+1.52	[+0.45, +2.59]	+0.09	+1.43	+0.41
InternVL3.5-4B / 43	+2.72	+1.49	[+0.05, +2.92]	+0.18	+1.31	+0.74
Qwen three-seed mean	+3.89	+2.27	–	+1.11	+1.16	+0.50
Qwen / 42 / Far	–	–	–	–	+1.47	+1.20

Table 5: Controlled comparisons on the Embodied Arena aggregate. B, L, C, J, and O denote Base, Letter, Content, Joint, and Open. C−O fixes answer targets; L−C and L−J fix candidates; L−O compares compact choice supervision with Open.

5.2 Choice Supervision Transfers beyond the Training Answer Format

ECT improved the Embodied Arena aggregate over Open in all three Qwen training runs and both additional models, with every individual 95% interval above zero (Table 5). The average Qwen gain was **+2.27 points**; the internal 8B and InternVL gains were +1.52 and +1.49. Because every condition is evaluated with each benchmark's native instructions and response format, the aggregate advantage is measured beyond the letter interface used during training. The result therefore supports transfer from ECT's choice supervision to comprehensive embodied evaluation.

5.3 Candidate Inputs Add Value with the Answer Target Fixed

Content and Open were supervised with identical answer strings and token sequences. Adding candidate text increased the Qwen mean by **1.11 points** across seeds (Table 5, C−O). The difference was positive in all three Qwen training runs and both additional models, ranging from +0.09 for the internal model to +2.23 for Qwen seed 44. This comparison locates a useful part of the supervision in the input representation: the alternatives change what the model can compare while leaving the required answer content unchanged.

The positive aggregate direction across these five evaluations shows that the benefit of candidate inputs persists across training runs and model families. Because Content and Open use the same answer target for every example, this paired result assigns their aggregate difference to the information supplied with the question rather than to the answer sequence. Figure 2 provides a qualitative illustration of the corresponding visual evidence allocation.

5.4 CONFUSABLE DISTRACTORS SUPPORT EMBODIED TRANSFER

(a) Complete benchmark scores and paired differences

Target	Candidates	ERQA	Cosmos	EmbSp.	VSI	W2P	VAB-P	OpenEQA	RoboVQA	UniEQA	Mean
Letter	Near	43.01	60.78	80.19	55.39	57.62	36.33	65.92	64.60	51.18	57.23
Letter	Far	45.87	58.63	76.76	54.78	54.05	33.67	66.18	64.14	51.27	56.15
Letter	Δ_{N-F}	-2.86	+2.16	+3.43	+0.61	+3.57	+2.67	-0.26	+0.46	-0.09	+1.08
Content	Near	45.11	57.45	80.44	54.96	54.52	38.00	65.01	64.45	50.54	56.72
Content	Far	42.33	55.88	75.77	53.54	52.86	34.33	63.91	64.88	48.62	54.68
Content	Δ_{N-F}	+2.77	+1.57	+4.67	+1.42	+1.67	+3.67	+1.10	-0.43	+1.92	+2.04
Joint	Near	43.66	60.98	80.52	55.82	52.14	35.33	65.36	65.22	51.33	56.71
Joint	Far	44.94	59.61	76.35	54.09	47.62	30.33	66.72	64.27	50.59	54.95
Joint	Δ_{N-F}	-1.28	+1.37	+4.18	+1.74	+4.52	+5.00	-1.36	+0.95	+0.74	+1.76

(b) Aggregate differences and evaluation uncertainty

Target	Near	Far	Δ_{N-F}	95% CI
Letter	57.23	56.15	+1.08	[+0.01, +2.16]
Content	56.72	54.68	+2.04	[+0.81, +3.28]
Joint	56.71	54.95	+1.76	[+0.42, +3.11]

Table 6: Distractor intervention on Qwen3.5-4B, seed 42. Near uses the confusable distractors from initial corpus construction. Each pair fixes the correct target and training settings.

