EKT-ROUTER: EXPERT–EVIDENCE ALIGNMENT FOR EMBODIED VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied vision–language models must answer questions from partial, evolving observations. Their success depends on both the reasoning capability selected and the evidence available to it. Expert routers adapt internal computation, while tool-augmented systems acquire external evidence; choosing these independently can leave a suitable expert without the evidence it needs. We introduce *EKT-Router*, a bounded framework that jointly selects a capability expert, optional knowledge, and perceptual tools. Outcome-derived signatures represent the overlapping strengths of General, Spatial Reasoning, and Spatiotemporal experts. A joint utility scores complete expert–evidence routes under a call budget. After acquisition, a packet verifier admits valid evidence with positive predicted expert-conditioned gain. The selected expert then answers once, reverting to the unchanged observation if all packets are rejected. Across 6 embodied QA benchmarks, *EKT-Router* achieves a mean score of 61.4, the highest among the compared systems, and averages 74.7 on 10 general vision–language benchmarks. With all external evidence disabled, question-conditioned expert selection alone improves over the best fixed expert by 1.75 points. Ablations show that learned routing is the dominant component and that knowledge and perception are complementary. Code is available at `https://anonymous.4open.science/r/EKT-Router-0182/`.

1 INTRODUCTION

Embodied vision–language models (EVLMs) answer questions about evolving scenes from partial observations. Such grounded understanding supports language-guided robotic planning and interaction (Liang et al., 2025). Queries require semantic knowledge, precise geometry, or temporal state tracking; decisive cues may occupy only a few pixels or span several frames (Majumdar et al., 2024; Sermanet et al., 2024; Yang et al., 2025). Success depends on both the selected capability and the evidence it receives. We study *expert–evidence alignment under partial observability*: matching a reusable capability with query-specific evidence, while retaining a generalist path and avoiding an unbounded agent loop.

Existing methods address different parts of this problem. Embodied generalists such as LEO and RynnBrain integrate perception, language, and action capabilities, whereas SpatialVLM, SpatialRGPT, and RoboSpatial use geometric supervision to strengthen spatial reasoning (Huang et al., 2024; Dang et al., 2026; Chen et al., 2024a; Cheng et al., 2024; Song et al., 2025). Mixture-of-experts methods provide conditional parameter paths for efficient specialization (Wu et al., 2024; Shen et al., 2024); ViperGPT and LLaVA-Plus instead compose external vision modules to recover information that a backbone cannot reliably infer (Surís et al., 2023; Liu et al., 2024a). These expert routers focus on selecting a parameter path, whereas the cited tool-augmented systems acquire evidence for a fixed or implicit reasoner. Jointly choosing the capability path and its external evidence makes their dependence an explicit part of the routing decision.

These coupled choices pose two challenges. ❶**Evidence value is expert-dependent.** Expert strengths overlap, and the same knowledge or tool output can provide different gains to different experts. Selecting the expert and evidence independently can therefore overlook useful combinations. ❷**Predicted evidence value need not match the usefulness of the returned packets.** A promising route may still yield incomplete or ambiguous evidence; deciding what to retain requires inspecting packet coverage, confidence, ambiguity, and provenance after execution. The challenge is to account for this

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.

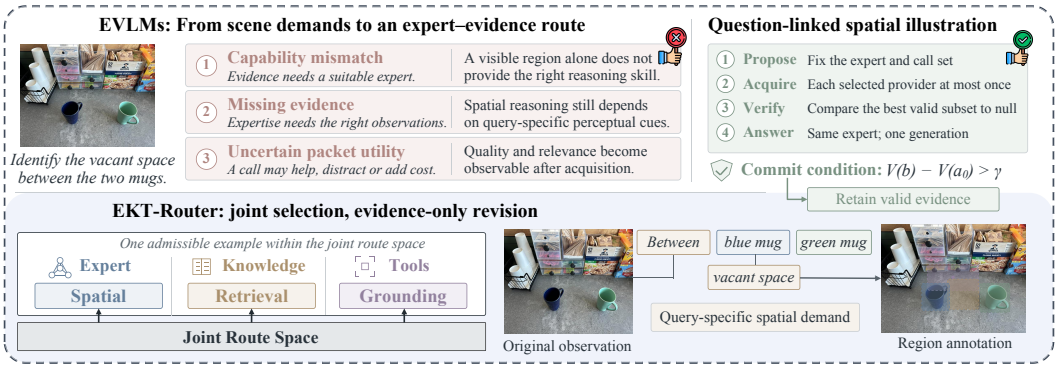


Figure 1: *EKT-Router* couples capability selection with evidence acquisition. Top: three conceptual failure modes of independent decisions. Bottom: the Route Space includes one expert, optional knowledge, and a budgeted tool subset, followed by acquisition, packet verification, and one answer.

newly available information without restarting expert selection or expanding the acquisition budget. Our design separates these uncertainties in time: select a complete expert–evidence route before acquisition, then revise only the evidence decision after inspecting the packets. Conditional evidence use thus requires neither repeated expert selection nor repeated answer generation.

EKT-Router implements this decision in two stages (Figure 1). Offline, independent capability post-training produces three Qwen3-VL-8B LoRA experts: General through continual pretraining (CPT), Spatial Reasoning through spatial supervision, and Spatiotemporal through video supervision. Diagnostic outcomes encode their overlapping strengths as continuous capability signatures, while paired executions supervise a complete-tuple route utility. Online, a frozen Qwen3-VL-2B provides a query-aware control view; the router selects one expert, optional Knowledge, and a budget-feasible Tool subset. Knowledge supplies task priors, while Tools return grounding, depth, or text as typed packets. A bounded verifier selects a valid evidence subset by its predicted score margin over the null reference. The fixed expert answers once from the observation and retained evidence; rejecting every packet recovers the evidence-free path. Our contributions are:

1) We formulate expert–evidence alignment under partial observability as one decision coupling a capability path with optional knowledge and a budget-feasible tool set.

2) We introduce *EKT-Router*, which derives continuous expert signatures and learns complete-route utility from answer outcomes. Typed packet verification refines evidence use after acquisition and preserves an exact evidence-free fallback.

3) Across 16 benchmarks, *EKT-Router* achieves the best embodied QA average among the compared systems while matching strong general VLMs. Matched controls show that expert selection pays off even without external evidence, and ablations identify learned routing as the dominant component and knowledge and perception as complementary.

2 Related Work

Embodied question answering. Early work studied situated question answering and instruction-following navigation (Das et al., 2018; Anderson et al., 2018). Foundation-model benchmarks now evaluate open-vocabulary episodic memory, spatial representation, and interactive or long-horizon reasoning (Majumdar et al., 2024; Yang et al., 2025; Ni et al., 2025). Spatially specialized VLMs improve geometric and grounded reasoning (Chen et al., 2024a; Cheng et al., 2024). The diversity of image, multi-frame, and video questions motivates measuring where specialists are complementary. Existing EQA models nevertheless remain limited in *turning partial visual observations into reliable evidence for spatial and cross-frame reasoning* (Majumdar et al., 2024; Yang et al., 2025).

