SEEING THE WHOLE SCENE: SCALING VISION-LANGUAGE MODELS FOR ULTRA-HIGH-RESOLUTION EARTH OBSERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultra-high-resolution (UHR) Earth observation imagery enables fine-grained reasoning over large areas, but rapidly growing visual sequences make full-attention remote sensing vision-language models (RS-VLMs) difficult to scale. Existing UHR RS-VLMs typically control cost through multiscale tiling, adaptive region selection, or token pruning and merging. These strategies bound computation by reducing the visual sequence, requiring increasingly aggressive visual compression as image size grows and potentially discarding fine spatial detail or query-relevant information. We introduce SourceScale-VL, a post-training-converted RS-VLM that preserves the complete post-encoder visual sequence and scales it through ring-style sequence parallelism, compressing visual interactions rather than visual content. SourceScale-VL replaces quadratic visual propagation with linear attention and introduces a magnitude-aware Referential Spectrum Residual to selectively retain Exact text-to-visual interactions where linear attention lacks sufficient source-response capacity. To preserve source-specific behavior during conversion, we further introduce Source-Factored On-Policy Distillation, which aligns source-wise output contributions on student-generated states. On the held-out conversion-evaluation split, SourceScale-VL reaches 57.46% accuracy, outperforming the same-budget RSR-Hybrid by 0.43 percentage points while reducing SCE from 0.063 to 0.041. On XLRS-Bench-lite and LRS-VQA, it achieves 54.64% and 33.58% aggregate accuracy, slightly surpassing Teacher (Exact) by 0.14 and 0.18 percentage points, respectively. With four A800 GPUs, SourceScale-VL supports $1.12\times$ longer visual sequences than sequence-parallel Exact attention while reducing peak memory by 13.8% and TTFT by 15.5%.

1 INTRODUCTION

Ultra-high-resolution (UHR) Earth observation imagery simultaneously preserves fine-grained objects and their spatial organization over large geographic extents. This combination is essential for Earth-observation tasks in which small targets or local structures must be interpreted together with broader scene context and spatial relationships, including infrastructure inspection, urban monitoring, transportation analysis, and disaster assessment Xiao et al. (2025); Wang et al. (2025b).

Remote sensing vision-language models (RS-VLMs) have expanded remote-sensing interpretation from task-specific recognition toward open-ended question answering, grounding, and multimodal reasoning Kuckreja et al. (2023); Zhan et al. (2024). Extending these capabilities to UHR imagery, however, produces rapidly growing visual sequences whose full-attention interactions impose substantial memory and computation costs during both training and inference Wang et al. (2025b;a).

Existing UHR RS-VLMs primarily control these costs by reducing the visual sequence presented to the language model. Representative approaches include multiscale tiling and encoding, adaptive region selection and zooming, and token pruning or merging Wang et al. (2025a); Luo et al. (2025); Dang et al. (2026); Wang et al. (2026). These strategies substantially improve tractability, but their scalability remains coupled to visual compression. Under a bounded token budget, increasing image resolution requires stronger reduction, which limits retained spatial detail and may exclude small or query-relevant regions. Adaptive search can recover selected high-resolution content, but the

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.

