KITE: SPECIALIZING FROZEN MULTIMODAL MODELS WITH A SHARED SIDE-INPUT EXPERT

Anonymous authors
Paper under double-blind review

ABSTRACT

Specializing a multimodal model with external evidence raises a twofold design challenge: how to integrate that evidence without consuming the host's input sequence, while preserving its behavior when specialization is inactive. We introduce Kite, a shared side-input expert that reads external evidence and host marker states at sparse depths and writes additive updates at those markers. Marker-free requests bypass the branch, preserving host outputs bit-exactly under the same execution configuration; a training-free, prompt-level router controls invocation. We instantiate Kite for Earth observation using frozen DINOv3 features and a frozen 35B host, training a 1.88B-parameter side expert. On a pre-locked new-site recognition test, Kite exceeds a host LoRA matched across the full training chain by +9.8 percentage points. On a separate pre-locked test, mismatched imagery reduces accuracy from 0.429 to 0.106, supporting evidence-sensitive recognition. Benchmark adaptation confined to the expert raises the official VRSBench-VQA ten-category mean from 0.627 to 0.772 and reaches 0.913 on RSVQA-LR. These gains incur additional side computation. We report its cost alongside controlled ablations and a pre-registered negative result on depth-structured evidence use.

1 INTRODUCTION

When one multimodal host serves both general and specialist requests, adding domain expertise should leave established behavior intact whenever the specialist is inactive. An explicit invocation boundary makes the extension easier to audit and disable without replacing the host. This motivates a modular specialist with a dedicated evidence pathway.

We seek three properties together: evidence outside the host token sequence, interaction at multiple depths, and explicit invocation with exact preservation when inactive. Visual prefixes can encode external evidence but consume sequence positions. LoRA (Hu et al., 2022), including conditional variants (Greenewald et al., 2025), adapts host transformations without itself providing a separate evidence interface. Side networks such as Ladder Side-Tuning (Sung et al., 2022) connect across layers; our design instead reuses one evidence-reading expert at sparse depths.

We introduce Kite, a method for specializing frozen multimodal models with a *shared side-input expert*. External evidence is represented as a variable-length feature set outside the host's token sequence. At sparse depths, lightweight interfaces read the host's marker states, query the evidence, and pass both to one shared expert trunk. The trunk's output is written back as an additive update at designated marker positions. The host's native computation remains available, and marker-free requests skip the branch entirely, preserving host behavior bit-exactly under the same execution configuration (Thm. 1).

Earth observation (EO) provides a testbed with fine-grained visual evidence in native-resolution, often submeter imagery, represented by an encoder distinct from the host's vision tower (Fig. 1). A frozen 35B host and DINOv3 encoder (Siméoni et al., 2025) supply host states and evidence to a shared 2B trunk, with 1.88B trainable parameters (5.4% of the host). Training integrates expert and native vision before benchmark adaptation (§3.4); a frozen-host router controls invocation (§4.4).

This instantiation strengthens fine-grained expertise with unchanged host weights. On a pre-locked new-site test, Kite exceeds a host LoRA matched across the full training chain by +9.8 percentage points; on the separate 198-question evidence-control subset, accuracy is 0.429 with correct imagery

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.

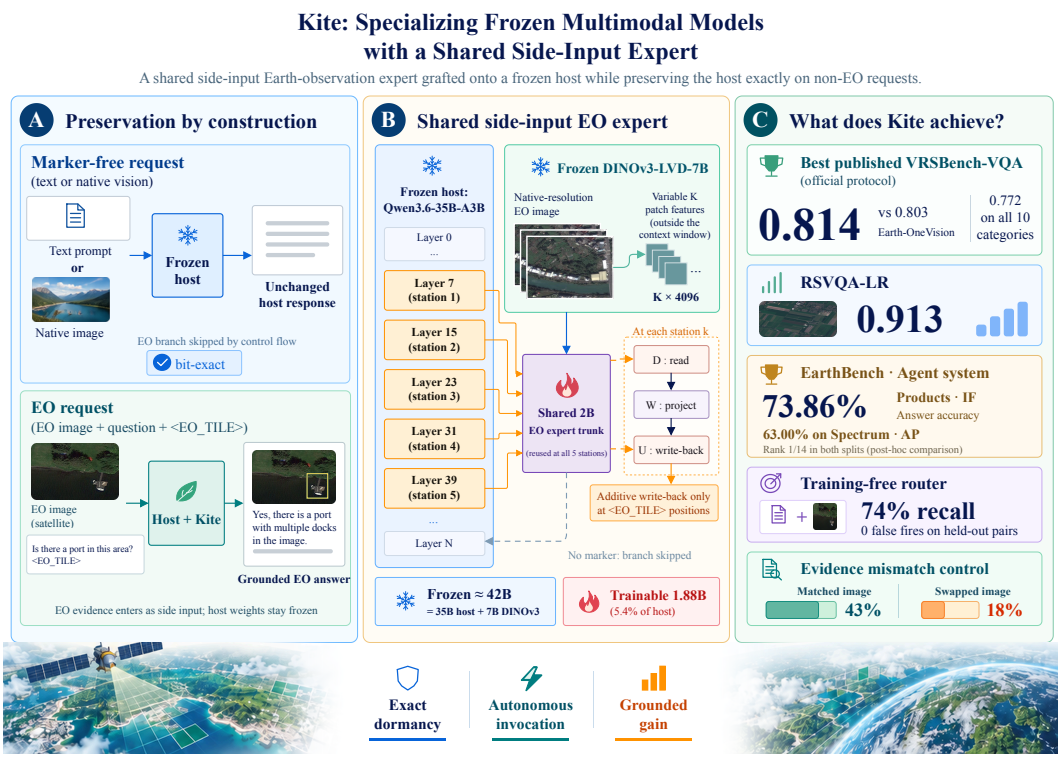


Figure 1: **Kite instantiated for Earth observation.** **(A)** Marker-free requests preserve the frozen host exactly; EO requests invoke the side-input expert. The port example and box are illustrative, not a measured model response or predicted localization. **(B)** Frozen DINOv3 patches feed one shared expert at five host depths; write-back is restricted to <EO_TILE> positions. **(C)** Selected results. VRSBench's 0.814 comparison uses seven reported categories (0.772 on all ten). EarthBench reports agent-system answer accuracy on Products–IF and Spectrum–AP; ranks are post-hoc comparisons with 13 supplied same-mode systems, not official submissions. The 43%/18% mismatch control uses the stage-1 corner-development subset ($n = 110$, native vision off; §4.3).

and 0.106 with mismatched imagery (Table 2). Broader benchmark adaptation raises the official VRSBench ten-category mean from 0.627 to 0.772 and reaches 0.913 on RSVQA-LR (§4.2). The expert adds parameters and side computation; we report these costs alongside the gains.

Our contributions are:

- A shared side-input architecture with bit-exact preservation when inactive (Thm. 1) and a training-free invocation router.
- An EO instantiation and three-stage recipe: establishing the expert before native-vision fusion avoids the channel collapse observed under co-training in our setup (§3.4; Fig. 6).
- Controlled LoRA comparisons, evidence-mismatch interventions and station ablations (§5), together with a pre-registered negative result on progressive evidence consumption (App. A).

2 Related Work

Frozen hosts, zero-initialized branches, conditional adapters. Keeping the language model frozen while attaching vision dates to Frozen (Tsimpoukelli et al., 2021); Flamingo (Alayrac et al., 2022), LLaMA-Adapter (Zhang et al., 2023) and BLIP-2 (Li et al., 2023) use zero-initialized or interposed modules. Zero initialization controls the initial effect; Kite's control-flow bypass preserves marker-free execution throughout training (Thm. 1). LoRA (Hu et al., 2022) and bottleneck adapters (Houlsby et al., 2019) provide parameter-efficient adaptation. Unmerged LoRA

can be disabled per request; Activated LoRA (Greenewald et al., 2025) acts after an invocation marker and reuses the prefix KV cache. Conditional activation and prefix reuse are therefore not unique to Kite (Cor. 1). Its distinction is one shared expert reading external evidence and host states at sparse depths. Ladder Side-Tuning (Sung et al., 2022) targets training-memory savings with layerwise side computation and connections at every host layer; LoRAMoE (Dou et al., 2024) and MoE-Adapters (Yu et al., 2024) add experts to protect prior knowledge; in remote sensing, RSAdapter (Wang & Ghamisi, 2023) inserts adapters into a frozen VQA model and RSUni-VLM (Liu & Lian, 2024) routes granularity-specific experts inside the host.

EO vision–language models. GeoChat (Kuckreja et al., 2024), RSGPT (Hu et al., 2023), EarthGPT (Zhang et al., 2024), LHRS-Bot (Muhtar et al., 2024) and SkySenseGPT (Luo et al., 2024) adapt language models for EO; SkyMoE (Liu et al., 2026) and RS-MoE (Lin et al., 2024) build specialist MoEs with trainable routing. Kite adds EO capability with an explicit inactive-path preservation guarantee and evaluates against the host's own vision path (§4.2). EO embedding models (AlphaEarth (Brown et al., 2025), TESSERA (Feng et al., 2025), Prithvi (Jakubik et al., 2023), SatMAE (Cong et al., 2022), Scale-MAE (Reed et al., 2023)) occupy a different layer of the stack and could supply alternative evidence features.

3 METHOD

3.1 PROBLEM SETTING: ADAPTING A FROZEN HOST UNDER A PRESERVATION CONSTRAINT

We are given a frozen, causally masked multimodal host f_{host} and a specialist task with external evidence z. An evidence encoder ϕ supplies features $E = \phi(z) \in \mathbb{R}^{K \times d_v}$, where K may vary across inputs. Kite learns a shared expert trunk $\mathcal{T}$ and read/write interfaces that connect this evidence to the host's hidden states at selected depths. The host's parameters and native input pathways remain frozen. We impose a *preservation constraint*: for any input x without the expert's designated invocation markers, including text and native multimodal inputs,

$$f_{\text{Kite}}(x) \;\equiv\; f_{\text{host}}(x) \qquad \text{bit-exactly, not approximately.} \tag{1}$$

Kite satisfies Eq. 1 by control flow (Thm. 1) and we verify it empirically to $\Delta = 0$ (§4).

3.2 KITE: AN ADDITIVE SHARED EXPERT AT SPARSE DEPTHS

Kite augments a selected set of host depths $\mathcal{S}$, called *stations*, with one shared expert trunk. Each station has its own read, evidence-projection and write-back interfaces; the trunk's parameters are reused across stations.