Routed parameter-efficient specialization. LoRA separates adaptation from frozen backbone weights (Hu et al., 2022); routed extensions distribute examples or tokens across multiple low-rank paths. Representative designs include hierarchical composition, anti-forgetting plugin mixtures,

token-level routing for conflicting data, and sparse multimodal experts (Wu et al., 2024; Dou et al., 2024; Chen et al., 2024b; Shen et al., 2024). Continual and embodied variants further allocate adapters by instruction, layer, or task (Ge et al., 2025; Jia et al., 2025; Luo et al., 2026). RouteLLM learns model selection from preferences (Ong et al., 2025). Route-to-Reason jointly selects a model and a reasoning strategy under a budget (Pan et al., 2025), and MASER routes spatial questions among modality adapters on a shared embodied backbone (Raj & AV, 2026). These methods select models, parameter paths, or internal strategies. EKT-Router *instead scores a complete route: a capability path together with the non-parametric evidence that the path will consume.*

Tool-augmented multimodal reasoning. Modular systems recover missing information through programs, retrieval, or vision APIs (Surís et al., 2023; Lu et al., 2023; Gao et al., 2023). LLaVA-Plus likewise composes external perception modules (Liu et al., 2024a). Recent work develops spatial toolkits (Ruan et al., 2026), call-level credit assignment (Feng et al., 2026), self-regulated visual tools (Chen et al., 2026), and latent control for agent decisions (Ghasemabadi et al., 2026). *Beyond selecting tools,* EKT-Router *distinguishes between acquiring evidence and accepting it, with its utility evaluated for the selected expert.*

3 METHOD

EKT-Router makes a two-stage decision under partial information. The Utility Router first selects an expert and typed Knowledge and Tool actions. Once the packets return, the Packet Verifier revises the evidence decision on the same utility scale, keeping the expert fixed. Verification uses packet attributes, not candidate answers; only the final expert generates an answer. Figure 2 connects this online procedure to offline supervision.

3.1 JOINT EXPERT–EVIDENCE FORMULATION

Given $x = (q, o)$, *EKT-Router* answers question q from a partial visual observation o (an image, multiple frames, or a video) by jointly selecting a reasoning expert and its external evidence. We denote the answer by $\widehat{y}$ and the training reference by y.

We form each expert by applying an independently trained LoRA update to the shared Qwen3-VL-8B backbone with parameters θ_0:

$$\Theta_e = \theta_0 + \Delta_e, \qquad e \in \mathcal{E} = \{\mathrm{gen}, \mathrm{spa}, \mathrm{st}\}, \tag{1}$$

where Δ_e is zero on unadapted weights. General uses continued pretraining; Spatial Reasoning and Spatiotemporal use spatial and video supervision. Their measured strengths overlap rather than form a hierarchy. Routing activates one adapter, with all expert parameters frozen.

Knowledge providers $\mathcal{K}$ supply reasoning templates or retrieved text; Tools $\mathcal{T}$ extract scene evidence through grounding, depth estimation, or OCR. With $\mathcal{K}_0 = \mathcal{K} \cup \{\varnothing_{\mathcal{K}}\}$ allowing no knowledge, the joint route space is

$$\begin{aligned} z &= (e, a) \in \mathcal{E} \times \mathcal{A}, \qquad a = (k, S), \\ \mathcal{A} &= \{(k, S) : k \in \mathcal{K}_0,\ S \subseteq \mathcal{T},\ C_T(S) \le B_T\}. \end{aligned} \tag{2}$$

Here S is a possibly empty tool subset with cost budget B_T, and $C(a) = C_K(k) + C_T(S)$ with zero null costs. The null action $a_0 = (\varnothing_{\mathcal{K}}, \varnothing)$ preserves the original-input path for every expert.

Selected non-null providers, indexed by $\mathcal{I}(a)$, return evidence packets:

$$\begin{aligned} \Omega_a &= \{\omega_j\}_{j \in \mathcal{I}(a)}, \qquad \Omega_{a_0} = \varnothing, \\ \omega_j &= (v_j, q_j^{\mathrm{cov}}, q_j^{\mathrm{conf}}, q_j^{\mathrm{amb}}, p_j, C_j), \end{aligned} \tag{3}$$

where the fields record content, coverage, confidence, ambiguity, provenance, and cost. The verifier inspects acquired packets before deterministic verbalization. Thus routing chooses what to acquire for an expert; verification decides which returned evidence that fixed expert receives.

3.2 OUTCOME-DERIVED EXPERT SIGNATURES

A categorical expert name cannot express overlapping strengths. We instead evaluate each expert without external evidence on a diagnostic collection partitioned into D capability probes $\{\mathcal{G}_d\}_{d=1}^{D}$.

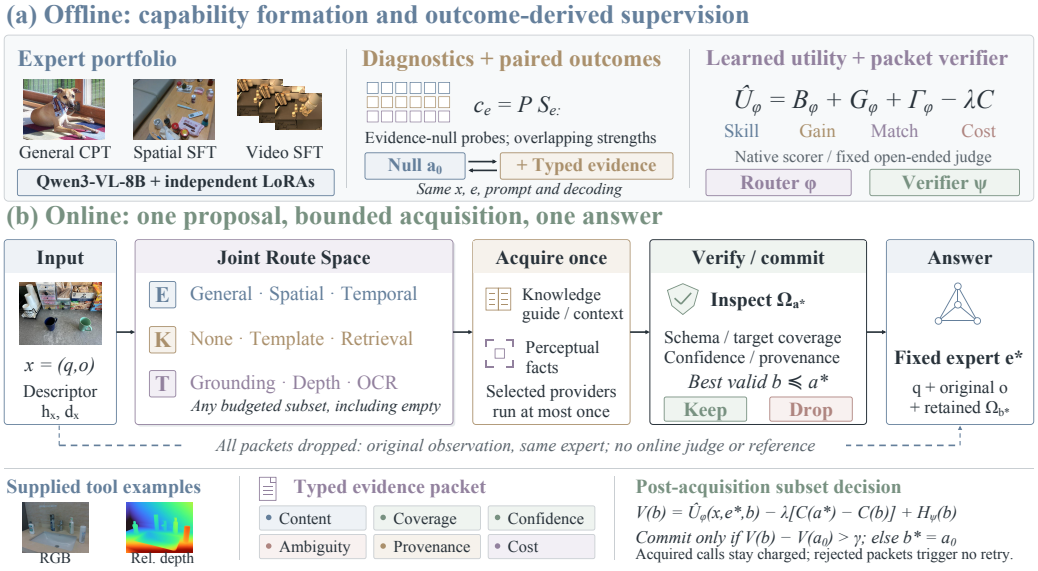


Figure 2: **Overview of *EKT-Router*.** *(a) Offline:* independent adaptations form the expert portfolio; evidence-free diagnostics yield capability signatures, and paired answer outcomes supervise the router and verifier. *(b) Online:* ϕ jointly selects an expert and budgeted evidence actions. Each selected provider runs at most once. ψ scores valid packet subsets against the null reference, retaining evidence only above margin γ. The fixed expert answers once from the original observation and retained evidence; the dashed path preserves the evidence-free fallback.

After normalizing scores within each probe, we form

$$S_{ed} = \frac{1}{|\mathcal{G}_d|} \sum_{i \in \mathcal{G}_d} \widetilde{R}_d(y_i, \widehat{y}_i(e, a_0)) \,, \tag{4}$$

and define the capability signature as $c_e = PS_{e:} \in \mathbb{R}^r$, where P is a learned low-rank projection. Probe-wise normalization makes accuracy, exact match, and answer-judge scores comparable despite their different numerical ranges.