Near exceeded Far for each output target (Table 6), by **1.08 points for Letter**, **2.04 for Content**, and **1.76 for Joint**. Each individual 95% interval was above zero. The consistent aggregate direction across three targets establishes that distractor construction matters even when incorrect answers receive no direct target loss.

The Content comparison is particularly informative. Both models generate the same correct text, and that text appears in both candidate sets, yet changing the other options changes transfer. The training value of a candidate set therefore depends on its incorrect alternatives as well as its correct answer. Near's advantage supports the confusable candidate inputs used by ECT.

5.5 COMPACT LETTER TARGETS YIELD THE STRONGEST AGGREGATE

Holding the candidate sets fixed, Letter scored above both Content and Joint in all three Qwen training runs and both additional models (Table 5, L−C and L−J). The average Qwen advantage was **+1.16 points over Content** and **+0.50 over Joint**. Under Far, Letter retained gains of +1.47 over Content and +1.20 over Joint. Adding answer text to the target therefore did not improve aggregate transfer in these experiments.

Target	Examples	Tokens	Mean	Median	P95	Max	Top 10% share
Letter	10,000	20,000	2.00	2.00	2	2	10.00%
Content	10,000	87,646	8.76	5.00	24	121	32.44%
Joint	10,000	109,287	10.93	8.00	26	123	27.83%
Open	10,000	87,646	8.76	5.00	24	121	32.44%

Table 7: Distribution of supervised Qwen target lengths over the shared 10K questions.

Letter supplied two supervised tokens per question, including the terminator (Table 7). Content averaged 8.76 tokens, with its longest tenth of examples contributing 32.44% of supervised tokens. Joint averaged 10.93 tokens, with 27.83% from its longest tenth. Equation 4 therefore gives Letter equal coefficients on mean token losses, whereas Content and Joint weight them by target length. Target content and length weighting vary together in these comparisons (Section 3.4).

Candidate content thus improved transfer when provided as input, while reproducing it in the target did not improve on Letter. Together, the performance and target-length analyses support Letter as the most effective output format tested in ECT.

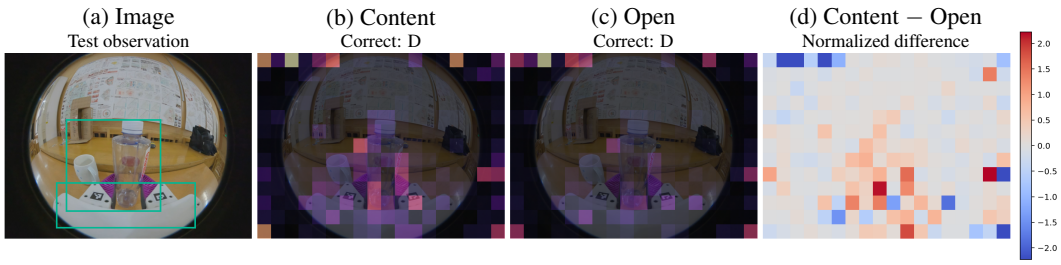


Figure 2: Visual attention for Qwen models trained with Content and Open on ERQA question 290. From left to right: test image, Content attention, Open attention, and the normalized Content–Open difference (red: Content; blue: Open). Both models answer correctly. Full inputs, outputs, and readout details appear in Appendix E.2.

5.6 A Visual Example of Learning with Input Candidates

Question framing can alter visual attention in VLMs (Fan et al., 2026), motivating our analysis of visual evidence allocation after training with different candidate inputs.

Figure 2 qualitatively illustrates the Content–Open difference. Although both models answer correctly, Content assigns 37.29% of image attention to the bottle, mug, and gripper/contact region, versus 31.57% for Open; the region covers 24.79% of the visual grid. This increased focus on objects named by the candidates complements the aggregate gain. Appendix E.2 provides full details.