Evidence as side input. The feature set E is routed directly to the stations, outside the host's token sequence. The prompt carries M designated marker tokens after the question, allowing the expert to read question-conditioned host states (Prop. 1). Thus the side branch adds M sequence positions independently of K; any inputs consumed by the host's native path retain their usual sequence cost.

Station computation. At a station $k \in \mathcal{S}$, let $h \in \mathbb{R}^{B \times T \times d}$ be the hidden states entering the selected native computation block F_k, and $m \in \{0,1\}^{B \times T}$ the marker mask. Suppressing the station index on the interfaces, the augmented block computes

$$y \;=\; \underbrace{F_k(h)}_{\text{frozen native block}} \;+\; m \odot g(h) \cdot \mathrm{U}\big(\mathcal{T}\big(\mathrm{D}(h),\, \mathrm{W}(E)\big)\big), \tag{2}$$

where $\mathrm{D} : d \to d_e$ reads the residual stream (d_e is the trunk's hidden width), $\mathrm{W} : d_v \to d_e$ projects the evidence features, $\mathcal{T}$ is the shared expert trunk, and $\mathrm{U} : d_e \to d$ writes back. U is zero-initialized (Flamingo-style), so the branch initially contributes zero. The native block retains its weights and computation rule. When the expert is active, marker updates can affect subsequent text states through causal attention; bit-exact equality with the host applies to marker-free requests and positions before the first marker (§3.3).

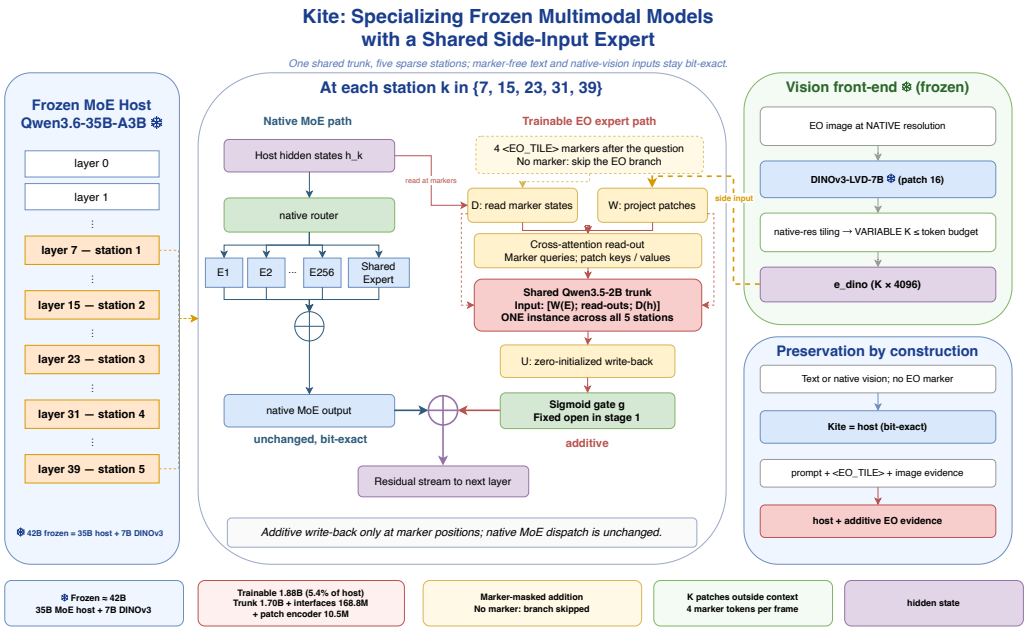


Figure 2: **Kite instantiated for EO.** Frozen DINOv3 patches and host marker states feed one shared trunk at five MoE depths (§3.2). Updates are written only at <EO_TILE> positions; marker-free inputs skip the branch (Thm. 1). The expert trains 1.88B parameters (5.4% of the frozen host).

Cross-attention read-out. At each station, marker states plus learned slot embeddings query $\mathrm{W}(E)$ to produce read-outs $r_{1..M}$. The trunk consumes $[\mathrm{W}(E)\,;\, r_{1..M}\,;\, \mathrm{D}(h)_{1..M}]$, or $K + 2M$ tokens; attention structure also determines computation cost (App. E). Read-out normalization and scaling remain inside the marker mask (App. D). Replacing the patch path with read-outs alone collapses recognition (App. E).

Sharing and conditional invocation. The same trunk $\mathcal{T}$ is applied at every station to that station's projected evidence and current marker states. Its parameter count is independent of the number of stations; interfaces and computation are incurred at each station. If no marker is present, the implementation skips the branch by control flow. For automatic invocation, a prompt-level router can decide whether to insert markers before the main forward pass; our deployment uses the frozen host for this decision (§4.4).

EO instantiation. Our frozen Qwen3.6-35B-A3B host (Qwen Team, 2025) uses MoE stations at layers $\{7, 15, 23, 31, 39\}$. Frozen DINOv3-7B features ($d_v = 4096$, patch size 16) retain native resolution within a 1024-patch budget, with larger images downscaled. Each frame uses $M = 4$ <EO_TILE> markers and a shared Qwen3.5-2B trunk. The 1.88B trainable parameters comprise the trunk, station interfaces and patch encoder; implementation details are in App. D.

3.3 Properties of the Construction

All statements below are architectural and hold for any weights.

Theorem 1 (Bit-exact behavioral superset)**.** *For any marker-free input x (no invocation marker, or no evidence E supplied), $f_{\mathrm{Kite}}(x) = f_{\mathrm{host}}(x)$* bitwise*, in floating-point arithmetic.*

The proof (App. D) is control-flow identity: for the same backend, precision and kernel configuration the two models execute the same operations in the same order. The marker's vocabulary row preserves original-vocabulary logits (App. C); mathematical equality from a zero-valued branch alone does not guarantee bitwise equality.

Corollary 1 (Prefix locality)**.** *For marker-bearing inputs, all positions strictly before the first marker emit bit-identical hidden states and logits to the host's. Consequently any KV-cache prefix computed by the host remains valid for Kite, and vice versa.*

Proposition 1 (Gating measurability)**.** *Under causal masking, any gate or branch computed at position i is a function of tokens at positions $\leq i$ only. Hence a marker placed* before *the question makes question-conditioned gating impossible — an information-theoretic property of the placement, not a training difficulty.*

Proposition 1 determines marker placement (App. A).

3.4 TRAINING

Training has three stages under one objective (App. D); each stage initializes from the previous one and the host stays frozen throughout.

Stage 1: EO expert. Trains the trunk, station interfaces and patch encoder (the marker's embedding row stays frozen, App. C) on the generated corpus of §4.1; the gate is fixed open and the host's native vision path is off, so the expert alone must carry the EO evidence. Frozen-corpus recipe comparisons use 500 steps, where development metrics plateau; training exposure and optimization details are in Apps. B.1 and D.

Stage 2: perception scheduling. The native vision path is switched on (image tokens in the sequence, 409,600 pixels), the gate becomes a trainable sigmoid supervised by same-image pairs (§4.4), and the expert continues training on the frozen corpus for 2,000 steps at half the stage-1 learning rates. We call the result the joint model; the gate is kept as a diagnostic, the deployed invocation mechanism is the router of §4.4.

Stage 3: benchmark fine-tuning. For a public benchmark, the joint model is fine-tuned on that benchmark's training split with the gate frozen (8,000 steps for VRSBench, 12,000 for RSVQA-LR). Rows labelled *in-domain* in the tables skip stage 2 and fine-tune the stage-1 expert directly.

4 EXPERIMENTS

We report task performance and preservation under the settings below; §5 contains the ablation study. Additional system evaluations, costs and per-type scores are in Apps. F, E and G.

4.1 EXPERIMENTAL SETUP

Training data and checkpoints. With the host frozen, stage 1 trains the EO path on 2.27M then 133,333 rows from fMoW-RGB (Christie et al., 2018) and DOTA (Xia et al., 2018); stage 2 enables native vision and stage 3 adapts the expert on benchmark training splits. Target-benchmark zero-shot excludes that benchmark's training data (§B).

Evaluation data and tasks. Public tests cover 37,409 VRSBench questions, 10,004 RSVQA-LR questions, and ordinary/Philadelphia RSVQA-HR (§G.2); VRSBench diagnostics use 2,400 rows. Grounded development uses 886 image-disjoint fMoW questions, including 110 20-class object-identification questions (§B).

Data separation. Development sites overlap frozen-corpus training by 26.4%; recipe selection uses this set. Pre-locked new-site tests retain 377/932 questions (198/458 corner questions) under distinct answer-independent filters (§B).

Baselines and controlled comparisons. Baselines are the text-only/native-vision host, mismatched EO evidence, a trunk-free connector, a one-station expert and rank-64 host LoRA. LoRA controls match benchmark adaptation or the full training chain; initialization, data and step budgets are specified in §B.

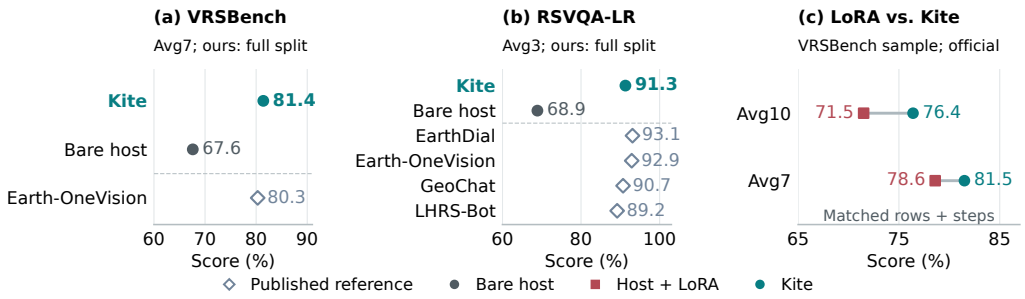


Figure 3: Benchmark landscape and matched adaptation (%). (a,b) Full-split host/Kite results and published references, not matched reruns; category means and horizontal ranges differ by panel. (c) Official-protocol LoRA/Kite scores on the same 2,400-row sample, with matched benchmark rows and steps; raw answers can benefit from substring matching. Full training metadata and diagnostic protocols: Table 8; extracted-answer comparison: §5.1.