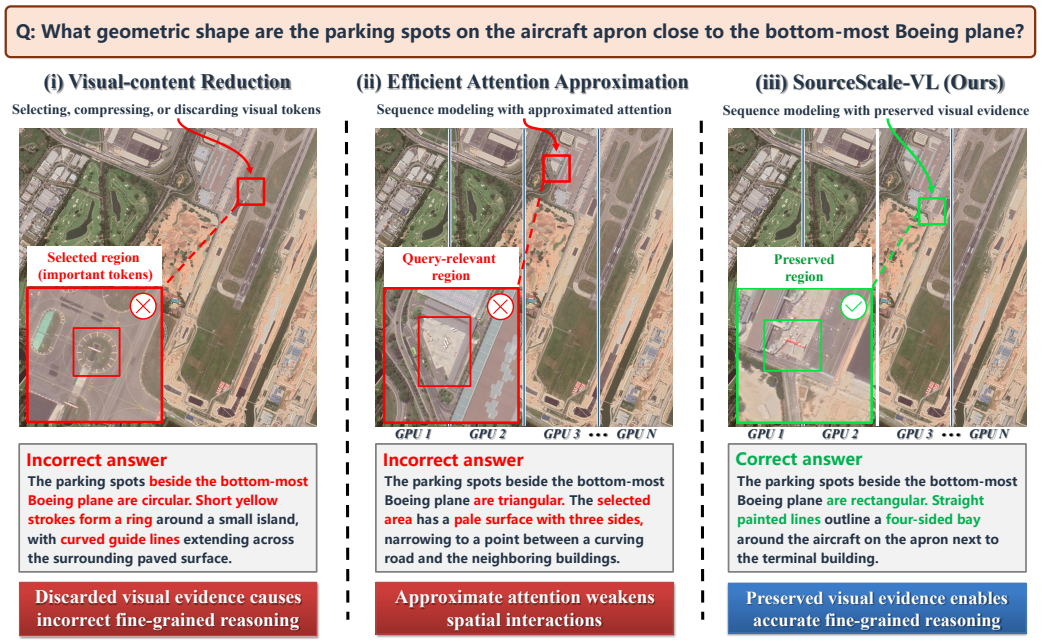

Figure 1: **Motivation of SourceScale-VL.** Existing approaches either reduce visual content or approximate full attention. SourceScale-VL preserves complete visual tokens and selectively compresses visual interactions while maintaining critical source-dependent retrieval.

resulting visual context remains dependent on the search or selection policy Wang et al. (2026); Liu et al. (2026). The central limitation is therefore not whether large images can be processed, but whether visual coverage can continue to grow with image resolution without relying on increasingly restrictive content reduction.

Sequence parallelism and efficient attention provide complementary mechanisms for scaling long visual sequences without first reducing their visual content. Sequence parallelism distributes long sequences across devices Liu et al. (2023), while linear attention reduces interaction complexity by compressing key-value information into compact states Katharopoulos et al. (2020). Recent work has increasingly combined these mechanisms, extending distributed execution to linear and hybrid attention and enabling substantially longer sequences with reduced computation and memory Xue et al. (2024); Sun et al. (2024; 2025); Chou et al. (2025). These advances establish distributed efficient attention as a promising alternative to visual-token reduction. However, existing designs primarily optimize sequence scalability, system efficiency, or aggregate model fidelity Zhang et al. (2024); Goldstein et al. (2025); Li et al. (2025), without explicitly characterizing which visual interactions can be compressed while preserving fine-grained retrieval. This limitation is particularly relevant to UHR imagery, where numerous small and visually similar instances must be distinguished according to the language query. The central challenge is therefore not merely to scale the visual sequence, but to compress its interactions while preserving query-conditioned distinctions among spatial sources. Figure 1 illustrates this distinction between visual-content reduction and interaction conversion.

We introduce SourceScale-VL, obtained through Factor-OPD, a source-aware post-training conversion procedure for visual-prefix VLMs. The conversion preserves the post-encoder visual sequence, token order, positional information, and pretrained backbone, while an exact sequence-parallel reference provides the full-attention target. Factor-OPD linearizes visual-to-visual propagation to remove the dominant quadratic cost and leaves text-to-text reasoning unchanged. For text-to-visual retrieval, it derives a magnitude-aware Referential Spectrum Residual (RSR) that identifies layer and visual-KV groups whose query-conditioned source-response structure exceeds the representational capacity of the target linear branch, retaining exact retrieval only for these capacity-critical groups. The remaining groups are converted to compact linear states. Source-Factored On-Policy Distillation further aligns source-wise output contributions on student-generated states, preserving

source-specific behavior during conversion. The resulting architecture retains the complete visual sequence while allocating exact capacity only where required.

We evaluate SourceScale-VL on XLRS-Bench-lite Wang et al. (2025b) and LRS-VQA Luo et al. (2025) with LongVILA-R1-7B Chen et al. (2025) as the Exact-attention reference. SourceScale-VL achieves 54.64% and 33.58% aggregate accuracy, slightly surpassing Teacher (Exact) by 0.14 and 0.18 percentage points, respectively. On the held-out conversion-evaluation split, it reaches 57.46% accuracy, remaining within 0.36 points of the Exact teacher while reducing SCE by 34.9% over the same-budget RSR-Hybrid. With four A800 GPUs, it supports $1.12\times$ longer visual sequences while reducing peak memory and TTFT by 13.8% and 15.5%, respectively.