6 Conclusion

In this work, we introduced ECT, a data-centric framework that jointly designs capability coverage and choice supervision. ECT uses Near candidate sets with confusable distractors and supervises the correct Letter as a compact target. We instantiated ECT with a 10K corpus and matched controls. Across three Qwen training runs and two additional models, ECT achieved the highest Embodied Arena aggregate among the four formats. Controlled comparisons further show that candidates add value when the answer target is fixed, Near distractors outperform Far alternatives across all three choice targets, and reproducing answer text in the target does not improve on compact letter supervision. ECT therefore provides a practical principle for embodied SFT: retain confusable semantic alternatives in the input and supervise the final decision with a compact letter target. The framework can be instantiated with other model families and corpus scales while retaining this supervision principle.

AI Use Statement

We used Qwen models to assist with the data generation procedures described in Section 3 and Appendix A, and OpenAI Codex to assist with literature search and language polishing. The authors reviewed and verified the resulting data, text, and references.

Ethics Statement

The study uses existing visual datasets for embodied question answering. Source media remain subject to their original licenses; derived questions and source indices are intended for release in accordance with those terms. Synthetic candidates and model judgments can inherit source errors or biases. Physical deployment requires validation beyond the offline perception and reasoning tasks evaluated here.

Reproducibility Statement

Section 3 and Appendices A–C specify data construction, training, and scoring. Tables 3, 4, and 6 report all 26 model configurations. The accompanying source package includes scores at full precision, interval endpoints, statistics of target lengths, and a script that rebuilds the numerical tables.

Training records, construction instructions, adapters, and evaluation code are being prepared for release subject to licenses for the source media.

REFERENCES

Mohammad Qazim Bhat, Yufan Huang, Niket Agarwal, Hao Wang, Michael Woods, John Kenyon, Tsung-Yi Lin, Xiaodong Yang, Ming-Yu Liu, and Kevin Xie. VLM-AutoDrive: Post-training vision-language models for safety-critical autonomous driving events. *arXiv preprint arXiv:2603.18178*, 2026. doi: 10.48550/arXiv.2603.18178. URL https://arxiv.org/abs/2603.18178.

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. SIMS-V: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025. URL https://arxiv.org/abs/2511.04668.

Alec Bunn, Sarah Wiegreffe, and Ben Bogin. Fine-tune on the format: First improving multiple-choice evaluation for intermediate LLM checkpoints. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 511–521. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.gem-1.46/.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024. doi: 10.18653/v1/2024.acl-short.33. URL https://aclanthology.org/2024.acl-short.33/.

Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8197–8214, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.546/.

Wan-Cyuan Fan, Jiayun Luo, Declan Kutscher, Leonid Sigal, and Ritwik Gupta. Tinted frames: Question framing blinds vision-language models. *arXiv preprint arXiv:2603.19203*, 2026. doi: 10.48550/arXiv.2603.19203. URL https://arxiv.org/abs/2603.19203.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, et al. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. URL https://arxiv.org/abs/2503.20020.

Xu Guo, Qiming Ge, Jian Tong, Kedi Chen, Jin Zhang, Xiaogui Yang, Xuan Gao, Haijun Lv, Zhihui Lu, Yicheng Zou, and Qipeng Guo. Rethinking multiple-choice questions for RLVR: Unlocking potential via distractor design. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20092–20113. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1003. URL https://aclanthology.org/2026.findings-acl.1003/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Sutej Kulgod, Sean Ye, Sanchit Tanwar, and Christoffer Heckman. Reducing text bias in synthetically generated MCQAs for VLMs in autonomous driving. *arXiv preprint arXiv:2602.17677*, 2026. doi: 10.48550/arXiv.2602.17677. URL https://arxiv.org/abs/2602.17677.

Gokul Karthik Kumar, Iheb Chaabane, and Kebin Wu. VisCon-100K: Leveraging contextual web data for fine-tuning vision language models. In *Data Science: Foundations and Applications. PAKDD 2025, Part VII*, volume 15876 of *Lecture Notes in Computer Science*, pp. 389–401. Springer, 2025. doi: 10.1007/978-981-96-8298-0_31. URL https://link.springer.com/chapter/10.1007/978-981-96-8298-0_31.