The frozen query encoder maps x to h_x; a learned router mapping predicts demand d_x. We define the evidence-free expert score as

$$B_\phi(x, e) = d_x^\top W_B c_e + b_e. \tag{5}$$

Expert signatures thus derive from diagnostic outcomes, not hand-written levels. Extending the portfolio requires new diagnostics and fitting the affected routing components. General, Spatial Reasoning, and Spatiotemporal remain descriptive names for the current paths.

3.3 Interaction-Factorized Route Utility

We score a route through three distinct terms:

$$\widehat{U}_\phi(x, e, a) = B_\phi(x, e) + G_\phi(x, a) + \Gamma_\phi(x, e, a) - \lambda C(a). \tag{6}$$

Here B measures evidence-free competence, G is an action value shared across experts, and Γ captures input-dependent compatibility between expert and evidence. We anchor $G_\phi(x, a_0) = \Gamma_\phi(x, e, a_0) = 0$ and impose the centering $\sum_{e \in \mathcal{E}} \Gamma_\phi(x, e, a) = 0$, so G is the expert-mean action value and Γ the expert-specific deviation from it.

Let r_a be a learned typed-action embedding with $r_{a_0} = 0$. We set $G_\phi(x, a) = d_x^\top W_G r_a$, then form a raw interaction and subtract its expert mean:

$$\begin{aligned} \overline{\Gamma}_\phi(x, e, a) &= d_x^\top W_\Gamma \big[(W_E c_e) \odot (W_A r_a)\big], \\ \Gamma_\phi(x, e, a) &= \overline{\Gamma}_\phi(x, e, a) - \frac{1}{|\mathcal{E}|} \sum_{e' \in \mathcal{E}} \overline{\Gamma}_\phi(x, e', a). \end{aligned} \tag{7}$$

Here $\odot$ denotes element-wise multiplication. Without Γ, evidence value is independent of the expert. With Γ, the same grounding action can have different predicted value for Spatial Reasoning and Spatiotemporal. We define acquisition gain relative to the selected expert's evidence-free route:

$$\widehat{\Delta}_\phi(x, e, a) = \widehat{U}_\phi(x, e, a) - \widehat{U}_\phi(x, e, a_0). \tag{8}$$

Before acquisition, this expert-dependent margin estimates evidence value without observing the quality of the returned packets.

3.4 Route Supervision from Answers

EQA datasets provide answers but no route labels. For training example i, we execute selected route pairs under identical prompts and decoding, using a benchmark-native scorer whenever one exists and a fixed answer judge only for open-ended evaluation. The normalized realized utility is

$$\begin{aligned} u_i(e, a) &= \widetilde{R}_i(y_i, \widehat{y}_i(e, a)) - \lambda C(a), \\ \delta_i(e, a) &= u_i(e, a) - u_i(e, a_0). \end{aligned} \tag{9}$$

For a fixed expert, δ_i measures the gain from evidence; null executions compare expert capabilities without external evidence.

Using $u_i(z) = u_i(e, a)$ and $\widehat{U}_\phi(x, z) = \widehat{U}_\phi(x, e, a)$ for $z = (e, a)$, let $\mathcal{P}_i$ contain ordered route pairs (z^+, z^-) for which $u_i(z^+) > u_i(z^-) + \epsilon$. We optimize the Bradley–Terry preference loss (Bradley & Terry, 1952)

$$\mathcal{L}_{\text{pref}} = -\sum_i \sum_{(z^+, z^-) \in \mathcal{P}_i} \log \sigma\left(\frac{\Delta \widehat{U}_i}{\tau}\right), \quad \Delta \widehat{U}_i = \widehat{U}_\phi(x_i, z^+) - \widehat{U}_\phi(x_i, z^-), \tag{10}$$

with a lower-weight Huber term $\mathcal{L}_{\text{cal}}$ between predicted and realized gains. The full objective $\mathcal{L} = \mathcal{L}_{\text{pref}} + \alpha \mathcal{L}_{\text{cal}}$ combines route ranking with acquisition-gain supervision. A held-out routing split selects ϵ, λ, α, and the decision margin.

Packet outcomes supervise a residual correction for non-null evidence. For a proposed action a and valid retained subset $b \preceq a$ with $b \neq a_0$, define

$$\rho_i(b \mid a) = \widetilde{R}_i(y_i, \widehat{y}_i(e, b)) - \lambda C(a) - \widehat{U}_\phi(x_i, e, b) + \lambda[C(a) - C(b)]. \tag{11}$$

Thus, $\rho_i(b \mid a) = u_i(e, b) - \widehat{U}_\phi(x_i, e, b)$ measures the discrepancy between realized and predicted route utility. We train $H_\psi(h_x, c_e, r_b, \Omega_b)$ to predict this residual using a Huber loss over valid non-null subsets, while fixing $H_\psi(h_x, c_e, r_{a_0}, \varnothing) = 0$ for the evidence-free reference. The resulting retention margin tracks realized evidence gain when B_ϕ is calibrated to evidence-null outcomes (Appendix A).

3.5 Bounded and Verified Inference

At inference, the router first proposes

$$(e^\star, a^\star) = \arg \max_{e \in \mathcal{E},\, a \in \mathcal{A}} \widehat{U}_\phi(x, e, a). \tag{12}$$

One router pass scores all discrete tuples without generating candidate answers. Only the non-null components of $a^\star$ execute, each at most once, to produce $\Omega_{a^\star}$.

A validity predicate checks schema consistency, target coverage, finite confidence, and provenance; the null subset is always valid. Write $b = (k', S') \preceq a^\star = (k^\star, S^\star)$ when $k' \in \{\varnothing_\mathcal{K}, k^\star\}$ and $S' \subseteq S^\star$. Valid subsets are then rescored after acquisition:

$$\widehat{V}_{\phi,\psi}(b \mid x, e^\star, a^\star) = \widehat{U}_\phi(x, e^\star, b) - \lambda[C(a^\star) - C(b)] + H_\psi(h_x, c_{e^\star}, r_b, \Omega_b). \tag{13}$$

The cost adjustment fixes incurred cost at $C(a^\star)$, which cancels across valid subsets. Acquisition cost matters when choosing a route, but is already sunk when deciding which returned packets to retain. The valid maximizer is committed only if its predicted score margin over the null subset exceeds γ; otherwise $b^\star = a_0$. Knowledge and Tool packets can be rejected independently.

Finally, the fixed expert answers from the original observation and retained packets:

$$\widehat{y} \sim p_{\Theta_{e^\star}}(\cdot \mid q,\, o,\, \text{verbalize}(\Omega_{b^\star}))\,. \tag{14}$$

Empty packets add no text, so $b^\star = a_0$ recovers the same expert's evidence-free path. Verification generates no candidate answer, each provider runs at most once, and the answer is generated once under $\mathbf{1}[k^\star \neq \varnothing_\mathcal{K}] + |S^\star| \leq B_{\text{ext}}$.