Our contributions are summarized as follows:

- We formulate source-response fidelity as a criterion for efficient UHR VLM conversion, capturing whether query-conditioned distinctions among spatial visual sources are preserved when long-sequence interactions are compressed.
- We introduce SourceScale-VL through Factor-OPD, combining linear visual propagation, RSR-guided hybrid visual retrieval, and Source-Factored On-Policy Distillation to preserve the complete visual sequence while reducing interaction cost and maintaining source-specific responses.
- We demonstrate on XLRS-Bench-lite and LRS-VQA that SourceScale-VL preserves Exact-reference performance while improving source-response fidelity over converted baselines and reducing peak memory and TTFT.

2 Related Work

2.1 Ultra-High-Resolution Remote-Sensing Vision-Language Models

Recent UHR RS-VLMs aim to extend vision-language reasoning to large remote-sensing scenes while preserving fine-grained spatial information beyond the capacity of conventional visual encoders. Existing approaches mainly improve scalability through multiscale processing, query-guided token reduction, and adaptive visual acquisition. GeoLLaVA-8K Wang et al. (2025a) improves high-resolution reasoning through visual token pruning and anchored selection, while C2F (Coarse-to-Fine) Luo et al. (2025) performs hierarchical region refinement to progressively focus on informative areas. UHR-BAT Dang et al. (2026) introduces adaptive token compression for efficient high-resolution understanding, GeoEyes Wang et al. (2026) enables query-dependent visual information acquisition, and RADAR Liu et al. (2026) further explores query-conditioned relative attention and progressive visual evidence acquisition for large-scale remote-sensing reasoning. Although these methods substantially improve UHR visual reasoning, they mainly achieve scalability through selective processing, compression, or adaptive acquisition of visual information under constrained computation. As image resolution continues to increase, preserving complete spatial coverage and query-relevant fine-grained evidence while maintaining efficiency becomes increasingly challenging. This motivates approaches that scale vision-language models through efficient long-sequence modeling, improving computational efficiency while preserving richer visual context.

2.2 Efficient Attention for Long Visual Sequences

Efficient attention mechanisms, distributed sequence modeling, and post-training conversion provide complementary directions for scaling visual sequences beyond the capacity of conventional transformers. LongVILA Xue et al. (2024) demonstrates multimodal sequence parallelism for distributing extremely long vision-language sequences across multiple devices. LASP Sun et al. (2024) combines ring-style sequence parallelism with linear attention through compact recurrent states, while LASP-2 Sun et al. (2025) reduces communication overhead by exchanging sequence-length-independent memory states and further extends the framework to hybrid softmax-linear attention through LASP-2H. ZeCO Chou et al. (2025) improves distributed linear attention through communication-efficient collective operations. Meanwhile, post-training conversion methods aim to transform pretrained softmax attention into efficient alternatives without full retraining. LoLCATs Zhang et al. (2024) transfers pretrained attention to linear operators via attention matching