Guillaume Le Berre, Christophe Cerisara, Philippe Langlais, and Guy Lapalme. Unsupervised multiple-choice question generation for out-of-domain q&a fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 732–738. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.83. URL https://aclanthology.org/2022.acl-short.83/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

June M. Liu, Shaomian Zheng, He Cao, Dingnan Jin, Qing Cui, and Jun Zhou. Improving cross-format robustness in language models with multi-format training. *arXiv preprint arXiv:2606.11643*, 2026a. URL https://arxiv.org/abs/2606.11643.

Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jing-Shu Zheng, Zheqi He, JG Yao, Xi Yang, Bowen Qin, and Jiajun Zhang. Beyond multiple choice: Verifiable OpenQA for robust vision-language RFT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18859–18869, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Beyond_Multiple_Choice_Verifiable_OpenQA_for_Robust_Vision-Language_RFT_CVPR_2026_paper.html.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Majumdar_OpenEQA_Embodied_Question_Answering_in_the_Era_of_Foundation_Models_CVPR_2024_paper.html.

Fei Ni, Min Zhang, Pengyi Li, Yifu Yuan, Lingfeng Zhang, Yuecheng Liu, Peilong Han, et al. Embodied arena: A comprehensive, unified, and evolving evaluation platform for embodied ai. *arXiv preprint arXiv:2509.15273*, 2025. URL https://arxiv.org/abs/2509.15273.

NVIDIA, Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, et al. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. URL https://arxiv.org/abs/2503.15558.

OpenGVLab. InternVL3.5-4B-Instruct model card. https://huggingface.co/OpenGVLab/InternVL3_5-4B-Instruct, 2025. Accessed 2026-09-22.

Qwen Team. Qwen3.5-4B model card. https://huggingface.co/Qwen/Qwen3.5-4B, 2026. Accessed 2026-09-19.

Amirreza Rouhi, Parikshit Sakurikar, Satya Sai Reddy, Narsimha Menga, Anirudh Govil, Sri Harsha Chittajallu, Rajat Aggarwal, Anoop Namboodiri, and Sashi Reddi. PRISM: A multi-view multi-capability retail video dataset for embodied vision-language models. *arXiv preprint arXiv:2603.29281*, 2026. URL https://arxiv.org/abs/2603.29281.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J. Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, and Yuan Cao. RoboVQA: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652, 2024. doi: 10.1109/ICRA57147.2024.10610216. URL https://ieeexplore.ieee.org/document/10610216.

Bhavesh Sood and Jaromir Savelka. Punching above their weight: Classification-head fine-tuning of tiny language models (TLMs) for verifiable multiple-choice tasks. *arXiv preprint arXiv:2607.03801*, 2026. URL https://arxiv.org/abs/2607.03801.

TJURL-Lab. UniEQA dataset. https://huggingface.co/datasets/TJURL-Lab/UniEQA, 2025. Accessed 2026-09-24.

Zhecan Wang, Long Chen, Haoxuan You, Keyang Xu, Yicheng He, Wenhao Li, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. Dataset bias mitigation in multiple-choice visual question answering and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8598–8617. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.576. URL https://aclanthology.org/2023.findings-emnlp.576/.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4331–4344, 2024. doi: 10.18653/v1/2024.acl-long.237. URL https://aclanthology.org/2024.acl-long.237/.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

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. URL https://proceedings.mlr.press/v270/yuan25c.html.

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. URL https://openreview.net/forum?id=yngvAamNQi.

A Data Composition and Construction

A.1 Capability and Source Coverage

The capabilities in Table 1 organize the questions; detailed question types specify answer domains and construction rules. **Other** contains 1,189 examples across 12 question types covering object and text counting, pairing and partitioning, quantized directions, missing positions in structured layouts, and visual anomalies. Table 8 gives the 19 data sources represented in the training set.