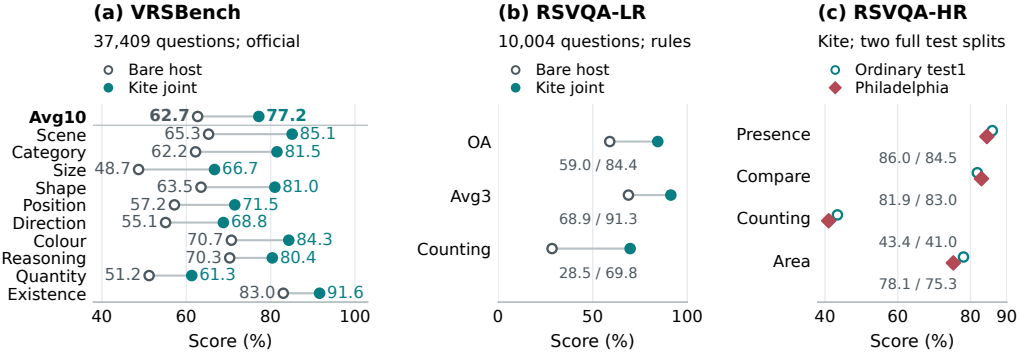


Figure 4: Full-split benchmark results (%). (a) VRSBench official scores, ordered by Kite's gain; Avg10 is the ten-type mean. Substring matching can favour long answers (extracted readings: App. G). (b) RSVQA-LR: overall accuracy (OA), official three-type mean (Avg3), and counting; both models answer the same 10,004 questions. (c) Benchmark-adapted Kite on ordinary RSVQA-HR test1 (222,684 questions) and Philadelphia (105,647). Numeric pairs follow legend order. Panel (c) compares test distributions, not models, with an expanded 38–90% horizontal range. Horizontal ranges differ across panels, which use their respective benchmark protocols.

Scoring and statistics. Grounded scores use exact/balanced accuracy or numeric error; 95% paired bootstrap intervals resample development images (10,000 draws) or probe sites (§B). Public results use official VRSBench (Li et al., 2024) and RSVQA-LR (Lobry et al., 2020) scoring; diagnostic judge/extractor protocols are distinguished in Table 8 and §B.

4.2 Public-benchmark results

VRSBench and RSVQA-LR. On full-split VRSBench, the joint model reaches 0.772 versus 0.627 for the bare host (ten-category means; paired per-question $\Delta = +0.135$, 95% CI [+0.131, +0.139]). Its seven-category mean is 0.814 versus Earth-OneVision's published 0.803 (Cai et al., 2026), a 1.1-percentage-point gain. Substring matching precedes the judge and can favour longer answers (extracted host/Kite readings: App. G). Full-split RSVQA-LR reaches 0.913 over three types, with counting rising from 0.285 to 0.698 through the native pixel path. Figures 3 and 4 show reference-model and per-type comparisons; Table 8 retains all training metadata and protocols. Additional zero-shot results and selected qualitative cases are in App. G.2.

Run-to-run reproducibility. A full-recipe rerun, with stage 1 restarted from scratch at 20,000 steps, reproduces the VRSBench sample result under the official protocol (0.763 vs. 0.764, same 2,400 rows, raw answers) and the RSVQA-LR full-split result (OA 0.845 vs. 0.844; three-type average 0.894 vs. 0.913, a gap of six of the 100 rural/urban questions).

High-resolution VQA with a frozen host. After 2,000 steps on the RSVQA-HR training split, the side expert reaches 83.94% presence/comparison mean on ordinary test1 and 83.75% on the geographically shifted Philadelphia split. Across all four question types, OA is 72.35% and 71.30%, respectively, on the full 222,684 and 105,647 active questions, with no inference errors. These are benchmark-adapted results; the host remains frozen. Full category scores and published-reference scope appear in App. G.2.

Agent-system transfer (EarthBench). On EarthBench, Kite in direct mode reaches 63.00%/68.18% answer accuracy on Spectrum/Products in answer-provided (AP) mode, and 63.00%/73.86% in instruction-following (IF) mode. Against bare Qwen3.6-35B-A3B on the same server, these are gains of +17.00/+17.05 and +22.00/+22.73 percentage points, respectively. Inserting each run among 13 supplied same-mode systems places AP Spectrum and IF Products first of 14, and AP Products and IF Spectrum second; these are post-hoc comparisons, not official submissions. RGB answer accuracy also improves, but its strict tool-trajectory metrics decline (App. F).

4.3 Fine-grained recognition

Recognition and evidence dependence. Table 1 reports the stage-1 expert on the development set. Corner accuracy is 0.43 exact and 0.28 balanced, against 0.21 and 0.05 for the majority predictor. Mismatching EO evidence reduces exact accuracy to 0.18, a +0.25 drop relative to the +0.43 gain over the text-only host. The residual 0.18 may reflect task and format priors; wrong imagery can also mislead, so this intervention is not a strict decomposition of the gain. Giving the native host the 20 class names raises its score to 0.28 (text-only 0.07), but the expert retains a +0.15 margin (95% CI $[0.04, 0.25]$). Its mismatched score falls below this vocabulary-given baseline, supporting an evidence-dependent recognition gain.

Table 1: Fine-grained recognition on the pure development subset. Paired differences and 95% bootstrap intervals compare the expert with the text-only host, native vision and mismatched EO evidence. The second row gives both host baselines the 20 class names; n.s. means not significant.

Field	Kite	Δ vs. host-alone	Δ vs. native vision	Δ vs. mismatch
Corner object (20-way, exact)	0.43	+0.43 [0.34, 0.52]	+0.43 [0.34, 0.52]	+0.25 [0.16, 0.33]
host given the 20 names (multiple choice)	0.43	+0.36 [0.25, 0.46]	+0.15 [0.04, 0.25]	—
Season (temporal)	0.42	—	n.s.	+0.17 [0.01, 0.33]
Geodesic distance (median error)	—	$-1{,}349$ m	-116 m [-140, -77]	—

Recognition on additional sites. Table 2 brings together the independent new-site tests: the 198-question subset measures station and evidence effects; the 458-question subset compares Kite with host LoRA. On the former, mismatching imagery reduces accuracy from 0.429 to 0.106.

4.4 Exact preservation and autonomous routing

Exact dormancy. Comparing marker-free text and native-vision logits with an independently loaded pristine host in bf16 on the same backend, the maximum absolute logit difference is 0 for every trained model tested, as Theorem 1 predicts. On 20 held-out native-vision inputs it is 0 with identical greedy decodes.

Autonomous invocation. For tile-bearing requests, a one-word EO-relevance query to the frozen host determines marker insertion, costing one extra host prefill; requests without a tile bypass routing and preserve the host unconditionally. Same-image relevant/irrelevant pairs, image-disjoint from stage-2 training, evaluate the router and the independent sigmoid-gate diagnostic. On 1,094 held-out same-image pairs per class, the prompt-level router never fires on an EO-irrelevant question (100%) and fires on 74% of EO-relevant ones (balanced accuracy 0.869). Every miss is a season

Table 2: Independent new-site recognition, with both manifests locked before evaluation. A: 198 corner questions from the strict 377-question, 357-site subset. B: 458 corner questions from the LoRA-comparison subset (932 questions, 848 sites). The subsets use different exclusion lists (App. B); scores are compared within each panel. Differences are Kite minus control, with paired 95% intervals.

Model / evidence	Exact	Difference [95% CI]
A: station and evidence controls (accuracy, 0–1)		
Kite / correct image	0.429	—
Single-station continuation	0.071	+0.359 [0.288, 0.429]
Kite / mismatched image	0.106	+0.323 [0.242, 0.399]
B: full-chain LoRA control (accuracy in %, difference in pp)		
Kite (163/458 correct)	35.6%	—
Host + LoRA (118/458 correct)	25.8%	+9.8 [+6.6, +13.1]

question whose unredacted header already carries the date (51% of those fire; all other templates 100%). It fires on 97.6% of VRSBench sample questions (all but three misses are image-metadata questions), 100% of 500 RSVQA-LR questions, and 0% of 300 general prompts (95% upper bound 1.2%; 1.3% with an EO header). A false fire forfeits exact preservation; a miss returns the host's own answer.

5 Ablation Study

5.1 LoRA adaptation

Public-benchmark adaptation. With the same benchmark rows and steps, native-path LoRA scores 0.786 versus 0.815 for the joint model under the official sample protocol (seven-category means). Under the frozen-host judge with the shared extractor (H/S in Table 8), LoRA scores 0.739 versus 0.705 for the joint model over ten types; the paired joint$-$LoRA difference is -0.035 (95% CI [-0.057, -0.015]), significantly favoring LoRA. LoRA is stronger on counting, size and position, and weaker on colour (App. G). On 512-pixel chips, routed LoRA provides comparable gains and dormancy; Table 3 examines the sub-metre grounded setting.

Fine-grained recognition. On the independent LoRA-comparison subset, Kite gains +9.8 percentage points over host LoRA (Table 2). This specialist gain uses 1.88B trainable parameters versus 84.7M for LoRA; the estimated visual-path cost ranges from 6.8 to 27 TFLOP as the patch budget increases, versus 3.8 for native vision with LoRA (App. E).

5.2 Architecture and training ablations

Frozen-corpus training from scratch. In the exposure-matched upper block of Table 3, five stations reach 0.08 corner accuracy, below LoRA's 0.22 (paired $\Delta = -0.14$, 95% CI [-0.23, -0.05]) and level with one station (0.09; paired 95% CI [-0.06, 0.06] over all 150 corner rows). The connector and expert variants remain below the majority-class corner rate at this exposure; the five-station trunk shows no detected advantage.

Large-corpus initialization and station continuation. The five-station continuation reaches 0.43, improving over five stations from scratch by +0.35 (CI [0.25, 0.45]) with a paired 4,367 km smaller geocell error (CI [2,691, 6,108]). From the same five-station initialization, the station-31-only continuation reaches 0.10 (season 0.28 vs. 0.42). The independent test also favors five stations (Table 2A). The from-scratch five- and one-station arms remain level there (0.066 vs. 0.056; difference +0.010, CI [-0.040, 0.061]). Thus the continuation comparison favors five active stations under matched initialization and training exposure. All grafts reduce metric-distance error by about 1.35 km, and LoRA also reduces geocell error. The large-corpus stage is associated specifically with the five-station recipe's identity gain. The earlier 1/5/10-station loss study and these task-level comparisons are summarized separately in App. E.