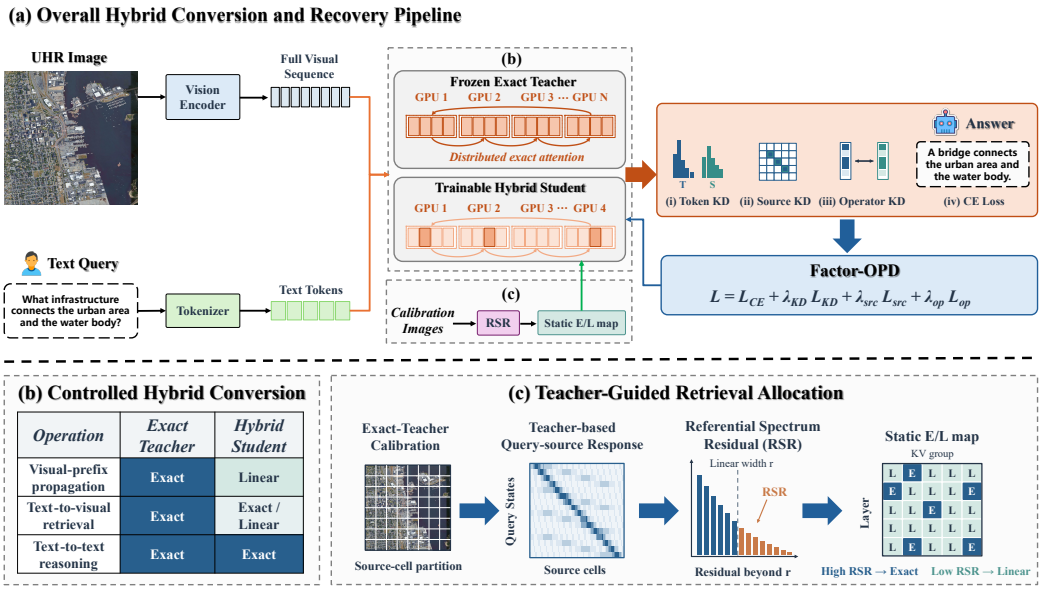

Figure 2: **SourceScale-VL hybrid conversion and recovery.** (a) Factor-OPD distills the Exact teacher into the hybrid student. (b) Hybrid attention combines Linear visual propagation with selective Exact/Linear retrieval. (c) RSR calibration determines the retrieval conversion map.

and lightweight recovery, while RADLADS Goldstein et al. (2025) and subsequent hybrid conversion approaches Li et al. (2025) improve recovery efficiency or selectively preserve softmax attention in capacity-sensitive components. However, existing approaches mainly optimize efficient attention from a sequence-level perspective, emphasizing scalability, communication efficiency, or global output recovery. They overlook the query-dependent nature of visual evidence and lack mechanisms to preserve critical spatial interactions during conversion. SourceScale-VL addresses this limitation through capacity-aware retrieval allocation and source-preserving conversion, enabling efficient scaling without sacrificing fine-grained visual evidence.

3 SourceScale-VL

3.1 Overview

SourceScale-VL preserves the complete post-encoder visual sequence while reducing interaction cost through linear visual propagation, RSR-guided Exact/Linear retrieval, and source-aware on-policy distillation. Figure 2 summarizes the framework; Secs. 3.2 and 3.3 detail hybrid conversion and recovery, respectively.

3.2 Source-Aware Hybrid Attention Conversion

SourceScale-VL converts visual interactions while leaving the original language-to-language attention operator unchanged. Visual-prefix propagation is replaced by causal linear attention to remove its quadratic dependence on visual sequence length. Language-to-visual retrieval is converted selectively because its query-conditioned source-response patterns determine the spatial evidence contributing to the final answer and exhibit heterogeneous approximation difficulty across layers and KV groups, while a width-r linear branch imposes a fixed representational constraint that cannot uniformly preserve such fine-grained dependencies.

Let $q_{t,h}, k_{i,g}, v_{i,g} \in \mathbb{R}^{d_h}$ denote the query at position t and query head h, and the key and value at position i and KV group g, where d_h is the head dimension. For interactions assigned to linear attention, we introduce the positive feature mappings

$$\phi_{Q,h}(q) = \frac{\mathrm{softplus}(R_{Q,h}q)}{\sqrt{r}}, \qquad \phi_{K,g}(k) = \frac{\mathrm{softplus}(R_{K,g}k)}{\sqrt{r}}, \tag{1}$$

where r is the feature width. $R_{Q,h} \in \mathbb{R}^{r\times d_h}$ and $R_{K,g} \in \mathbb{R}^{r\times d_h}$ are newly introduced trainable feature projections for the linear branch. The query mapping is head specific, while the key mapping is shared by the query heads associated with the same KV group, following the grouped-query structure of the carrier Ainslie et al. (2023).