Data source	Training questions
Charades/ActionGenome	1,805
RH20T	1,279
COCO	1,274
SMART-101	1,121
EPIC-KITCHENS	1,062
PixMo_web_image	1,008
ARKitScenes	962
PointOdyssey	711
OpenImages	347
AlphaMaze	174
VIEW2Space	69
BlendedMVS	54
CA-1M	39
CLEVRER	31
Humanbased-AI-Appliance-Knobs	18
MapTrace	17
RoboFAC	15
AnomReason	13
REBOOT	1

Table 8: Counts by data source in the 10K training corpus. A source can contribute examples for several question types.

Each example retains a source identifier, source split, scene/group identifier, media hashes, capability label, detailed question type, common stem, four ordered options, and correct answer. All supervision variants inherit the same record ordering and media references. Removing textual options does not remove labels that are already rendered within the visual observation. Constraints on the answer domain are placed in the shared stem so that the target retains the same meaning across formats. Source indices allow visual media to be referenced when their licenses restrict redistribution.

A.2 Construction of the Original Corpus and Near Candidates

The original corpus is constructed through three complementary routes (Table 9). Released multiple-choice questions retain their published question and correct answer, together with three original distractors. When a source provides more than four options, we retain the correct answer and select three of its incorrect options. This route preserves comparisons already defined by the source task.

Construction route	Examples	Share
Released MCQ adaptation	2,879	28.79%
Annotation-grounded construction	3,740	37.40%
Image-conditioned model construction	3,381	33.81%
Total	10,000	100.00%

Table 9: Construction routes for the original 10K corpus and its Near candidate sets.

Annotation-grounded construction uses task templates or constrained model wording while determining answers from structured annotations. Count questions use published counts or computed set sizes and form alternatives from nearby valid quantities. Point, geometry, and multi-view questions

use persistent tracks, part and instance annotations, camera poses, or depth to identify the correct correspondence, with other marked points, directions, or relations serving as alternatives. Text questions compute answers from source transcriptions, while map questions validate paths against source topology. These procedures keep all candidates within the answer domain defined by the question.

Image-conditioned construction uses Qwen3.7-plus (version `qwen3.7-plus-2026-05-26`) to generate a question, four candidates, a correct index, and supporting visual evidence from the image or ordered frames, the intended question type, and task conditions when needed. The generation request permits abstention when the visual material is insufficient or ambiguous. A separate answer-hidden request reviews answer uniqueness, visual support, and alignment with the intended capability; disputed examples are revised, inspected, or rejected. Across the three routes, source-label validation, deterministic computation checks, format checks, and model-assisted semantic review screen the examples. Each retained question has four distinct options and one correct answer, and correct positions are balanced across A, B, C, and D.

A.3 Construction by Question Type

The construction rules preserve the requested answer domain. For example, a question asking which object was manipulated requires an object name, whereas a question asking what happened requires an action description. Temporal qualifiers such as *before*, *after*, and *immediately* remain part of the common question. Viewpoint, coordinate, unit, and discretization conventions are also shared. The taxonomy determines which distinctions Near candidates express, such as object identity, action feasibility, spatial relations, or temporal order.

A.4 Far Construction, Validation, and Coverage

We construct Far by replacing Near distractors with alternatives that preserve the answer domain but clearly conflict with the visual evidence. Table 10 summarizes the rules across capabilities. Qwen3.8-Max receives the visual input, common stem, correct answer, and requirements for the question type, and generates complete, fluent candidate answers.