Table 3: Architecture and training ablations. Upper block: frozen-corpus training from scratch. Lower block: matched continuations from the same five-station large-corpus checkpoint. Corner exact accuracy uses 110 paired development questions (majority baseline 0.21). Geocell is median great-circle error; distance is the change in median error from the host's 1,404 m. Textual paired distance comparisons use medians of per-image differences. Baselines are defined in §4.1.

Graft	Large corpus	Trainable	Corner (exact)	Δ corner vs. host	Geocell (km)	Δ distance
Host alone (EO inert)	—	0	0.00	—	—	—
LoRA on the host's native path	no	84.7M	0.22	+0.22 [0.15, 0.30]	1,711	−1,364 m
One-station connector, no trunk	no	44.3M	0.00	+0.00 [0.00, 0.00]	3,479	−1,360 m
Single-station expert	no	1.74B	0.09	+0.09 [0.05, 0.15]	2,783	−1,364 m
Kite (shared, 5 stations)	no	1.88B	0.08	+0.08 [0.04, 0.14]	4,805	−1,364 m
Single-station expert	yes	1.74B	0.10	+0.10 [0.05, 0.15]	5,160	−1,364 m
Kite (shared, 5 stations)	yes	1.88B	0.43	+0.43 [0.34, 0.52]	1,112	−1,349 m

5.3 Evidence-mismatch analysis on benchmark images

On the 2,000 diagnostic rows of VRSBench's ten official types, the joint model scores 0.716 versus 0.680 for native vision (paired $\Delta = +0.036$, 95% CI [+0.015, +0.057]); extracted answers score 0.705 versus 0.679 (CI [+0.006, +0.044]). Replacing evidence with the preceding image gives 0.714; using a random image from another scene and source dataset (DOTA↔DIOR) gives 0.715. These swaps barely change mean accuracy, but the preceding-image swap changes correctness on 4.9% of rows (51 lost, 47 gained), versus 30.9% for the EO-only expert (467 lost, 152 gained). The EO-only in-domain expert (0.663 on this sample) falls to 0.504 with another scene's evidence (paired $\Delta = +0.159$, 95% CI [+0.135, +0.180]; category, scene and existence all drop). The two arms differ in task, training and native-path state, so their different swap responses do not isolate a mechanism for the joint model's gain.

5.4 Learned-gate ablation

The independent sigmoid gate learns an open−close separation of +0.216 (median +0.244; Fig. 5), but training or freezing it during fine-tuning makes no measurable difference (0.721 vs. 0.716, CI [-0.013, +0.023]). Forcing it open on VRSBench likewise changes accuracy little (0.518 vs. 0.516). It is therefore retained as a diagnostic; deployed invocation uses the prompt-level router.

6 Discussion and Conclusion

Kite adds specialist capabilities through a shared side-input expert, improving public benchmarks and independent new-site recognition with unchanged host weights and exact marker-free preservation. These gains incur additional computation. Benchmark gains and evidence dependence remain distinct: the joint model's VRSBench mean changes little under mismatch.

Limitations. Evaluation covers one host and optical EO imagery; transfer to other hosts, SAR and multispectral imagery remains to be evaluated. EO-only counting, fine-grained position readout and added computation remain limitations (Apps. E and G).

AI use statement

Generative AI tools (Claude) were used for code generation, experiment orchestration and log analysis, for literature search, and for drafting and editing the text. No research idea, claim or reported number originates from a generative AI tool without author verification: all AI-generated code was reviewed and tested by the authors, and every number in the text is traced to an experiment log by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

All imagery behind the corpus, the held-out sets and the benchmarks comes from published datasets (fMoW, DOTA, DIOR, VRSBench, RSVQA) used under their academic licences at resolutions that are already public; no new imagery was collected and no personal data is involved. Overhead perception has dual uses; the graft adds no sensing capability beyond the public imagery and the frozen public encoder it reads, and its bit-exact dormancy makes its presence in a deployed host auditable. The generated corpus will be released under the licence terms of its source datasets.

Reproducibility statement

The corpus construction rules and acceptance checks are in App. B, the objective, optimisation and evaluation details in App. D, and the exact benchmark protocols in §4.1 and Table 8; the full recipe was rerun end to end once (§4.2). Code, the generated corpus and the held-out set will be released on publication.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv:2507.22291*, 2025.

Miaoxin Cai, Guanqun Wang, Wei Zhang, Guangyao Zhou, et al. Earth-onevision: Extending remote sensing multimodal large language models to more sensor modalities and tasks. *arXiv:2606.10819*, 2026.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

Yezhen Cong, Samar Khanna, Chenlin Meng, et al. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, 2022.

Muhammad Sohail Danish et al. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024. ICCV 2025.

Shihan Dou, Enyu Zhou, Yan Liu, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *ACL*, 2024.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv:2506.20380*, 2025.

Kristjan Greenewald, Luis Lastras, Thomas Parnell, Ambrish Radhakrishnan, Sanjeeb Roy, et al. Activated lora: Fine-tuned llms for intrinsics. *arXiv preprint arXiv:2504.12397*, 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yuan Hu, Jianlong Yuan, Congcong Wen, et al. Rsgpt: A remote sensing vision language model and benchmark. *arXiv:2307.15266*, 2023.

Johannes Jakubik, Sujit Roy, C. E. Phillips, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv:2310.18660*, 2023.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, et al. Geochat: Grounded large vision-language model for remote sensing. In *CVPR*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, 2024. arXiv:2406.12384.

Hui Lin, Danfeng Hong, Shuhang Ge, Chuyao Luo, et al. Rs-moe: A vision-language model with mixture of experts for remote sensing image captioning and vqa. *arXiv:2411.01595*, 2024.

Jiaqi Liu, Lang Sun, Ronghao Fu, and Bo Yang. Towards faithful reasoning in remote sensing: A perceptually-grounded geospatial chain-of-thought for vision-language models. *arXiv:2509.22221*, 2025.

Jiaqi Liu, Ronghao Fu, Lang Sun, Haoran Liu, et al. Skymoe: A mixture-of-experts vision-language foundation model for remote sensing. In *AAAI*, 2026. arXiv:2512.02517.

Xu Liu and Zhouhui Lian. RSUniVLM: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv:2412.05679*, 2024.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. arXiv:2003.07333.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv:2406.10100*, 2024.

Dilxat Muhtar, Zhenshi Li, Feng Gu, et al. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *ECCV*, 2024.

Qwen Team. Qwen3 technical report. *arXiv:2505.09388*, 2025.

Colorado J. Reed, Ritwik Gupta, Shufan Li, et al. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV*, 2023.

Oriane Siméoni et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *CVPR*, 2025.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *NeurIPS*, 2022.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021.

Yuduo Wang and Pedram Ghamisi. RSAdapter: Adapting multimodal models for remote sensing visual question answering. *arXiv:2310.13120*, 2023.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, et al. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *CVPR*, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*, 2023.

Wei Zhang, Miaoxin Cai, Tong Zhang, et al. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing. *IEEE TGRS*, 2024.

A A PRE-REGISTERED NEGATIVE RESULT: PROGRESSIVE EVIDENCE CONSUMPTION

Early in this project we hypothesized a fourth finding: that the stations consume EO evidence *progressively* by depth — shallow stations feeding recognition, deep stations feeding reasoning ("progressive consumption"). We pre-registered acceptance criteria before running the decisive experiments, committed to reporting the outcome either way, and the hypothesis was falsified. We document it here because the instruments are reusable, the intermediate findings strengthened the main results, and silent abandonment of a named hypothesis is how fields accumulate publication bias.

Pre-registered criteria. Progressive consumption would count as real only if, jointly: (1) the reasoning EO-dependence gap recovered from harmful to neutral; (2) recognition remained strongly EO-dependent; (3) the shallowest and deepest stations swapped their dominant question type (crossover); (4) the causal center-of-depth for recognition sat at least one station above reasoning.

Instruments. Three independent probes on a fixed four-type question set (recognition / attribute / spatial / reasoning): *station-drop causality* (disable one station's EO branch, measure gold-answer log-likelihood change); *gate-by-type* (per-station gate readout, bucketed by question type); *EO-dependence confound* (per-type log-likelihood gap between full-EO and host-only).

What the probes found along the way. The journey produced two findings that outlived the hypothesis. First, an information-theoretic bug in our own prompt design: with the <EO_TILE> marker placed *before* the question, the causal attention mask guarantees the branch input contains no question information — question-conditioned gating was architecturally impossible, not hard to learn. Moving the marker after the question unlocked conditioning (gate spread by question type: $0.0005 \rightarrow 0.029$), cleanly separating "cannot condition" from "does not close." Second, a competitive (softmax-share) gate saturates open at stations where native experts are weak; the ability to *condition* is not the ability to *close*, which motivated the independent sigmoid gate used in stage 2.

Verdict. Under full training with every fix applied, three probes agree: the causal depth profile is uniform across question types (center-of-depth spread 0.07 stations), the learned per-station gate profile is flat (five stations open and close in lockstep), and there is no question-type $\times$ depth interaction. The five stations behave as a *single, unstructured on/off unit* — there is no progressive, depth-structured consumption to report. What survives is precisely the multi-depth result of the main paper (§5.2): per-station *causality* is real (removing any early station costs 20–28 log-likelihood), but it carries no question-type structure. A global when-to-invoke gate is trainable (gate BCE $0.77 \rightarrow 0.01$; with same-image negatives and an independent sigmoid gate, open$-$close separation +0.216, §4.4), but it carries no depth structure: all five stations open and close together.

Why report this. The architecture and reported task gains do not depend on this hypothesis (progressive consumption in the pre-registration's labels); the falsification cost the paper nothing but an unearned mechanism story. We kept the pre-registered bar rather than softening it, and we recommend the same discipline for mechanism claims layered on adaptation architectures: phenomena that survive only under favorable analysis choices are not findings.

B CORPUS CONSTRUCTION AND ARRANGEMENT

B.1 TRAINING AND EVALUATION PROTOCOLS

Training data and checkpoints. Stage 1 uses a large generated corpus (2.27M rows) followed by a version-frozen corpus (133,333 rows, 20 slices), built from fMoW-RGB (Christie et al., 2018) and DOTA boxes for counting (Xia et al., 2018). In the frozen corpus, hard annotations supply 41% of rows; the remainder are assembled from them, with machine acceptance checks below. The host receives no weight updates. Stage 1 evaluates the EO path alone; stage 2 enables native vision; stage 3 adapts the side expert on each benchmark's training split. *Target-benchmark zero-shot* means

no training on that benchmark. Frozen-corpus recipe comparisons use 500 steps; longer-run and replication results are reported with the corresponding outcomes.