For visual-prefix propagation, the complete prefix is partitioned into ordered contiguous shards $\mathcal{P}_1, \ldots, \mathcal{P}_P$ across P devices. Let $t \in \mathcal{P}_p$ denote a visual-prefix position located on device p, and let t_p^{end} denote the last position of shard $\mathcal{P}_p$. The causal linear state admits the decomposition:

$$\begin{aligned} \bar{z}_{p,t} &= \sum_{\substack{i\in\mathcal{P}_p \\ i\le t}} \phi_K(k_i), & \bar{M}_{p,t} &= \sum_{\substack{i\in\mathcal{P}_p \\ i\le t}} \phi_K(k_i)v_i^\top, \\ Z_p &= \bar{z}_{p,t_p^{\text{end}}}, & B_p &= \bar{M}_{p,t_p^{\text{end}}}, \\ z_{p,t} &= \sum_{p'<p} Z_{p'} + \bar{z}_{p,t}, & M_{p,t} &= \sum_{p'<p} B_{p'} + \bar{M}_{p,t}. \end{aligned} \tag{2}$$

Here $(\bar{z}_{p,t}, \bar{M}_{p,t})$ summarize the current shard up to position t, while (Z_p, B_p) summarize the complete shard. A ring-style prefix scan propagates these boundary states across adjacent devices, so the communicated state depends on r and d_h rather than on visual sequence length Sun et al. (2024).

Language queries additionally attend to previous language tokens. For a language query $q_{t,h}$, let $(\mathbf{n}_{t,h}^{\text{txt}}, d_{t,h}^{\text{txt}})$ denote the Exact numerator and normalization denominator contributed by accessible language keys. For a visual retrieval group converted to Linear attention, we define

$$\begin{aligned} \mathbf{n}_{t,h}^{\text{vis}} &= \gamma_h M^\top \phi_{Q,h}(q_{t,h}), \quad d_{t,h}^{\text{vis}} = \gamma_h \phi_{Q,h}(q_{t,h})^\top z, \\ \mathbf{u}_{t,h} &= \frac{\mathbf{n}_{t,h}^{\text{txt}} + \mathbf{n}_{t,h}^{\text{vis}}}{d_{t,h}^{\text{txt}} + d_{t,h}^{\text{vis}} + \epsilon}. \end{aligned} \tag{3}$$

Here (M, z) summarize the complete accessible visual prefix, $\gamma_h = \exp(\beta_h) > 0$ is a learnable head-specific scale that calibrates the Linear visual statistics against the unchanged Exact language statistics, and $\epsilon > 0$ ensures numerical stability. Exact retrieval groups replace the visual terms in Eq. 3 with their corresponding softmax statistics, while both retrieval modes remain jointly normalized with the textual context.

We use superscripts T and S to denote the Exact reference and the converted student, respectively. To determine which retrieval groups require Exact attention, we characterize the query-conditioned spatial response of the reference model. For image s, its physical visual tokens are partitioned into J_s source cells $\{G_{s,1}, \ldots, G_{s,J_s}\}$ according to their coordinates in the original image. Let m index a collected query-head pair (t, h) associated with KV group g. The Exact attention mass assigned to source cell j is

$$A_{\ell g,mj}^{T,s} = \sum_{i\in G_{s,j}} \alpha_{\ell g,mi}^{T,s}, \tag{4}$$

where $\alpha_{\ell g,mi}^{T,s}$ denotes the Exact attention weight of visual token i at layer ℓ. Stacking K_s query-head rows yields $A_{\ell g}^{T,s} \in \mathbb{R}^{K_s\times J_s}$, which records how the same scene is retrieved across different language query states and query heads.

Let $Q_\phi \in \mathbb{R}^{K_s\times r}$ contain the mapped query rows, $Z = [z_1, \ldots, z_{J_s}] \in \mathbb{R}^{r\times J_s}$ the source-level key states with $z_j = \sum_{i\in G_{s,j}} \phi_{K,g}(k_i)$, D the diagonal matrix of joint normalization denominators, and Γ the corresponding row-wise scales γ_h.