Capability	Examples of answer domains	Far construction principle
Action	Observed action, manipulated object, contact prerequisite, feasible tool approach	Use an absent object/tool, an unobserved operation, or a movement incompatible with visible geometry.
State	Contact, support, containment, attribute change, task completion	Use missing parts, incompatible contact configurations, or a state change absent from the relevant frames.
Task	Recorded action order, next step, subgoal progress, recovery plan	Violate visible action dependencies or require unavailable objects while retaining the requested planning granularity.
Spatial	Relative position, orientation, capacity, metric relation	Preserve the relation or numerical domain while making it clearly inconsistent with the scene.
Multi-view	Corresponding objects, viewpoint change, camera relations	Preserve the actual view labels and compare identities or relations incompatible with the scene when the domain permits.
Pointing	Object/part localization, contact points, feasible locations	Retain the coordinate convention and use image locations away from the valid target or region.
Trajectory	Paths, motion direction, temporal correspondence	Use incompatible paths or motion descriptions within the original answer domain; retain cases with finite answer domains.
Other	Counting, pairing, partitioning, quantized directions, missing positions, visual anomalies	Preserve answer type and use quantities, objects, or relationships inconsistent with the observation.

Table 10: Answer domains and Far construction principles by capability. The accompanying data file preserves 109 construction records and their training counts, each corresponding to a capability and a question type.

Structural validation checks that there are exactly four distinct options and that the correct answer, its position, the common stem, and media remain unchanged. A separate review by a model checks that the proposed wrong options are incompatible with the scene while answering the requested kind of question. Targeted inspection handles ambiguous cases. Review labels and explanations are retained as construction metadata rather than supervised responses.

Table 11 reports replacement coverage by capability. Far changes at least one distractor in 6,098 of the 10,000 examples (60.98%). Specifically, 6,017 examples change three distractors, 71 change two, and 10 change one, giving $6{,}017 \times 3 + 71 \times 2 + 10 = 18{,}203$ replacements. The remaining 3,902 preserve their Near candidates: 3,874 have finite answer domains, 12 retain disputed source annotations symmetrically, and 16 reviewed candidate sets coincide with Near. All 10,000 records remain in each training condition.

Capability	Training	Changed	Unchanged	Coverage
Action	1,054	1,044	10	99.05%
State	1,000	972	28	97.20%
Task	1,200	1,043	157	86.92%
Spatial	1,300	250	1,050	19.23%
Multi-view	1,400	387	1,013	27.64%
Pointing	1,357	985	372	72.59%
Trajectory	1,500	353	1,147	23.53%
Other	1,189	1,064	125	89.49%
Total	10,000	6,098	3,902	60.98%

Table 11: Far replacement coverage across capabilities. Changed counts examples with at least one replaced distractor; unchanged examples retain their Near candidates. Coverage is the changed count divided by the training count within each capability.

B Training Settings

Setting	Qwen3.5-4B	Internal 8B	InternVL3.5-4B
Training seeds	42, 43, 44; Far 42	42	43
Training examples	10,000	10,000	10,000
Epochs / final steps	2 / 314	3 / 939	2 / 314
Effective batch	64	32	64
Learning rate	5×10^{-5}	10^{-5}	2×10^{-5}
LoRA rank / scaling	32 / 64	384 / 768	8 / 16
LoRA dropout	0.05	0.05	0.05
Optimizer	AdamW	Fused AdamW	AdamW
Weight decay	0.01	0.01	0
Warmup	3% of steps	10 steps	None
Schedule after warmup	Constant	Constant	Constant
Gradient norm clipping	1	1	1
Precision	BF16	BF16	BF16
Sequence limit	32,768	32,768	No additional trainer truncation
Checkpoint	Final epoch	Final epoch	Final epoch

Table 12: Training settings for each model family. Each format initializes independently from its family's common base.

All three model families adapt their language modules with LoRA and freeze the visual encoder and connector. Qwen's 248 target linear layers cover full-attention, linear-attention, and feed-forward modules. Internal 8B and InternVL target the q/k/v/o and gate/up/down projections. Qwen training uses gradient checkpointing and scaled dot-product attention. Seeds 43 and 44 use two GPUs, a microbatch of eight per GPU, and four accumulation steps, yielding $2 \times 8 \times 4 = 64$ examples per update.