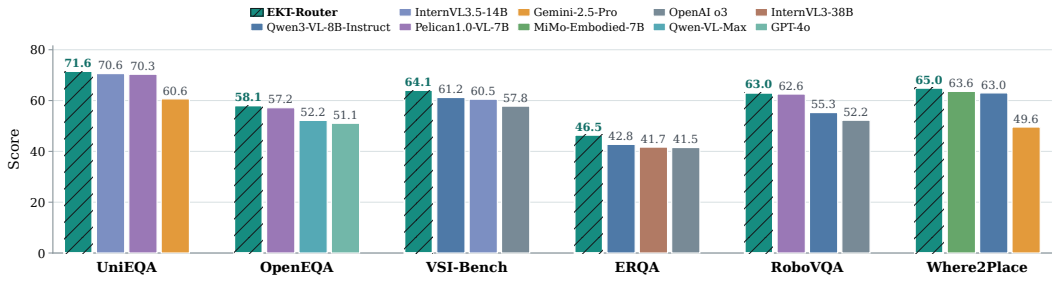


Figure 3: **Embodied QA across six benchmarks (Q1).** Each panel uses its native higher-is-better metric. Results combine reported and reproduced scores under the respective evaluation protocols.

Structural guarantee. Define the induced provider-call bound as $B_{\text{ext}} = \max_{(k,S)\in\mathcal{A}}\{\mathbf{1}[k \neq \varnothing_{\mathcal{K}}] + |S|\}$. Inference uses one route proposal, at most B_{ext} provider invocations, one subset-verification pass, and one answer generation. If no valid subset exceeds the null score by γ, then $b^\star = a_0$ and Equation 14 recovers the selected expert's original-input path. This follows from null validity, finite subset search, and a fixed expert. General remains available through (gen, a_0).

4 Experiments

We design experiments to answer four questions. **Q1:** Does expert–evidence alignment improve embodied QA over strong general, embodied, and proprietary models? **Q2:** Are the experts complementary, and does routing among them help without any external evidence? **Q3:** Which decisions (routing, knowledge, perception, supervision) account for the gain? **Q4:** Does specialization come at the cost of general vision–language competence?

4.1 Evaluation Protocol

The primary suite comprises six embodied QA benchmarks spanning single-image, multi-frame, and video inputs: UniEQA (TJURL-Lab, 2026), OpenEQA (Majumdar et al., 2024), VSI-Bench (Yang et al., 2025), ERQA (Gemini Robotics Team, Google DeepMind, 2025), RoboVQA (Sermanet et al., 2024), and Where2Place (Yuan et al., 2024). We retain each native higher-is-better metric and report the equal-weight mean over the six scores; a separate ten-benchmark suite assesses general reasoning, document understanding, and hallucination.

We compare with Qwen3-VL, InternVL3.5, and the embodied models MiMo-Embodied and Pelican-VL (Bai et al., 2025; Zhu et al., 2025; Hao et al., 2025; Zhang et al., 2025); proprietary models use their respective endpoints. The three experts share frozen Qwen3-VL-8B and identical decoding, with one independently trained LoRA checkpoint active per query; General is CPT-trained, not the unadapted backbone. The descriptor and experts stay frozen during router training, and only the media-disjoint routing split supervises the router.

Knowledge templates cover reference-frame normalization, spatial decomposition, and temporal ordering. Hybrid retrieval returns at most five passages from a capability-indexed corpus. Tools include Grounding DINO (Liu et al., 2024b), Depth Anything V2 (Yang et al., 2024), and OCR; any budget-feasible subset may be selected. Packets retain coordinates, frame indices, text, confidence, and provenance before joint verbalization for one fixed-expert answer.

Training and comparison protocols. The spatial and temporal adapters use scale 256, dropout 0.05, and two epochs of AdamW with learning rate 10^{-4}, weight decay 0.1, cosine scheduling, and 5% warm-up. A media-disjoint 1,020/180 split fixes early stopping, costs, margins, and commitment. Matched runs share prompts, zero-temperature decoding, and generation budgets; native scorers are used when available, and the frozen answer judge sees no route identity. Decision ablations reuse checkpoints, instances, decoding, and scorers; training controls change the named training stage. The six-benchmark comparison combines reproduced, official API, and reported results. Matched controls isolate contributions under a shared evaluation protocol, while cross-system comparisons describe overall performance.

Table 1: **General vision–language evaluation (Q4).** All metrics are higher-is-better. **Bold** identifies EKT-Router; underlining marks column maxima, including ties. Both may apply. MMB: MMBench v1.1; MMS: MMStar; MathV: MathVista-mini; TVQA: TextVQA; CQA: ChartQA; DVQA: DocVQA; HallB: HallusionBench.

	Visual reasoning					OCR / documents			Hallucination		
Model	MMB	MMS	MMMU	AI2D	MathV	TVQA	CQA	DVQA	HallB	POPE	Avg.
Open general VLMs											
LLaVA-OV-7B	80.8	61.7	48.8	82.0	63.2	77.6	80.0	87.5	47.3	87.5	71.6
MiniCPM-V2.6	78.0	57.5	49.8	82.1	60.6	80.1	79.3	90.5	48.1	87.0	71.3
Cambrian-8B	75.9	57.6	42.7	79.7	49.0	71.7	73.3	77.8	41.6	86.5	65.6
InternVL2-8B	79.4	61.5	51.2	83.8	58.3	77.4	83.3	91.6	45.0	86.4	71.8
InternVL2.5-8B	82.2	62.8	56.0	84.5	64.4	78.2	84.7	92.9	49.3	88.5	74.4
Qwen2-VL-7B	80.5	60.7	54.1	83.0	58.2	84.3	83.0	93.9	42.4	87.3	72.7
Qwen3-VL-8B	74.9	42.9	47.0	46.7	32.8	83.1	83.5	95.8	51.3	88.3	64.6
Embodied and spatial specialists											
SpatialBot-3B	58.2	39.1	31.4	56.7	35.0	50.2	48.1	56.3	36.2	81.5	49.3
SpatialRGPT-7B	72.3	52.8	43.5	74.6	49.1	70.5	68.4	78.0	41.0	85.7	63.6
ManipVQA-7B	69.5	50.2	40.3	71.5	46.8	67.0	65.3	75.2	38.8	83.7	60.8
LEO-7B	70.8	51.4	41.7	72.0	47.3	68.8	66.7	76.5	39.5	84.9	62.0
EKT-Router (8B)	**81.9**	**63.2**	**57.8**	**84.3**	**61.5**	**83.9**	**84.2**	**93.8**	**47.9**	**88.3**	**74.7**

4.2 Q1: Embodied QA Results

Finding 1: an 8B router surpasses larger and proprietary systems. *EKT-Router* attains the highest six-benchmark average, 61.4, ahead of embodied specialists (MiMo-Embodied, Pelican-VL), general models up to 38B, and proprietary APIs including Gemini-2.5-Pro and o3, while activating a single 8B expert per query. It ranks first on all six benchmarks, leading the runner-up by +0.4 to +3.7 points, with the largest margins on ERQA (46.5 vs. 42.8) and VSI-Bench (64.1 vs. 61.2). Against its own General path under the matched protocol, *EKT-Router* gains +1.78 on average, driven by RoboVQA (+7.78), ERQA (+3.75), VSI-Bench (+2.94), and Where2Place (+2.00). These tasks require temporal state tracking, multiview grounding, and spatial affordance, precisely where a specialist supplied with grounding or depth packets can resolve cues that a generalist sees but cannot localize.

The gains are selective. On UniEQA (-3.37) and OpenEQA (-2.43), dominated by open-vocabulary semantics and episodic recall, the General path remains stronger; such queries benefit little from geometric evidence, which is why General is kept as an explicit route. Q2 and Q3 decompose the aggregate gain into expert selection and evidence use.