$$A^S = D^{-1}\Gamma Q_\phi Z, \qquad \operatorname{rank}(A^S) \le r. \tag{5}$$

Thus, Linear source responses are restricted to an r-dimensional basis. To quantify teacher-response structure beyond this limit, we remove query- and source-independent main effects using $\widetilde{A} = H_K A_{\ell g}^{T,s} H_J$, where $H_K = I_{K_s} - \frac{1}{K_s}\mathbf{1}\mathbf{1}^\top$ and $H_J = I_{J_s} - \frac{1}{J_s}\mathbf{1}\mathbf{1}^\top$ center the query and source dimensions. Let $\sigma_1 \ge \sigma_2 \ge \cdots$ denote the singular values of $\widetilde{A}$. For width r, we define

$$R_{s\ell g}(r) = \sqrt{\frac{\sum_{a>r} \sigma_a^2(\widetilde{A})}{K_s J_s}}. \tag{6}$$

By the Eckart–Young theorem Eckart & Young (1936), the numerator in Eq. 6 is the minimum squared residual of any rank-r approximation. Large Referential Spectrum Residual (RSR) therefore indicates source-specific structure beyond Linear capacity; matrix-size normalization preserves residual magnitude rather than only relative spectral energy.

We average $R_{s\ell g}(r)$ over calibration examples to obtain one score $R_{\ell g}$ for each layer–KV-group pair, termed a retrieval atom. Given an Exact budget K_{exact}, the highest-RSR atoms retain Exact retrieval and the remainder use Linear retrieval. The resulting E/L map is fixed after calibration and introduces no input-dependent routing at inference.

3.3 Source-Factored On-Policy Distillation

After the hybrid architecture is fixed, Factor-OPD recovers the converted model with source-factored supervision. Conventional distillation of logits or aggregated hidden states Hinton et al. (2015) can preserve the final response while allowing errors from different spatial sources to compensate for one another. This ambiguity is particularly relevant to UHR imagery containing many repeated and visually similar regions.

While A measures the attention mass assigned to each source, C retains the value-bearing contribution of that source. For model $M \in \{T, S\}$, let $C^M_{\ell,t,j} \in \mathbb{R}^{d_{\text{model}}}$ denote the contribution of source cell $G_{s,j}$ to language state t after joint normalization and the frozen attention output projection at layer ℓ. Instead of matching only the sum of these contributions, we align them before spatial aggregation:

$$\mathcal{L}_{\text{src}} = \mathbb{E}_{\ell,t}\left[\frac{\sum_{j=1}^{J_s}\left\|C^S_{\ell,t,j} - \text{sg}\left(C^T_{\ell,t,j}\right)\right\|_2^2}{\sum_{j=1}^{J_s}\left\|C^T_{\ell,t,j}\right\|_2^2 + \epsilon}\right], \tag{7}$$

where sg denotes stop-gradient. Source-wise supervision prevents contribution errors across spatial regions from canceling after aggregation and complements RSR by preserving source-specific value contributions.

Distillation is performed on student-generated states Agarwal et al. (2023). The student first generates an autoregressive prefix $\widetilde{y}_{<t}$, after which both models evaluate the same image, question, and prefix. Token-level and source-level targets are therefore defined on states that the converted model actually visits during inference, reducing the mismatch introduced by teacher-forced distillation alone.

Let $\mathcal{Q}$ denote the subset of training examples for which the Exact reference produces a correct answer under the corresponding task evaluator. Ground-truth cross-entropy and operator alignment are computed over all training examples, while token-level and source-level distillation are restricted to $\mathcal{Q}$. To prevent aggregate operator drift during joint recovery, we retain an output-level alignment objective. Let $O^M_{\ell,t}$ denote the complete attention output after the frozen output projection and before residual addition. We define

$$\mathcal{L}_{\text{op}} = \mathbb{E}_{\ell,t}\left[\frac{\left\|O^S_{\ell,t} - \text{sg}\left(O^T_{\ell,t}\right)\right\|_2^2}{\left\|O^T_{\ell,t}\right\|_2^2 + \epsilon}\right]. \tag{8}$$

The recovery objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KD}}\mathbb{E}_{n\in\mathcal{Q}}\left[\mathcal{L}_{\text{KD},n}\right] + \lambda_{\text{src}}\mathbb{E}_{n\in\mathcal{Q}}\left[\mathcal{L}_{\text{src},n}\right] + \lambda_{\text{op}}\mathcal{L}_{\text{op}}, \tag{9}$$

where $\mathcal{L}_{\text{KD}}$ is the token-level distillation loss evaluated on student-generated states. This joint recovery combines task supervision with exact-model guidance while preserving the source-specific behavior affected by attention conversion.