Evaluation data and tasks. Public evaluations use the full 37,409-question VRSBench split, 10,004-question RSVQA-LR split, and the ordinary and Philadelphia RSVQA-HR splits (§G.2). Diagnostic VRSBench comparisons use a 2,400-row sample; its ten official types contain 2,000 rows. Grounded development questions use 886 of 1,200 rows from 15,263 fMoW sequence/test images, excluding training images. Of these, 110 ask for the category of the object in a named corner over 20 classes; one labelled object per chip makes this identification, not localisation. Geocell predicts a 10° latitude–longitude cell. Answer-determining header fields are redacted. Recipe selection used the development set, so its intervals exclude selection uncertainty. Site overlap, the 97 site-disjoint corner questions, both pre-locked new-site filters and public-benchmark exposure are detailed in the audits below; Table 2 reports the new-site results.

Baselines and controlled comparisons. We compare the text-only host, its native vision path, and the EO expert with another image's evidence (mismatch). Simpler grafts use a station-31 connector without a trunk, a one-station expert, or rank-64 LoRA on all language-model linear layers. Table 3 separates frozen-corpus training from scratch from continuations of the same five-station large-corpus checkpoint; the single-station continuation activates only station 31 in the latter. The benchmark LoRA matches benchmark rows and adaptation steps. The host-LoRA control matches Kite across the full training chain, including the large-corpus stage, frozen-corpus training and later dense continuation. Matching covers training data, stages and step budgets.

Scoring and statistics. Categorical fields use exact match and balanced accuracy; numeric fields use median error and tolerance hit rates, with mode/constant-median baselines. Geocell error is the great-circle distance between predicted and true cell centres. Development comparisons use paired image bootstrap intervals (10,000 resamples); later probes use paired site resampling, with the probe-specific draw counts given below. Each corner question has its own site. We report 95% intervals; intervals excluding zero indicate a detected difference. VRSBench's official protocol applies substring/exact matching before its gpt-4o-mini judge (Li et al., 2024). Diagnostic readings use a lenient frozen-host judge (thinking disabled), with and without a shared text-only short-answer extractor. Category sets and protocols are labelled in Table 8; these readings are not interchangeable. RSVQA-LR uses its official rule-based scoring (Lobry et al., 2020).

B.2 Corpus construction and separation audits

Source imagery and metadata. Training rows are generated over fMoW-RGB (globally distributed sub-meter imagery with per-image ground-sample distance, timestamp, geodesic coordinates and, for many instances, multi-view series) and, for counting, DOTA (oriented object boxes). fMoW is *not* north-up (median off-nadir $\sim$20°), so every directional label is computed from geodesic coordinates, never from image orientation; precise coordinates are always stripped from prompts.

The original corner labels have a median offset of 0.041 of the image width from the nearest midline (85% within 0.1); the off-centre recrops place every anchor at least 0.15 from both midlines.

Two corpus generations. A first, large generation (2.27M rows, twelve template families) is the large corpus of stage 1. A second, version-frozen generation supplies the subsequent frozen-corpus training stage: 133,333 rows in 20 slices, frozen under a versioned manifest (Table 4). Three findings of the audit shaped it. (1) *Held-out contamination*: rows had been sampled without image-level exclusion, so 21%–41% of held-out images also appeared in training; we now enforce image-level disjointness at generation time and report the original grounded diagnostics on the pure subset (§4.1). (2) *Sub-patch position labels*: fMoW crops are object-centred, so "which corner" labels had a median anchor offset of 0.041 of the image width — below one patch; position slices are now generated from off-centre recrops with a uniform target quadrant and a minimum offset of one patch, with a ninth "centre" class for near-centre anchors. (3) *Degenerate fields*: two fields whose mode exceeded 90% were removed from evaluation rather than reported.

Direct hard labels (detection boxes or official metadata) account for 41% of rows, position recrops derived from boxes and geometry for 30%, assembled long reports for 27% and constructed gate

Table 4: Frozen corpus (133,333 rows). "Hard" = derived directly from detection boxes or official metadata; "assembled" = several hard facts composed by rule, every sentence backed by one; "constructed" = paired controls. No row is model-generated.

Slice	Task	Rows	Label source
offc / offc9	corner / 9-way position after off-centre recrop	20,000 / 20,000	hard (box + geometry)
report / report_lite	long report, $\geq$3 / 2 facts + inferred relation	10,549 / 25,751	assembled
direction	inter-object bearing (8-wind, geodesic)	12,109	hard (boxes)
temporal	season	9,296	hard (metadata)
count / count_tiles	whole-image / 1024-tile count, 36% zero negatives	9,322 / 2,070	hard (DOTA boxes)
caption / bbox / mmm / distance	category / extent / geo-band / pixel$\times$GSD distance	3,729 / 3,701 / 3,693 / 2,308	hard
change (order / season / development)	multi-view series	3$\times$2,000	hard (metadata)
corner_anchor / spatial	corner object / pair relation	1,875 / 240	hard
gate (needing / irrelevant / mixed)	same-image EO-relevant vs. irrelevant	1,094 / 1,094 / 502	constructed

pairs for 2%. By token count the corpus is roughly 55% long reports, 30% short QA and 15% position, with at most two short questions per image per epoch. The spatial-reasoning slice is small because 93.5% of the source pairs turned out to involve a false-detection pseudo-entity and were discarded.

Acceptance rules. Every slice passes a machine check before it may enter a training mix: label geometry perceptible (anchor $\geq$1 patch from any decision boundary); answer text and structured value rendered by the same function; required metadata prefix present; answer value absent from prompt (leakage $\leq$5%); no augmenter pseudo-entity as gold; encoder cache coverage 100%; image-level held-out disjointness; per-field mode $\leq$60%; and no in-place overwrite of a frozen corpus version. Several rules were written after the incident they now prevent.

Prompts. Each row renders a metadata prefix (geocell, sensor, time, ground-sample distance, pixel size), the question, then 4 `<EO_TILE>` markers per image frame. The fields that determine or geometrically imply a row's answer are redacted in that row's prefix (`[GEOCELL]` for geocell and recognition questions; `[TIME]` and `[SUN]` for season questions), and a post-check drops any row whose redacted block is not fully blanked. During training the prefix is randomly reordered or partially dropped (format augmentation), and long reports are requested with both caption-style and report-style wordings.

Public-benchmark adaptation. For in-domain benchmark training we use each benchmark's official training split with test images excluded at the image level (RSVQA-LR: 67,228 questions over 672 images; VRSBench: 85,813 questions over 20,262 images). Because these benchmarks' answers are single words, we additionally synthesize a multi-fact description per image from its own question–answer pairs (facts only; a fact-coverage filter removes 17% of outputs) so that in-domain training does not erode long-form ability. The corpus generators, acceptance checker and held-out lists will be released.

Held-out audit. The held-out images are drawn from the fMoW *sequence* and *test* splits (15,263 images); the large corpus draws only from the train and validation splits, so it is disjoint from the held-out set at the image level and at the site level (verified against its 128,605 images: zero overlap). The frozen corpus also draws rows from the sequence and test splits; our original held-out split shared 21%–41% of its *images* with those slices (rows had been sampled without image-level exclusion), so all numbers below use the *pure* subset — 886 of 1,200 rows whose images appear in no training slice — and image-level disjointness is now enforced at generation time. One residual exposure remains and is disclosed: fMoW sites are multi-temporal, and 26.4% of the held-out sites have a view from another date in the continued-training slices (different image, same location). Same-source crops do not exist (one image per fMoW view). Re-scoring on the pure subset preserved every core conclusion and downgraded one borderline comparison.

Site-disjoint re-scoring. Because fMoW category is a per-site label, we re-score the grounded probes on the held-out images whose site has no view of any date in the frozen corpus ($n = 97$ of 110 corner questions). The expert's corner accuracy is 0.43 exact (0.29 balanced; majority 0.24), the paired differences are +0.43 (CI $[0.34, 0.53]$) over the host alone, +0.28 (CI $[0.16, 0.39]$) over the mismatch control, +0.20 (CI $[0.10, 0.29]$) over the LoRA and +0.34 (CI $[0.24, 0.44]$) over the

five-station expert trained from scratch, and the geocell error is 3,140 km below the LoRA's (CI [1,481, 5,097]). Every conclusion of Tables 1 and 3 holds on this subset.

Pre-locked new-site test. The separate 1,250-question probe supplies the new-site test. The test manifest and site exclusions were locked before evaluation. The strict subset uses a rule-based exclusion without consulting answers or model scores: remove every fMoW site in the 1.83M-row large corpus, the 133,333-row frozen corpus, ten earlier probe-input files, or any identifiable fMoW path in 524 earlier result JSONL files. This leaves 377 questions at 357 sites: 198 corner-object, 99 metadata, 68 temporal, and 12 distance questions. All seven compared arms have one unique prediction for each retained question. The corner questions occupy distinct sites; for other fields we resample sites (5,000 paired bootstrap draws). The historical site-overlap audit is limited to records with identifiable image provenance.

Independent-site LoRA comparison. For the host LoRA comparison we use a separate, answer-independent site filter to the same 1,250-question probe. It removes sites used in the audited 1.83M-row large corpus and 133,333-row frozen corpus, as well as sites represented in the earlier site-filtered evaluation records and the earlier grounded-recognition probe input. The retained 932 questions span 848 sites and include 458 corner-object questions at distinct sites. The Kite and host LoRA final checkpoints each produced one answer for every retained question, with no inference errors; corner objects were scored by the same typed exact-match rule. The paired confidence interval in Table 2 resamples sites 10,000 times. This subset uses a different exclusion scope from the 377-question check above; its manifest was also locked before evaluation.