For Qwen, W in Equation 4 covers all examples in an accumulation window. InternVL averages answer token losses within each microbatch before accumulation, so W denotes one microbatch. All formats within each family use the same implementation. Exposure counts and effective batch sizes are matched within each comparison; prompt lengths and counts of supervised tokens differ by design.

The Qwen model responds with thinking disabled. At evaluation, the maximum image budget is 262,144 pixels and videos use up to 64 sampled frames. InternVL uses its native 448-pixel processing with one tile per image or sampled frame. Within each family, image processing, frame sampling, decoding settings, and response extraction are shared across training conditions.

C Evaluation and Statistical Protocol

The benchmark sources are ERQA from Gemini Robotics (Gemini Robotics Team et al., 2025), Cosmos-Reason1 (NVIDIA et al., 2025), EmbSpatial-Bench (Du et al., 2024), VSI-Bench (Yang et al., 2025), Where2Place from RoboPoint (Yuan et al., 2025), VABench from FSD (Yuan et al., 2026), OpenEQA (Majumdar et al., 2024), RoboVQA (Sermanet et al., 2024), and UniEQA (TJURL-Lab, 2025). Table 13 specifies the evaluation subsets and aggregations used in this study.

Benchmark	Questions	Response	Native aggregation
ERQA	400	Choice	Equal mean over eight capability categories.
Cosmos	510	Choice	Example accuracy.
EmbSpatial	3,640	Choice	Example accuracy.
VSIBench	5,130	Choice/numeric	Mean over eight task families, retaining native accuracy or numerical scoring.
Where2Place	100	Point	Equal mean over seen and unseen subsets.
VABenchPoint	300	Point	Point validity for each example.
OpenEQA	1,636	Open text	Example mean after normalizing 1–5 judgments.
RoboVQA	1,921	Open text	Mean over scores for ten task types after normalizing 1–5 judgments.
UniEQA	5,278	Open text	Mean over twelve dimensions, with three judgments per answer.
Total	18,915		

Table 13: The common evaluation panel. UniEQA uses fixed reconstructed question IDs across model conditions. All benchmark scores are expressed on a 0–100 scale before the mean over nine benchmarks is calculated.

Every evaluation fixes question IDs, visual observations, response instructions, extraction logic, and scoring across the compared models. Training with Letter does not impose responses consisting only of letters at evaluation. Open-ended benchmarks still request generated answers, and pointing tasks still request coordinates. This distinction is necessary for evaluating transfer across answer interfaces.

For choice questions, an unambiguous option label or matching answer content identifies the selected answer. Responses that cannot be mapped unambiguously to an option score zero. Pointing tasks use a common coordinate parser and scale convention. If a response gives an explicit bounding box for a task requiring a single point, its center is used consistently. Parsing does not select coordinates using the ground truth. Invalid final answers score zero.

Qwen3.8-Max judges OpenEQA, RoboVQA, and UniEQA with thinking disabled. The judging prompt excludes model identity and training condition. OpenEQA and RoboVQA use ratings $u \in \{1, 2, 3, 4, 5\}$, normalized as $100(u-1)/4$. UniEQA uses its 0–1 rubric, averages three judgments per answer, and scales the dimension mean by 100. Scoring uses the same rubric and aggregation for every training condition.

For Letter–Open and Near–Far, each bootstrap replicate resamples scene or video groups with replacement while keeping the two conditions' outcomes paired. We use 20,000 replicates with random seed 20260923. Within each replicate, we recompute every benchmark score using the aggregation in Table 13 before calculating M and Δ. The reported 95% intervals are individual intervals

without adjustment for simultaneous comparisons. They reflect variation across evaluation scenes and videos for fixed checkpoints. The three Qwen3.5-4B seeds characterize variation due to training separately.

D VARIABILITY ACROSS TRAINING SEEDS

Tables 3, 4, and 6 report the complete benchmark results for all 26 model configurations. Table 14 summarizes Qwen's variation across training seeds. Each standard deviation is computed over the three trained models, separately from the intervals obtained by resampling evaluation data in Tables 5 and 6.