4.3 Q2: Expert Complementarity

Finding 2: no single expert dominates. In Figure 4, General leads UniEQA and OpenEQA, Spatial Reasoning leads RefSpatial and VABench-Point, and Spatiotemporal leads RoboVQA and VSI-Bench; these rank reversals persist from 2K to 8K steps rather than converging to one best checkpoint. Specialization thus creates stable, query-dependent complementarity, the precondition for routing to pay off. The diagnostic also shapes the portfolio: the Pointing/Grounding candidate overlaps with Spatial Reasoning, so localization is exposed as a Tool that any expert can consume. Unlike expert-only routers such as MoLE or MASER, evidence useful to several experts is shared as an action rather than duplicated as another expert.

Finding 3: routing alone converts complementarity into accuracy. To separate expert choice from evidence, we disable Knowledge and Tools for every candidate on nine benchmarks (the six primary tasks plus Cosmos, RefSpatial, and VABench-Point). Question-conditioned selection reaches 60.24 versus 58.49 for the best fixed expert, a +1.75 gain whose media-grouped 95% CI $[1.25, 2.26]$ excludes zero. With no external evidence involved, this gain is attributable to the router: signatures and utility recover which expert suits each query, which no globally best model can.



Figure 4: **Expert complementarity (Q2).** *Left:* the three deployed experts reverse rank on six representative benchmarks from the matched control. *Right:* candidate-checkpoint differences persist across training. Deployment retains General, Spatial Reasoning, and Spatiotemporal; the diagnostic Pointing/Grounding candidate becomes a selectable Tool action. W2P: Where2Place; RefSp: RefSpatial; VAB-Pt: VABench-Point; CV-B: CV-Bench; VSI-B: VSI-Bench.

4.4 Q3: Mechanism Attribution and Sensitivity

Finding 4: learned routing is the dominant component. Figure 5 (left) ablates each decision relative to the 60.2 full system. Removing the router causes the largest drop (-6.1, to 54.1), and random routing is worse still (-7.8, to 52.4), 1.7 points below using no router at all. Owning several adapters is thus harmful when they are mismatched to queries; the portfolio pays off only when a learned utility assigns each query to the right expert. This also rules out the explanation that the Q1 gains merely reflect ensembling several adapters.

Finding 5: knowledge and perception are super-additive. Removing Knowledge (-2.3), perception (-2.8), or both (-4.4) all hurt. From the evidence-free variant (55.8), restoring both channels recovers 4.4 points, 0.7 more than the sum of single-channel recoveries (1.6 + 2.1). The channels fix different failures: Knowledge supplies procedures such as reference-frame normalization, whereas perception supplies observation-bound measurements. A router that can request both for the same expert exploits this synergy, which expert-only routing cannot access.

Finding 6: outcome supervision beats hand-written maps and joint training. Heuristic capability labels score 58.3 (-1.9), so measured expert strengths are more informative than intuitive task-to-expert assignments. Joint end-to-end training (57.1, -3.1) and joint expert updates (56.2, -4.0) are both worse than freezing the experts: updating experts while learning the router shifts the targets the router ranks. Decoupling capability formation from route learning is thus a deliberate design choice, not merely a cost saving.

Figure 5 (right) shows saturation rather than fragility. In a diagnostic sweep, accuracy rises from 55.3 at LoRA rank 4 to 60.2 at rank 16. Three experts improve over two by 2.3 points, while a fourth adds 0.2, matching the three capability regimes in Q2. Scores stay within 59.8–60.5 across answer judges, and halving the route labels costs only 1.1 points (59.1), so the router is data-efficient and its gains are not an artifact of one evaluator.

Selective evidence acquisition. Each channel helps where its semantics match the query: Knowledge supplies procedures, grounding exposes small entities, depth resolves relative distance, and OCR recovers answer-bearing text. The same channels can also hurt, through scene-incompatible retrieval, repeated-instance confusion, or over-weighted low-confidence measurements. This is why *EKT-Router* treats acquisition and acceptance as separate decisions: the verifier can discard such packets after they return and fall back to the evidence-free path, whereas tool-augmented pipelines that always consume tool outputs inherit these errors.

4.5 Q4: General Vision–Language Performance

Finding 7: specialization without sacrificing generality. Embodied and spatial specialists typically trade general competence for geometric skill: in Table 1, the strongest of them, SpatialRGPT-7B (63.6), trails InternVL2.5-8B (74.4) by 10.8 points. *EKT-Router* avoids this trade-off. It averages

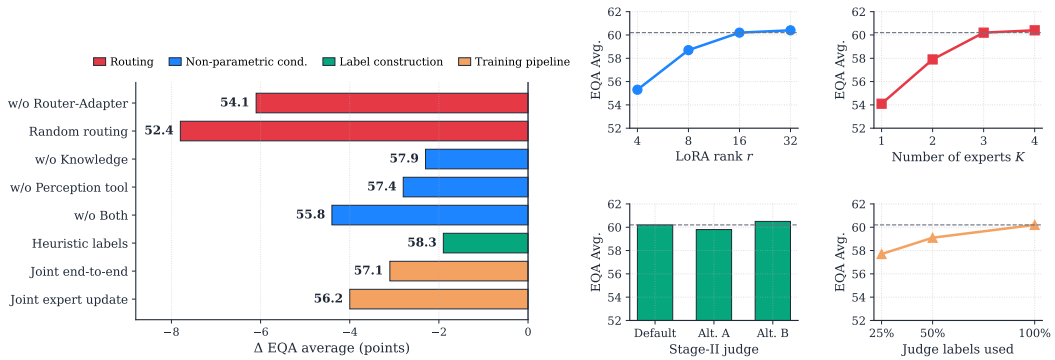


Figure 5: **Ablations and sensitivity under the separate frozen protocol (Q3).** *Left:* changes from the 60.2 full-system score. *Right:* LoRA rank, expert count, answer judge, and route-label coverage. Dashed lines mark the full system.

74.7, on par with the strongest open general VLM, and achieves the best MMStar (63.2) and MMMU (57.8) among all models, surpassing InternVL2.5-8B by +0.4 and +1.8 on these two reasoning-heavy benchmarks. It exceeds SpatialRGPT-7B by 11.1 points on average and on every benchmark, from +2.6 on POPE to +15.8 on ChartQA and DocVQA, and the embodied generalist LEO-7B by 12.7 points. The largest margins fall on documents and charts, where spatial fine-tuning usually degrades most. The reason is structural: specialization lives in separate adapters, and the CPT-trained General expert remains an explicit route for non-embodied queries. A single deployed model thus serves both embodied and general workloads.

4.6 Inference Cost and Auditability

Bounded, query-adaptive execution. Each query runs a fixed pipeline: one descriptor pass, one route proposal, one subset verification, and one answer generation. Unlike iterative agents, whose tool calls and generations grow with the reasoning trace, *EKT-Router* invokes at most B_{ext} providers and generates exactly one answer, regardless of query difficulty. Cost also adapts to demand: queries that need no evidence take the null route and trigger no retrieval or perception call, while the others invoke at most one Knowledge provider and a budget-feasible Tool subset.

Lightweight and auditable control. The answer path activates one frozen 8B backbone with a single expert adapter, and a frozen 2B descriptor supplies the control view. The only trainable controllers, the Utility Router and Packet Verifier, total about 1.245M parameters, less than 0.02% of the answer model, so their forward passes are negligible next to 8B answer generation. Because every decision is explicit, logs record the descriptor, route utilities, selected tuple, packet provenance and confidence, and the answer, so each failure can be traced to demand estimation, acquisition, compatibility, or generation, an attribution that end-to-end agents cannot readily provide.