VRSBench evaluation hygiene. The diagnostic VRSBench sample contains 2,400 rows (the first 200 per question type in file order; the file carries twelve type labels, of which the ten official types give 2,000 rows), and 95.6% of those rows are DOTA-sourced chips. The main benchmark results use the full official split. Separately, 52.6% of the DIOR-sourced validation chips have a pixel-identical window in the public DIOR training images (perceptual hash, Hamming distance 0), so every model trained on DIOR/DOTA detection data — including the full-fine-tuned specialists in Table 8 — has image-level exposure with different labels. Our box-derived training slice touches only 59 sampled rows; on the pure subset (DOTA plus non-overlapping DIOR, n=2,341) the gain of the box-derived data is unchanged (paired row-level $\Delta = +0.037$, 95% bootstrap CI [+0.017, +0.056]).

C THE MARKER'S VOCABULARY ROW

One bookkeeping detail matters for exact host preservation. <EO_TILE> is a new special token. The host's tokenizer has 248,077 tokens while its input and output embedding tables (untied) are padded to 248,320 rows; Kite resizes both tables to 248,077+1 rows, so the marker occupies a slot the host never used and every real token keeps its row, and the new row is frozen at the vocabulary mean. Over the original 248,077 tokens the logits are bitwise those of the host; at decoding we mask the marker's own logit, so the output distribution over the original vocabulary is also identical, and the marker can never be emitted.

D IMPLEMENTATION, PROOFS AND OPTIMISATION

EO implementation. We instantiate the construction with Qwen3.6-35B-A3B (Qwen Team, 2025), a frozen host with 40 decoder layers, 256 routed experts, top-8 dispatch, 35B parameters and approximately 3B active per token. The selected blocks F_k are its MoE blocks at layers $\{7, 15, 23, 31, 39\}$ (5 stations; Fig. 2). A frozen DINOv3-7B encoder supplies image features with $d_v = 4096$ and patch size 16, retaining native resolution within a 1024-patch budget and downscaling larger images. Ground-sample distance and pixel size enter the textual metadata prefix. Each image frame uses $M = 4$ literal <EO_TILE> markers, giving $K + 8$ trunk input tokens per station. The read-out adds 2-D Fourier positional features to keys and values; replacing the patch path by read-outs alone collapses recognition (App. E). The shared trunk is the Qwen3.5-2B decoder stack (1.70B), called through `inputs_embeds` without an embedding table or LM head. Station interfaces total 168.8M parameters; a 10.5M patch encoder has two transformer layers of width 512 and

eight heads. The trainable surface is 1.88B; host and evidence encoder remain frozen, and the gate is fixed open in the first training stage.

Cross-attention read-out and station computation. Let $I = \{i : m_i = 1\}$ be the M marker positions and $h_I \in \mathbb{R}^{M\times d}$ their gathered residual states. At each station, each marker reads the evidence once: $r_j = \mathrm{Attn}(q_j, \mathrm{W}(E) + p, \mathrm{W}(E) + p)$, where q_j combines the marker state with a learned slot embedding and p contains 2-D Fourier positional features. The shared trunk consumes $[\mathrm{W}(E)\,;\, r_{1..M}\,;\, \mathrm{D}(h_I)]$: K projected patches, M read-outs and M projected marker states. Its last M outputs are length-normalised, rescaled by a learned ratio of $\|h_I\|$, mapped by U and scatter-added at I with weight $g(h_I)$ (Eq. 2); other positions receive no direct update. Input length is $K+2M$, linear in K for fixed M; computational cost also depends on the trunk's attention structure (App. E).

Proof of Theorem 1. With $m \equiv 0$ each station takes the inert control path and returns $\mathrm{MoE}_{\mathrm{native}}(h)$; every other host module is untouched. The executed computation graph is therefore *identical* to the host's — the same operations in the same order — so equality holds bit for bit, not merely in exact arithmetic.

Proof of Corollary 1. Causal masking: computations at position i depend only on positions $\le i$, and the additive branch first acts at the first marker.

Proposition 2 (Removability)**.** *Suppressing markers at the prompt layer restores f_{host} bit-exactly at inference time (Theorem 1); forcing $g \equiv 0$ with markers present restores it up to the identity $a + 0 = a$, and bit-exactly only if the implementation skips the branch by control flow. The expert is a runtime switch: rollback requires no weight change or redeployment.*

Proposition 3 (Identity at initialization)**.** *With* U *zero-initialized, f_{Kite} equals f_{host} on* all *inputs at step 0 (up to the IEEE identity $a + 0 = a$). Training therefore starts at the host and never leaves the preservation-feasible set: Eq. 1 is not a constraint the optimizer trades against, but an invariant of the search space.*

Objective and optimisation. The objective is token-level cross-entropy with every target-value span up-weighted $8\times$, plus two numeric auxiliaries that alter training only. The host's tokenizer emits digits one token at a time, so at every label position whose target is a digit we take the softmax restricted to the ten digit ids and penalise the Wasserstein-1 distance to the target digit (the sum over the ordered digits of the absolute CDF difference), averaged over digit positions and weighted by 0.3. A regression head on the last station's marker representation outputs $(\mu, \log\sigma^2)$ per numeric kind (distance, count) and is trained with the Gaussian negative log-likelihood $\log\sigma + (y-\mu)^2/2\sigma^2$, $\log\sigma^2$ clamped to $[-6, 4]$. The task losses are combined by uncertainty weighting (Kendall et al., 2018), $\sum_k e^{-s_k} L_k + s_k$ with learnable $s_k \in [-3, 3]$; the router load-balancing term keeps a fixed weight. We use AdamW with two learning-rate groups (10^{-5} trunk, 10^{-4} interfaces) and fp32 master weights for every trainable parameter — in pure bf16 at these rates the trunk's updates fall below the representable resolution and it silently stops moving — scheduled sampling for the first half of training, and prompt-format augmentation (prefix-line shuffling and dropping) that halves sensitivity to prefix formatting (App. E). The host runs in bf16 under FSDP; the replicated trunk is synchronized by explicit all-reduce. Nothing about inference changes beyond the additive branch itself.

Why bitwise. An implementation that *added a zero-valued branch* ($y = \mathrm{native} + 0$) would be mathematically equal but need not be bitwise equal once kernels, fusion or reduction orders differ; exactness comes from control-flow inertness, so equality holds bitwise on a fixed backend, precision and kernel configuration, and as a mathematical identity across configurations.

E Ablations, negative results and supplementary comparisons

Station-count evidence across training settings. Table 5 collects the station-count ablations. The early study varies station count and placement under the same 150k-row recipe (batch size 8; 300 stage-0 and 2,500 stage-1 steps). The last logged losses (step 2,480 of 2,500) favor multiple injection depths; ten stations attain a lower training loss than five. These are individual training-log

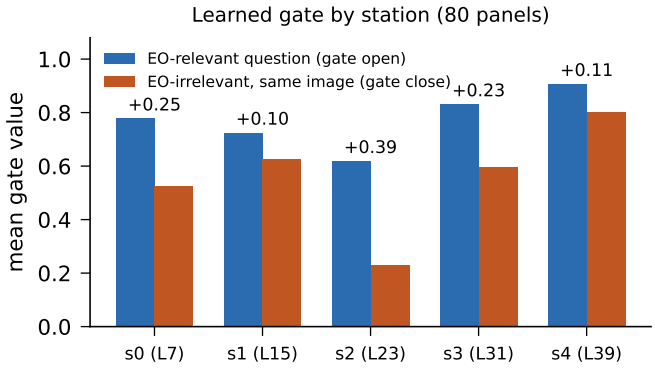


Figure 5: **Learned gate by station.** The stage-2 sigmoid gate, trained on same-image pairs, opens more on the EO-relevant question than on the EO-irrelevant one at every station (mean over 80 training panels; the difference is largest at the middle station), while the five stations move together — the gate is a diagnostic of separability, not a depth structure (App. A).

observations, not averaged test scores. The later recognition results measure task accuracy: five stations outperform one after large-corpus initialization, including on the independent test, whereas frozen-corpus training from scratch yields no detected station-count advantage. These comparisons test station count and placement; they do not compare shared and independent trunks at a fixed station count.

Training-loss readings. Losses are read from the original logs at step 2,480/2,500: 1.62 (L31), 0.99 (L23), 0.26 (five stations) and 0.12 (ten stations). A lower training loss does not establish better held-out performance.

Table 5: Station-count ablations by protocol. Each row is a within-setting comparison; losses and accuracies across rows are not interchangeable. Early losses are rounded from the last logged step of each completed run.

Training / evaluation setting	Results
Early recipe; last logged training loss	1 station (L31): 1.62; 1 station (L23): 0.99; 5 stations: 0.26; 10 stations: 0.12
Large-corpus initialization; development corner exact	1 station: 0.10; 5 stations: 0.43 (Table 3)
Frozen corpus from scratch; development corner exact	1 station: 0.09; 5 stations: 0.08; no detected difference (§5.2)
Large-corpus initialization; independent corner exact ($n = 198$)	1 station: 0.0707; 5 stations: 0.4293; paired difference +0.359, 95% CI [0.288, 0.429] (Table 2A)

Table 6: VRSBench by question type: Kite (in-domain) vs. bare host. The two paths are complementary.

Type	Kite	Bare host	Type	Kite	Bare host
Existence	0.915	0.765	Quantity	0.360	0.480
Category	0.785	0.735	Position	0.525	0.650
Scene	0.875	0.805	Colour	0.820	0.850
Reasoning	0.630	0.715	Size	0.590	0.580
Shape	0.685	0.670	Direction	0.445	0.550

Data, robustness and native-path contributions. On 400 held-out images, entity recall of free-form reports is 0.297 (caption-style prompt) / 0.316 (report-style prompt), with 0.29 of reports fully

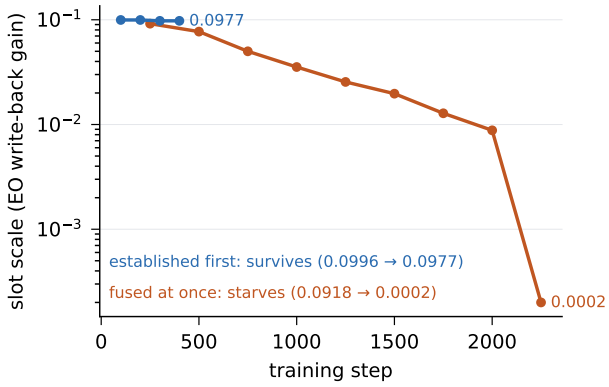


Figure 6: **Channel competition.** The EO write-back gain (slot scale) during co-training with the host's native vision path. A fresh EO channel started from the large-corpus checkpoint together with native vision starves to zero within 2,250 steps; an EO channel established first and then joined by native vision keeps its gain. Establish, then fuse.

correct, versus 0.222 / 0.00 for the large-corpus checkpoint; the gain comes from asking the same content with both prompt styles during training (a model trained with report-style prompts only scored 0.135 under caption-style prompts). Under four prompt-format perturbations, the gap between best- and worst-case accuracy on corner identification drops from 0.17 to 0.09 with format augmentation. Stage 2 lifts VRSBench quantity from 0.360 (expert alone) to 0.52, at parity with the host (0.480), by introducing the native pixel path; occlusion-based counting variants (after SkyMoE) and wider read-outs did not improve counting.