Training	ERQA	Cosmos	EmbSp.	VSI	W2P
Base	39.78	61.96	74.97	47.04	48.10
Letter	43.42±1.54	60.65±0.41	79.93±0.46	54.80±0.98	56.98±3.85
Content	44.39±0.63	57.12±1.88	77.66±2.40	54.31±0.90	52.94±1.98
Joint	44.02±0.44	60.52±0.49	80.25±1.15	55.64±1.36	53.89±1.53
Open	39.66±1.40	60.92±0.79	75.22±0.14	54.99±0.59	53.49±6.97

Training	VAB-P	OpenEQA	RoboVQA	UniEQA	Mean
Base	30.67	66.20	61.77	50.12	53.40
Letter	37.33±0.88	66.08±0.90	65.58±0.91	50.87±0.30	57.29±0.41
Content	37.11±1.26	65.34±0.35	65.34±1.10	50.95±0.40	56.13±0.84
Joint	35.56±0.69	65.31±0.39	65.44±0.19	50.55±0.99	56.80±0.22
Open	36.56±0.69	64.99±0.39	61.72±0.64	47.64±0.23	55.02±0.76

Table 14: Descriptive Qwen mean and sample standard deviation over seeds 42, 43, and 44. The base is a single shared checkpoint. These standard deviations are distinct from paired confidence intervals from evaluation resampling.

E SUPERVISION AND VISUAL READOUT

E.1 MEASUREMENT OF SUPERVISED TARGET LENGTHS

For Table 7, we count supervised target tokens using the Qwen tokenizer, including the terminator and excluding masked prompt tokens. The length histogram yields the total, mean, median, 95th percentile, and maximum over all 10,000 examples. To compute the Top 10% share, we sort targets by length and divide the token count of the longest 1,000 by the total. Content and Open have identical target strings and token sequences, so these statistics agree exactly. Section 3.4 derives the objective decomposition used to interpret the measurements.

E.2 ATTENTION READOUT FOR ERQA QUESTION 290

The example compares Qwen seed 42 Content and Open checkpoints. Both receive the same image and all four options. The question is *Which statement is correct?* The options are:

A. Robot is in contact with the water bottle cap.
B. Robot is in contact with the mug body.
C. Robot is in contact with the mug handle.
D. Robot is in contact with the water bottle body.

The response from the model trained with Content is *Robot is in contact with the water bottle body.* The response from the model trained with Open is *D. Robot is in contact with the water bottle body.* Both select the correct content, D.

We read attention at the final prompt query position before the first answer token and average equally across heads and full-attention layers. Other layer types in the hybrid model are excluded from this matrix readout. Weights on visual tokens are mapped to a 13×18 grid. Let V index the visual tokens,

$N = |V|$ be their count, and w_j be the attention weight for token j. The displayed normalized density is $\rho_j = N w_j / \sum_{k \in V} w_k$. The difference map is $\rho_j^{\text{Content}} - \rho_j^{\text{Open}}$, using a symmetric color range. Red indicates greater Content density and blue greater Open density. The overlays for the individual models use separate color mappings.

The inspected region is the union of normalized windows $(0.25, 0.36, 0.63, 0.85)$ and $(0.21, 0.70, 0.77, 0.94)$, covering the objects and gripper/contact area. It was identified through assistant visual inspection and occupies 24.79% of visual tokens. The corresponding shares of image attention are 37.29% and 31.57%. They are normalized by total image attention rather than attention to both image and text. The case was selected after inspecting the results and serves as a qualitative illustration of the comparison of input candidates.

E.3 Further Comparisons

Qwen uses three seeds; additional models and Far conditions use one. The base comparison evaluates the complete 10K ECT training design as a whole; effects of capability allocation require experiments that vary the data mixture. Calibrated distractor difficulty, counterbalanced option permutations, independent judges, and visual interventions could further resolve learning mechanisms. Evaluation covers embodied perception and reasoning; transfer to action policies remains to be tested.