5 Conclusion

Embodied QA needs a reasoning capability and evidence suited to the same query; choosing them independently can pair an expert with unusable evidence. *EKT-Router* aligns the two in one bounded route: outcome-derived signatures describe what each expert can do, a factorized utility predicts which evidence helps which expert, and packet verification decides what to keep after acquisition, with an exact evidence-free fallback. With one 8B expert active per query, *EKT-Router* achieves the best embodied QA average among the compared systems while matching strong general VLMs. Matched controls show that query-conditioned expert selection pays off on its own, and ablations show that learned routing is the dominant component and that knowledge and perception are complementary. These results suggest that, for embodied models, deciding **which expert should see which evidence** matters as much as strengthening either one. Appendix D outlines future directions.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, language polishing, and LaTeX formatting. They did not conduct the reported experiments or generate the reported measurements. The authors take responsibility for the final text, claims, code, and results, including the verification of primary sources, experimental artifacts, equations, citations, and AI-assisted edits.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a. URL https://arxiv.org/abs/2401.12168.

Shaoxiang Chen, Zequn Jie, and Lin Ma. LLaVA-MoLE: Sparse mixture of LoRA experts for mitigating data conflicts in instruction finetuning MLLMs. *arXiv preprint arXiv:2401.16160*, 2024b.

Xiuwei Chen, Quanlin Chen, Wentao Hu, Zisheng Chen, Kun Xiang, Zehua Ma, Mingyang Zhang, Jianhua Han, Hanhui Li, Hang Xu, and Xiaodan Liang. Beyond the eye: Efficient multimodal reasoning via self-regulated implicit visual tools. *arXiv preprint arXiv:2607.11106*, 2026. URL https://arxiv.org/abs/2607.11106.

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, pp. 135062–135093, 2024. URL https://arxiv.org/abs/2406.01584.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, et al. RynnBrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2018. URL https://arxiv.org/abs/1711.11543.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1932–1945, 2024.

Mingkuan Feng, Jinyang Wu, Hao Gu, Fangrui Lv, Ruihan Jin, Chuyuan Zhang, Zhengqi Wen, and Jianhua Tao. TACO: Tool-augmented credit optimization for agentic tool use. *arXiv preprint arXiv:2606.30251*, 2026. URL https://arxiv.org/abs/2606.30251.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. AssistGPT: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum LoRA experts for continual multimodal instruction tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 19011–19033, 2025.

Gemini Robotics Team, Google DeepMind. Gemini robotics: Bringing AI to the physical world. Technical report, Google DeepMind, 2025.

Amirhosein Ghasemabadi, Ruichen Chen, Bahador Rashidi, and Di Niu. Multi-head latent control: A unified interface for LLM agent decision making. *arXiv preprint arXiv:2607.14277*, 2026. URL https://arxiv.org/abs/2607.14277.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, et al. MiMo-Embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

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://openreview.net/forum?id=nZeVKeeFYf9.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 20413–20451, 2024.

Ziqi Jia, Anmin Wang, Xiaoyang Qu, Xiaowen Yang, and Jianzong Wang. Hierarchical-task-aware multi-modal mixture of incremental LoRA experts for embodied continual learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 28415–28427, 2025.

Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large model empowered embodied ai: A survey on decision-making and embodied learning, 2025. URL https://arxiv.org/abs/2508.10399.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. In *Computer Vision – ECCV 2024*, volume 15105, pp. 126–142, 2024a. URL https://arxiv.org/abs/2311.05437.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, 2024b.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yuankai Luo, Woping Chen, Tong Liang, and Zhenguo Li. CORAL: Scalable multi-task robot learning via LoRA experts. *arXiv preprint arXiv:2603.09298*, 2026.

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*, 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, et al. Embodied arena: A comprehensive, unified, and evolving evaluation platform for embodied AI. *arXiv preprint arXiv:2509.15273*, 2025.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

Zhihong Pan, Kai Zhang, Yuze Zhao, and Yupeng Han. Route to reason: Adaptive routing for LLM and reasoning strategy selection. *arXiv preprint arXiv:2505.19435*, 2025. URL https://arxiv.org/abs/2505.19435.

Hilton Raj and Vishnuram AV. MASER: Modality-adaptive specialist routing for embodied 3d spatial intelligence. *arXiv preprint arXiv:2606.02463*, 2026. URL https://arxiv.org/abs/2606.02463.

Shouwei Ruan, Bin Wang, Zhenyu Wu, Qihui Zhu, Yuxiang Zhang, Hang Su, and Yubin Wang. World2Mind: Cognition toolkit for allocentric spatial reasoning in foundation models. *arXiv preprint arXiv:2603.09774*, 2026. URL https://arxiv.org/abs/2603.09774.

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 *IEEE International Conference on Robotics and Automation*, 2024. URL https://arxiv.org/abs/2311.00899.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. MoME: Mixture of multimodal experts for generalist multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15768–15780, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. URL https://arxiv.org/abs/2303.08128.

TJURL-Lab. UniEQA dataset, 2026. Hugging Face dataset repository. Accessed 2026-07-26.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWvKBCYh4S.

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*, pp. 10632–10643, 2025. URL https://arxiv.org/abs/2412.14171.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024.

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 for robotics. *arXiv preprint arXiv:2406.10721*, 2024. URL https://arxiv.org/abs/2406.10721.

Yi Zhang, Che Liu, Xiancong Ren, Hanchu Ni, Shuai Zhang, Zeyuan Ding, et al. Pelican-VL 1.0: A foundation brain model for embodied intelligence. *arXiv preprint arXiv:2511.00108*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A EXPERT–EVIDENCE ROUTING DETAILS

This section details the information available at each decision stage, the relation between acquisition and retention scores, and the finite search that implements Section 3. The selected expert and original observation remain fixed throughout evidence verification.

A.1 INFORMATION AVAILABLE TO THE TWO DECISIONS

The portfolio contains General, Spatial Reasoning, and Spatiotemporal experts. Each is an independently adapted path on the shared backbone, and its label describes its principal supervision rather than an exclusive task assignment. Evidence-free probe outcomes in Equation 4 characterize overlapping strengths. Every expert can be selected with the null action; General can also be paired with Knowledge or Tools.

Before acquisition, the router observes the question and original visual input through the frozen descriptor, together with expert signatures and candidate action embeddings. Provider outputs are unavailable at this stage: $\widehat{U}_\phi(x, e, a)$ predicts how useful an action will be for an expert. After acquisition, the verifier additionally observes the returned packet content and attributes. It chooses a subset of these packets without changing the expert or acquiring further evidence. Reference answers and answer-judge outputs are used for offline supervision, not either online decision.

A.2 PACKET SEMANTICS AND VALID SUBSETS

A packet separates evidence content from coverage, confidence, ambiguity, provenance, and cost. Coverage associates content with its target regions or frames; provenance associates it with a provider and source observation or passage. Confidence and ambiguity describe the evidence returned by that provider. These attributes are inspected before deterministic verbalization, so rejecting a packet removes its added evidence without altering the original observation.