Why not co-train with native vision (channel competition; Fig. 6). Adding the host's native image tokens alongside the EO branch and training both from the large-corpus checkpoint drove the EO branch's write-back scale from 0.0918 to 0.0002: the in-distribution channel absorbs the residual and the new one starves. Keeping the EO channel alive with input dropout does not help, because the frozen fusion point still does not consume it. The same experiment started from the fully trained EO expert leaves the scale unchanged (0.0996→0.0977): a channel whose consumption is already established is not displaced. This "establish, then fuse" order is the design principle behind §4.4.

Longer stage-3 training (negative result). Continuing the VRSBench stage from 1.5 to 2 epochs of the training split (on one GPU at batch 4) does not raise official-protocol accuracy on the 2,400-row VRSBench sample (raw answers): 0.750 against 0.764 for the reported run. The stage-3 budget is not what limits the fine-tuned model.

Position readout (negative result; criteria fixed before the runs). ten single-variable arms — replacing vs. augmenting the patch path, patch dropout, an auxiliary direction head, positional features in attention values — never lifted position readout above chance on a well-posed probe (nine-way grid after off-centre recrop: 0.12 vs. chance 0.11; four-way quadrant through the auxiliary head: 0.27 vs. 0.25, n=726). In the early arms positional features entered only queries and keys, so attention could be *conditioned* on position while the retrieved value carried no coordinate; adding them to the values (the final recipe) did not change the outcome. We therefore state the result as an empirical one: in every configuration we tested, fine-grained position readout through the expert failed; we do not claim it cannot be learned. Coarse natural-language spatial questions (VRSBench position/direction, Table 6) are unaffected. A second cause was in the data: fMoW crops are object-centred, so the original position labels were sub-patch (median offset 0.041 of the image width); off-centre recrops fixed the labels but not the readout.

What the graft costs. To complement parameter counts, we give the forward-pass ledger per image, computed from the model configurations. The expert trunk (24 decoder layers, hidden 2048) costs 2.6 GFLOP per token and runs over $K + 8$ tokens at each of 5 stations; the frozen DINOv3-7B

encoder costs 13.4 GFLOP per patch. At $K = 1{,}024$ an EO query therefore costs about 27 TFLOP (encoder and trunk in equal parts), at $K = 256$ about 6.8 TFLOP, against 3.7 TFLOP for the host's native vision path on a 409,600-pixel image (vision tower plus 400 image tokens through the 3B-active host) and 3.8 TFLOP for that path with the LoRA — a factor of 7 at the largest patch budget and 1.8 at the smallest. Two things reduce it in deployment: the encoder side (13.7 TFLOP of the 27) is a pure function of the image and is pre-computed once for a static archive, as in all our training; and the branch runs only when the router fires, whose own cost is one extra host prefill of the question. Memory adds 3.8 GB of expert weights in bf16. Measured end to end on one B200 through a vLLM plugin, an EO query on a 512-pixel tile takes 1–3 s. The LoRA is cheaper on every line; what the graft buys is native-resolution evidence with the host's weights untouched.

Numeric auxiliaries. On a distance pilot, the ordinal digit loss raised hit@10% from 1.00% to 3.92% and lowered median relative error from 4.77 to 3.78; both auxiliaries are on in the final recipe.

F AGENT-SYSTEM EVALUATION ON EARTHBENCH

Kite on EarthBench. These runs compare Kite in direct mode with its bare host, Qwen3.6-35B-A3B. The official three-part evaluator scores 247 of the 248 completed EarthBench questions: one question lacks official ground truth. Spectrum, Products and RGB contain 100, 88 and 59 questions, respectively. Each mode compares Kite with the bare host on the same questions and hardware. Table 7 reports answer accuracy and gains; ranks insert each Kite run into the 13 same-mode systems supplied with EarthBench and are not official leaderboard submissions. The 247-question aggregate is diagnostic only.

Table 7: EarthBench answer accuracy under the official three-part evaluator (Kite vs. bare Qwen3.6-35B-A3B). AP = answer provided; IF = instruction following. Inserted ranks compare with 13 supplied same-mode systems; the full-set row has no official rank.

Mode	Split (n)	Bare Qwen3.6	Kite (rank/14)	Difference (pp)
AP	Spectrum (100)	46.00%	63.00% (1/14)	+17.00
	Products (88)	51.14%	68.18% (2/14)	+17.05
	RGB (59)	28.81%	40.68% (9/14)	+11.86
	All (247)	108/247 (43.72%)	147/247 (59.51%)	+15.79
IF	Spectrum (100)	41.00%	63.00% (2/14)	+22.00
	Products (88)	51.14%	73.86% (1/14)	+22.73
	RGB (59)	28.81%	33.90% (9/14)	+5.08
	All (247)	103/247 (41.70%)	148/247 (59.92%)	+18.22

The two runs completed all 248 questions; AP explicitly called Kite on 74 questions (436 calls), and IF on 65 questions (398 calls). On Spectrum and Products, several strict trajectory metrics also improve. On RGB, however, TIO/TEM/Parameter fall by -18.99/-18.32/-21.65 points in AP and -13.86/-18.84/-21.95 points in IF. The answer gains therefore do not imply uniformly better planning.

G DETAILS OMITTED FROM THE MAIN TEXT FOR SPACE

This appendix provides supplementary results, category-level comparisons and evaluation details under the protocols of §4.1.

G.1 COMPLETE BENCHMARK LANDSCAPE

G.2 SUPPLEMENTARY RESULTS AND LIMITATIONS

Additional benchmark context. The bare host itself exceeds published LoRA-tuned specialists on VRSBench under its official protocol. Without VQA-style training, EO evidence lifts RSVQA-LR presence from 0.473 to 0.747 and GEOBench-VLM (Danish et al., 2024) counting from 0.009

Table 8: Public-benchmark landscape. Protocol column: O/F = VRSBench's official protocol (substring, exact match, then the gpt-4o-mini judge) on the full 37,409-question split; O/S = the same on our 2,400-row sample; L/S and H/S = our lenient frozen-host judge on that sample, without and with one short-answer extractor applied to every model; R/F and R/S = the official RSVQA-LR three-type average on the full 10,004-question split and on a 4,000-question sample; pub. = as published. VRSBench averages differ by category set: Avg10 = the ten categories of the VRSBench paper, Avg7 = the seven reported by Earth-OneVision (category, existence, position, quantity, scene, colour, image); RSThinker averages nine (no direction); Avg7 is given for official-protocol rows only. The H/S rows apply one answer extractor to every model.

Model	LLM	Trained	RS data	Protocol	VRSBench Avg10	Avg7	RSVQA-LR
Fully fine-tuned RS specialists							
Earth-OneVision	Qwen3 2B	full	34M	pub.	—	0.803	0.929
EarthDial (Soni et al., 2025)	4B	full	—	pub.	—	0.392	0.931
RSThinker (Liu et al., 2025)	GLM-4.1V 9B	SFT + GRPO	380k CoT	pub.	0.759 (9 cat.)	—	—
SkySenseGPT (Luo et al., 2024)	7B	full	1.4M	pub.	—	0.486	—
LoRA-tuned RS specialists							
GeoChat (Kuckreja et al., 2024)	Vicuna 7B	LoRA r64	318k	pub.	0.606	—	0.907
LHRS-Bot (Muhtar et al., 2024)	LLaMA-2 7B	LoRA	—	pub.	—	—	0.892
General VLMs, zero-shot							
Bare host (native vision)	Qwen3.6 35B-A3B	**nothing**	0	O/F · R/F	0.627	0.676	0.689
Bare host (native vision)	same	**nothing**	0	H/S	0.679	—	—
Ours (host bit-exact) and the same-rows LoRA control							
Kite, zero-shot	Qwen3.6 35B-A3B	1.88B side expert	2.27M + 133k	L/S · R/S	0.403	—	0.482 (OA)
Kite, in-domain (EO only)	same	same	+ 86k / + 67k	L/S · R/S	0.663	—	0.742
Host + LoRA (same rows as joint)	same	84.7M LoRA	+ 86k	O/S	0.715	0.786	—
Host + LoRA (same rows as joint)	same	same	+ 86k	H/S	0.739	—	—
Kite joint	same	same	+ 86k	O/S	0.764	0.815	—
Kite joint	same	1.88B side expert	+ 86k / + 67k	O/F · R/F	**0.772**	**0.814**	**0.913**
Kite joint	same	same	+ 86k / + 67k	H/S	0.705	—	—

to 0.148 (5-way MCQ; published models score below chance). Figure 7 shows four selected short-answer comparisons across VRSBench and RSVQA-LR. The selected long-description comparison in Fig. 8 illustrates a corrected river-crossing relation; it does not establish overall caption superiority.

Scope and limitations. The full RSVQA-HR evaluation uses an HR-adapted checkpoint; it is not a zero-shot test of the stage-1 expert. In a multi-turn conversation, later turns after the first marker are perturbed; prefix preservation applies only before that marker (Cor. 1).

Simpler grafts (§5.2). The connector arm keeps the station-31 interface (DINOv3 projection, position-conditioned cross-attention read-out, patch encoder) and writes the read-out straight back into the residual stream with no trunk; it gets no corner answer right (paired Kite minus connector $= +0.43$, CI $[0.34, 0.52]$) and leaves the geocell error at 3,479 km. The LoRA trained on the frozen corpus reaches 0.22 on corner against 0.43 (paired Kite minus LoRA $= +0.21$, CI $[0.13, 0.30]$) with a 2,612 km larger geocell error (CI [1,343, 3,946]). The single-station expert initialised from the large-corpus checkpoint scores lower on the entity-bound tasks: corner 0.10 (paired five-station minus single-station difference $= +0.33$, CI $[0.24, 0.42]$), a 5,053 km larger geocell error, and season accuracy 0.28 vs. 0.42. Trained from scratch on the frozen corpus, five stations and one station score 0.08 and 0.09 on corner, 0.24 and 0.28 on season, and their geocell errors are 4,805 and 2,783 km; the paired geocell-error difference, frozen-corpus LoRA minus five-station expert, is -1,859 km (negative favors LoRA). The earlier station-count study in Table 5 measures training loss and predates the frozen corpus; the original-log readings give ten stations a lower last logged loss than five (protocol and scope: App. E).

What the mismatch control does and does not identify. Substituting another image's evidence tests whether the output depends on the image–question pairing. It is not a strict decomposition of the gain: a wrong image can mislead as well as uninform, and the evidence-dependent and adapter parts may interact, so the score under mismatch bounds the pairing-independent part in neither direction without further assumptions. Per question, the joint model's correctness flips on 4.9% of the ten-type sample rows under the preceding-image swap (51 lost, 47 gained) and on 5.1% under the different-scene swap (52, 50); the EO-only expert flips on 30.9% (467 lost, 152 gained). Note also that "sub-metre" describes ground sample distance and "512-pixel" image size: the grounded set uses native-resolution fMoW tiles (up to 1024 patches) with the native path off, the benchmark

chips 512-pixel tiles with it on, and the two regimes therefore differ in task, data and native-path state together. Cleaner interventions (evidence zeroed with markers present, same-category vs. different-category wrong images, batch permutation, local occlusion) are left to future work; the two swaps reported (preceding image; random image from a different scene and source dataset) agree on every model where both were run. On the grounded set every corner question sits on a distinct fMoW site, so resampling sites and resampling questions give the same intervals (+0.43 vs. host, +0.28 vs. mismatch, +0.20 vs. the LoRA, site-level). The three bare-host VRSBench numbers are 0.627 (official protocol, full split), 0.680 (lenient frozen-host judge, sample) and 0.679 (after the extractor) and are not interchangeable.

Paired short-answer examples. Figure 7 combines VRSBench questions 10990, 5799, and 35124 with RSVQA-LR question 23572. The complete short responses are quoted verbatim. Both arms receive the same image, question, and concise answer instruction; the Kite arm additionally uses its trained EO pathway. These selected examples do not measure population accuracy.

Selected long-description comparison. Figure 8 shows VRSBench caption example 3318. Both arms receive the same original image, benchmark question, and instruction to describe the scene in 3–4 factual sentences, including objects and spatial relationships without speculation. The bare arm is untuned Qwen3.6-35B-A3B; Kite uses the caption-adapted CAP checkpoint, distinct from the VRS5 checkpoint in Fig. 7. Both use a 409,600-pixel native-vision budget, disabled thinking, greedy decoding, and a 4,096-token ceiling. The displayed responses are complete and well below the ceiling. This example was selected after visual inspection of 12 additional paired caption probes, following five initial probes, including unsuccessful candidates. Dataset descriptions provide reference context rather than exhaustive verification of every claim. The figure supports the specific observation below, not an estimate of overall caption quality or an image-evidence causal effect.

Gate, router and evidence dependence on chips (§4.4, §5.3–5.4). The trained sigmoid gate's open−close separation is +0.216 (median +0.244) with 89% of training windows positive. Of the prompt-level router's misses on the held-out pairs every one is a season question whose unredacted header already carries the acquisition date (51% of those fire; land-cover and shadow questions fire at 100%). On external distributions (Table 9) the misses on the VRSBench sample are 54 of the 200 "image" questions (resolution, source, orientation — answerable from the header) and 3 of 200 reasoning questions; no other type is missed. Forcing the gate open on every VRSBench question leaves the stage-2 model at 0.518 against 0.516 with the learned gate. For the EO-only in-domain expert the mismatch control moves category from 0.79 to 0.24, scene from 0.88 to 0.41 and existence from 0.92 to 0.77; for the joint model the same swap leaves the ten-type mean at 0.714 (preceding image) and 0.715 (different scene and source dataset) against 0.716. Under the frozen-host judge on the sample the same-rows LoRA scores 0.739 against 0.716 for the joint model (0.739 vs. 0.705 after the extractor), above the host by +0.059 (CI [+0.037, +0.082]). Extracted readings: joint 0.705, host 0.679. The same-rows LoRA wins on quantity, size and position (0.645, 0.615, 0.805 vs. 0.52, 0.47, 0.69) and loses on colour (0.800 vs. 0.91); the paired joint−LoRA difference is -0.023 (CI [-0.044, -0.002]) under the frozen-host judge and -0.035 (CI [-0.057, -0.015]) after the extractor on both sides, significantly favoring LoRA in each case. These paired comparisons use the frozen-host judge (L/S and H/S), not the official gpt-4o-mini protocol (O/S).

Table 9: Prompt-level router on external question distributions (frozen host, text-only, same system prompt as §4.4).

Distribution	n	fires
VRSBench evaluation sample (VRS header + question)	2,400	97.6%
RSVQA-LR test questions (RSVQA header + question)	500	100%
General-purpose prompts, no header	300	0%
General-purpose prompts, fMoW-style header prepended	300	1.3%

Public benchmarks (§4.2). On the VRSBench sample the two paths win on different types: the expert on existence, category and scene, the native path on counting, colour and position (Table 6). Judges differ across papers by as much as 15–20 points on VRSBench; under our own strict frozen-host judge an earlier in-domain arm scored 0.586 where the lenient reading gave 0.624. Published

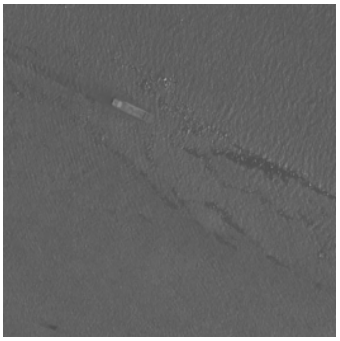
(a) Vehicle color
VRSBENCH | QID 10990



What color is the small vehicle?

Reference	blue
Bare Qwen	white
Kite	Blue

(b) Ship wake
VRSBENCH | QID 5799



Does the ship have visible wake patterns?

Reference	Yes
Bare Qwen	no
Kite	Yes, there is a visible wake pattern.

(c) Relative size
VRSBENCH | QID 35124



Is the top-most overpass larger or smaller than the bottom-most one?

Reference	Smaller
Bare Qwen	larger
Kite	smaller

(d) Building presence
RSVQA_LR | QID 23572



Is a medium building present?

Reference	no
Bare Qwen	Yes
Kite	No.

Figure 7: Selected short-answer comparisons: (a) vehicle color, (b) ship wake, (c) relative overpass size, and (d) building presence. Kite matches the dataset answer in each displayed pair. Both use a 409,600-pixel native-vision budget, disabled thinking, and greedy decoding with a 4,096-token ceiling; both finish naturally well before the ceiling. Kite uses the VRS5 checkpoint for (a–c) and the RSVQA3X checkpoint for (d). These are selected positives, not an accuracy estimate. The relative-size advantage is prompt-dependent: Kite's answer changes under an explanation prompt. Full-split results are in Fig. 4.

River crossing: a more accurate spatial relation

VRSBENCH_CAP | QID 3318 | P2404_0028.png



BARE QWEN

The image shows a sub-meter resolution overhead view of a river running diagonally from the lower left to upper right, flanked on its near bank by a row of low-rise buildings with white and gray roofs. Adjacent to these structures is a dense patch of green vegetation, likely trees or shrubs, which separates the built area from agricultural fields in the upper portion of the image. These fields are divided into rectangular plots, some appearing cultivated, others fallow or overgrown. On the far right edge, a paved road runs parallel to the river, with vehicles visible along its length, and additional buildings clustered near the top right corner.

KITE (CAP)

The image, sourced from GoogleEarth, shows a landscape with a river running through it. There is a bridge located at the bottom-right of the image. The bridge spans across the river, connecting the two sides of the landscape. The surrounding area includes various buildings and greenery.

Visible comparison: the right-edge bridge crosses the river, rather than running parallel to it.

Figure 8: Selected VRSBench long-description comparison (QID 3318). Kite describes a bridge crossing the river; bare Qwen incorrectly calls the right-edge road parallel to it. Complete responses under the shared prompt are shown; this case does not establish overall caption superiority.

zero-shot general VLMs on VRSBench: GPT-4V 0.656, Kimi-VL-Thinking 0.691; EarthMind reports 0.789 under its own protocol (its table scores GeoChat at 0.760 where the official protocol gives 0.606), so it is not placed in Table 8. An oracle that picks the better of the expert and the native path per question type reaches 0.70 on the VRSBench sample, only +0.02 above the native path alone, so type-level complementarity is real but small. The official protocol credits 72% of the joint model's answers and 53% of the bare host's by substring before any judge is consulted; the extracted readings are 0.705 (joint) and 0.679 (host).

RSVQA-HR full-split evaluation. We evaluate the fixed HR-adapted checkpoint on every active question in the official ordinary test and Philadelphia files. The checkpoint starts from S2T and receives 2,000 additional training steps on the HR training split, while the host weights remain frozen. This setting is *benchmark-adapted*, not target-benchmark zero-shot. Both splits have complete question-ID coverage and zero inference errors. For the four-type evaluation we score counts as exact integers and areas using the dataset's area bins. Avg2 averages presence and comparison only; AA4 averages all four types; OA weights every question equally.

Table 10: Completed RSVQA-HR results for the HR-adapted side expert with a frozen host. All scores are percentages.

Split	Questions	Presence	Comparison	Count	Area	Avg2	AA4	OA
Ordinary test1	222,684	86.00	81.88	43.41	78.08	83.94	72.34	72.35
Philadelphia	105,647	84.50	83.01	41.00	75.28	83.75	70.95	71.30

The earlier 500-question Philadelphia diagnostic overlaps the full split. Removing those questions leaves 105,147 items and gives OA 71.29% and AA4 70.94%. For a published ordinary-test1 reference, RSAdapter (Wang & Ghamisi, 2023) Table II reports OA 85.81% and AA4 85.52% after target-dataset training; its presence/comparison mean is 92.32%. These are external paper results, not matched reruns on our question manifest. Earth-OneVision (Cai et al., 2026) Table VI reports 86.36% for presence/comparison only, without identifying ordinary test1 versus Philadelphia in that table, so we do not treat it as an ordinary-test1 OA reference.