For the proposed action $a^\star = (k^\star, S^\star)$, each retained action $b = (k', S')$ satisfies $k' \in \{\varnothing_\mathcal{K}, k^\star\}$ and $S' \subseteq S^\star$. Retention operates on the provider packets defined in Equation 3. Let

$$\mathcal{B}(a^\star) = \{b \preceq a^\star : \mathrm{valid}(b, \Omega_b) = 1\}, \qquad a_0 \in \mathcal{B}(a^\star). \tag{15}$$

The deterministic predicate checks schema consistency, target coverage, finite confidence, and provenance. The learned verifier ranks subsets that pass these checks; validity and predicted utility therefore serve distinct roles.

A.3 NULL COST, RESIDUAL TRAINING, AND RETENTION SCORES

The cost convention is $C_K(\varnothing_\mathcal{K}) = C_T(\varnothing) = C(a_0) = 0$. Selecting a_0 before acquisition incurs no external-provider cost. Selecting a non-null $a^\star$ and later retaining a_0 still incurs $C(a^\star)$, because the providers have already executed. This distinction is implemented by the adjustment in Equation 13.

For a training proposal a and a valid non-null subset $b \preceq a$, the packet target in Equation 11 simplifies to

$$\begin{aligned} \rho_i(b \mid a) &= \widetilde{R}_i(y_i, \widehat{y}_i(e, b)) - \lambda C(b) - \widehat{U}_\phi(x_i, e, b) \\ &= u_i(e, b) - \widehat{U}_\phi(x_i, e, b). \end{aligned} \tag{16}$$

The two terms involving $C(a)$ cancel in this target. Thus H_ψ learns the remaining discrepancy between realized and predicted utility for retained non-null evidence, while the inference score separately accounts for the cost of the original proposal.

Writing ℓ_{Huber} for the Huber loss, the verifier objective has the form

$$\begin{aligned} \mathcal{L}_H &= \sum_{(i,e,a,b)\in\mathcal{D}_H} \ell_{\mathrm{Huber}}(H_\psi(h_{x_i}, c_e, r_b, \Omega_b) - \rho_i(b \mid a)), \\ \mathcal{D}_H &= \{(i, e, a, b) : b \preceq a,\ b \neq a_0,\ \mathrm{valid}(b, \Omega_b) = 1\}. \end{aligned} \tag{17}$$

Here $\mathcal{D}_H$ ranges over the collected training executions. The null correction is fixed to $H_\psi(h_x, c_e, r_{a_0}, \varnothing) = 0$ and is excluded from the residual loss. The null reference receives no learned residual correction.

Algorithm 1 Bounded expert–evidence inference

Require: Query $x = (q, o)$, frozen experts, feasible actions $\mathcal{A}$, margin γ
1: Encode x with the frozen query descriptor.
2: $(e^\star, a^\star) \leftarrow \arg\max_{e \in \mathcal{E},\, a \in \mathcal{A}} \widehat{U}_\phi(x, e, a)$
3: Execute each non-null provider in $a^\star$ once to obtain $\Omega_{a^\star}$.
4: Enumerate $\mathcal{B}(a^\star)$, including the always-valid null subset a_0.
5: Compute $v_b \leftarrow \widehat{V}_{\phi,\psi}(b \mid x, e^\star, a^\star)$ for every $b \in \mathcal{B}(a^\star)$.
6: $\widehat{b} \leftarrow \arg\max_{b \in \mathcal{B}(a^\star)} v_b$
7: **if** $v_{\widehat{b}} - v_{a_0} > \gamma$ **then**
8: $\quad b^\star \leftarrow \widehat{b}$
9: **else**
10: $\quad b^\star \leftarrow a_0$
11: **end if**
12: Generate one answer with $\Theta_{e^\star}$ from q, unchanged o, and $\mathrm{verbalize}(\Omega_{b^\star})$.

Substituting Equation 6 into Equation 13 gives

$$\begin{aligned}\widehat{V}_{\phi,\psi}(b \mid x, e^\star, a^\star) = B_\phi(x, e^\star) + G_\phi(x, b) + \Gamma_\phi(x, e^\star, b) \\ + H_\psi(h_x, c_{e^\star}, r_b, \Omega_b) - \lambda C(a^\star).\end{aligned} \tag{18}$$

Every retained subset, including a_0, is charged the same incurred cost. Using the zero anchors for G, Γ, and H at a_0, its predicted margin over null is

$$\begin{aligned}&\widehat{V}_{\phi,\psi}(b \mid x, e^\star, a^\star) - \widehat{V}_{\phi,\psi}(a_0 \mid x, e^\star, a^\star) \\ &\quad = G_\phi(x, b) + \Gamma_\phi(x, e^\star, b) + H_\psi(h_x, c_{e^\star}, r_b, \Omega_b).\end{aligned} \tag{19}$$

The router uses cost when choosing acquisition; the verifier uses returned content to choose what to retain. Shared acquisition cost and the null anchors yield this identity for the predicted score margin.

Null calibration. Write R_b and R_0 for the normalized scores of the same query and expert with retained subset b and with no evidence. If H_ψ fits Equation 16 exactly, the margin in Equation 19 equals $R_b - \lambda C(b) - B_\phi(x, e^\star)$. It matches the realized gain $R_b - \lambda C(b) - R_0$ only when the evidence-free score is calibrated, $B_\phi(x, e^\star) \approx R_0$. The preference loss alone constrains utility differences, not this absolute level; the gain term $\mathcal{L}_{\mathrm{cal}}$ and the held-out margin γ absorb shared offsets but not query-dependent miscalibration of B_ϕ. An equivalent formulation that removes this dependence trains H_ψ on the paired residual $\delta_i(e, b) - \widehat{\Delta}_\phi(x_i, e, b)$, for which an exact fit gives the realized gain directly. Reporting held-out calibration of B_ϕ against evidence-null outcomes, per expert, is therefore part of verifying the retention rule.

A.4 Finite Search and Original-Input Fallback

Let $\mathcal{S}_T = \{S \subseteq \mathcal{T} : C_T(S) \le B_T\}$. The router scores $|\mathcal{E}|(|\mathcal{K}| + 1)|\mathcal{S}_T|$ discrete tuples before acquisition. If the selected route contains $m = \mathbf{1}[k^\star \neq \varnothing_{\mathcal{K}}] + |S^\star|$ non-null providers, it has at most 2^m retained subsets before validity filtering. The deployed menu of three Tool types and at most one Knowledge provider therefore permits at most four provider invocations and sixteen retained subsets per proposal; the configured budget can reduce these bounds. Scoring an additional subset does not invoke a provider or generate a candidate answer.

Algorithm 1 makes the two decisions explicit. The null subset guarantees a defined fallback even when all non-null packets fail validation. If no valid subset clears the commitment margin, the retained packet set is empty and Equation 14 becomes $p_{\Theta_{e^\star}}(\cdot \mid q, o)$, the same selected expert conditioned on the original input. This identity preserves the input path independently of which providers were proposed.

Table 2: **Model, adaptation, and routing configuration.** Spatial and temporal adaptation settings are separate from the diagnostic LoRA-rank sweep in Q3.

Component	Setting
Answer portfolio	Shared Qwen3-VL-8B; one active expert adapter per query
General adaptation	Continued pretraining
Query descriptor	Frozen Qwen3-VL-2B
Spatial/temporal LoRA	Rank 128, scale 256, dropout 0.05
Adaptation optimizer	AdamW; two epochs; learning rate 10^{-4}; weight decay 0.1
Adaptation schedule	Cosine decay with 5% warm-up
Routing split	1,020 fitting cases and 180 held-out cases; media-disjoint
Matched decoding	Shared prompts and generation budgets; temperature zero
Knowledge retrieval	Five passages; evaluation answers and held-out scene descriptions excluded
Perceptual providers	Grounding DINO, Depth Anything V2, and OCR

B Training and Evaluation Details

B.1 Adaptation and Routing Stages

Expert adaptation precedes router fitting. General uses continued pretraining; Spatial Reasoning and Spatiotemporal use spatial and video supervision, respectively. The completed experts and the query descriptor remain frozen during router fitting. Evidence-free diagnostic outcomes define the capability profiles, while paired route executions provide answer-derived utility and packet-correction targets. Table 2 collects the configuration used in the main text.

For a fixed expert and input, paired evidence-on and evidence-null executions use the same prompts and decoding. Equation 9 combines their native normalized scores with action costs. Route pairs separated by more than ϵ supervise preference ranking, and the lower-weight Huber term fits realized acquisition gains. The packet objective in Equation 17 additionally uses the returned evidence attributes to correct non-null route predictions. The held-out routing cases set operating costs and margins; benchmark test scores do not select these operating points.

B.2 Comparison Units

Q1 reports the equal-weight mean of the six primary embodied QA benchmark scores, using each benchmark's native higher-is-better metric. Q2 adds Cosmos, RefSpatial, and VABench-Point to those six tasks and disables both external evidence channels for every compared expert. Q4 uses the separate ten-task general vision–language suite in Table 1. These suite means aggregate benchmark scores, rather than pooling all questions into one accuracy.

The six-task full-system mean in Q1 is 61.40. In the nine-task evidence-null control, question-conditioned selection reaches 60.24 versus 58.49 for the best fixed expert, yielding the reported 1.75-point improvement. Q3 reports component changes relative to its own 60.2 full-system reference under the separate controlled protocol. The Q3 effects are within-protocol comparisons; subtracting its absolute scores from Q1 or Q2 would combine different evaluation settings.

Decision ablations hold expert checkpoints, instances, decoding, and evaluators fixed; training and rank controls change the named adaptation configuration. The Q2 media-grouped bootstrap interval, $[1.25, 2.26]$, describes uncertainty in the selection gain under that matched evaluation. It does not measure variation across independently retrained portfolios. Cross-system comparisons combine reproduced, official API, and reported results under their respective settings.

B.3 Mechanism-Level Interpretation of Q3

The Q3 controls are interpreted as interventions on distinct decision interfaces. The routing intervention removes query-dependent demand matching while preserving the expert bank; the full learned route's advantage over the random and no-router conditions therefore tests whether the assignment rule accounts for the gain. The evidence interventions preserve the selected expert and vary only the available information: Knowledge supplies procedures or semantic priors, whereas perception

Table 3: **Evidence channels and representative interpretation issues.** The channels supply different information and require different consistency checks.

Channel	Information supplied	Representative interpretation issue
Knowledge	Reasoning procedures or retrieved semantic context	A passage may assume a reference frame or scene condition inconsistent with the observation.
Grounding	Target localization in an image or frame	Repeated instances or a small target may create ambiguous associations.
Relative depth	Relative-distance structure within the observation	Relative depth alone does not specify a calibrated metric distance.
OCR	Visible text associated with a region	A transcription or region association may be incorrect.

supplies measurements tied to the current observation. The observed 4.4-point recovery with both channels exceeds the 3.7-point sum of the single-channel recoveries by 0.7 points, consistent with these information sources covering different failure modes.

The supervision controls provide a second attribution check. A heuristic route does not receive answer-derived credit, while joint expert updates alter the answer functions at the same time as the route model. The frozen-expert, outcome-derived setting thus keeps the answer path fixed while learning the route utility, making its comparison with these controls more interpretable. This is an attribution rationale rather than a claim that the controls isolate all optimization effects.

The reported component removals do not independently identify the interaction term Γ or the Packet Verifier. A fully isolated audit would reuse the same expert checkpoints, route candidates, provider outputs, and decoding while comparing additive utility without Γ, utility with Γ but without packet correction, and a sequential expert-then-evidence selector. Such a comparison would distinguish expert–evidence compatibility from generic tool use and post-acquisition filtering; until then, the present evidence supports the interface-level interpretation above and not a separate causal estimate for either unablated component.

C Evidence Interfaces and Interpretation

C.1 Channel Semantics

Knowledge and perception contribute different information. Reference-frame, spatial-decomposition, and temporal-ordering templates supply reasoning procedures, while retrieved passages supply semantic context. Perceptual providers instead extract information tied to the observed image or frames. The selected expert determines how these complementary inputs are used to answer the query. Table 3 gives representative interpretation issues that can arise in each channel.

The RGB and relative-depth panels in Figures 1 and 2 illustrate how an observation can produce a distinct perceptual representation. Its packet retains source and target associations so the verifier can examine that representation before it is added to the answer context. The original visual input remains available whether the packet is retained or rejected.

C.2 Interpreting Acquisition and Retention

An evidence comparison is conditioned on the same expert, question, observation, and decoding. Its evidence-on and null-evidence conditions differ in the added packet content, whereas comparing different experts changes the reasoning path as well. The distinction matters for attribution: evidence utility is assessed by the former comparison, and expert complementarity by the latter under matched evidence conditions.

If a retained subset fails the validity predicate, it is excluded from the candidate set. If a valid packet is irrelevant or conflicts with other evidence, its usefulness must instead be assessed through the predicted expert-conditioned margin. A retained packet can still be misunderstood by the answer

expert, and an empty retained set can still leave an underspecified question. These cases distinguish packet validity, evidence usefulness, and answer reasoning, which are separate stages of the route.

D Limitations and Future Work

EKT-Router treats *which expert should see which evidence* as a first-class decision. We see this as a first step toward embodied agents that manage their own capabilities and information, and outline four directions that it opens.

From a fixed portfolio to an open capability market. The current portfolio contains three experts and three tools. Because experts are represented by outcome-derived signatures rather than names, a new skill can in principle be added by profiling it on diagnostic probes, without redesigning the router. Scaling this idea to tens of community-contributed adapters and tools, spanning manipulation, navigation, audio, and tactile sensing, would turn the router into a marketplace that composes capabilities on demand, and would test whether expert–evidence alignment continues to pay off as the space of routes grows combinatorially.

From single-round routing to closed-loop active perception. *EKT-Router* decides once which evidence to acquire. Embodied agents, however, can move, look again, and query new viewpoints. Extending the route utility to sequential decisions, where the verifier's judgement of returned packets informs the next acquisition or a change of viewpoint, would connect expert–evidence alignment with active perception and information-gathering policies, while retaining the explicit budget and fallback that make the current design auditable.

From question answering to planning and action. We use answer outcomes as the supervision signal because EQA provides them at scale. The same formulation applies wherever a downstream outcome can score a route: task success in vision–language–action policies, plan feasibility in long-horizon manipulation, or safety constraints in navigation. Routing experts and evidence for action generation, rather than answers, is a natural path from embodied understanding to embodied control.

Toward self-improving routing. Every executed route produces a labeled outcome. Deployed at scale, these logs form a continual training signal from which signatures, utilities, and verification can be refined online, and from which gaps in the portfolio can be detected and filled with new experts. Understanding how such a system co-evolves its capabilities and its routing, and how to keep that process stable and interpretable, is an exciting direction for future work.