4 Experiments

4.1 Experimental Setup

Datasets. We use the remote-sensing visual question answering corpus adopted by GeoLLaVA-8K Wang et al. (2025a) and partition it by physical image identity before adaptation. The corpus

covers recognition, counting, object properties, spatial relationships, route planning, and complex scene reasoning. We reserve approximately 2K question-answer pairs from 500 unseen images for controlled conversion evaluation, 64 image-disjoint UHR scenes for RSR calibration, and 50 independent UHR scenes for Exact-restoration oracle analysis. All remaining images form the training partition used for remote-sensing adaptation and Factor-OPD conversion. The calibration subset is used only to compile the static Exact/Linear allocation, while the oracle subset is used only for selector analysis.

Metrics. We report accuracy for all 13 tasks in XLRS-Bench-lite Wang et al. (2025b) and compute the overall score as their macro average. OLUC is evaluated by exact matching between the predicted and reference answer sets. For LRS-VQA Luo et al. (2025), we use the official evaluator and report Average Accuracy (AA) on FAIR1M, Bridge, and STAR together with their unweighted mean. For controlled conversion analysis, teacher-student consistency is measured using token-level $\mathrm{KL}(p_T \| p_S)$ and Source Contribution Error (SCE) from Eq. 7. Both metrics are evaluated on identical student-generated prediction states and on Teacher-qualified examples, defined as samples answered correctly by Teacher (Exact) under the corresponding task evaluator. Efficiency is measured using peak per-device memory, time to first token (TTFT), and maximum feasible post-encoder visual sequence length.

Baselines. Application-level comparisons include general-purpose VLMs (Qwen2.5-VL Bai et al. (2025b) and Qwen3-VL Bai et al. (2025a)), a matched-training baseline with Qwen3-VL adapted by RS-SFT, and UHR RS models (GeoLLaVA Wang et al. (2025a), UHR-BAT Dang et al. (2026), C2F Luo et al. (2025), DeepEyes Zheng et al. (2025), and GeoEyes Wang et al. (2026)). Controlled comparisons use the same remote-sensing-adapted LongVILA backbone Xue et al. (2024), complete visual sequence, initialization, and conversion budget. Teacher (Exact) preserves the original attention operators. Propagation-only replaces visual-prefix propagation with linear attention while keeping Exact visual retrieval. All-Linear further converts all visual retrieval groups. AT-Linear applies attention-output transfer under the linear architecture. KL-Hybrid and RSR-Hybrid retain $K_{\mathrm{exact}} = 28$ retrieval groups selected by KL-based and RSR-based allocation, respectively. SourceScale-VL adopts RSR-based allocation and further applies Source-Factored On-Policy Distillation.

Implementation Details. We first adapt LongVILA-R1-7B Chen et al. (2025) to the training partition using LoRA Hu et al. (2021) and freeze the resulting checkpoint as Teacher. Factor-OPD subsequently updates only the newly introduced feature projections R_Q, R_K, and calibration scales β, while the adapted carrier and retained Exact retrieval groups remain frozen. The primary configuration uses $r = 32$ and $K_{\mathrm{exact}} = 28$, corresponding to a 25% Exact budget over the 112 retrieval atoms. The RSR allocation map is compiled once from the calibration scenes and remains input independent during inference. Experiments use up to 12 NVIDIA A800 80GB GPUs across three four-GPU nodes. Fixed-input efficiency profiling uses four GPUs, while distributed scaling uses 1, 2, 4, 8, and 12 GPUs. All runs use BF16 precision, batch size one, identical visual inputs, and a fixed 32-step decoding protocol.

4.2 Main Results

Overall comparison Table 1 reports fine-grained performance on XLRS-Bench-lite and LRS-VQA. SourceScale-VL achieves 54.64% XLRS Avg. and 33.58% LRS Avg., slightly surpassing Teacher (Exact) by 0.14 and 0.18 percentage points, respectively, despite linearizing visual-prefix propagation and 75% of the visual retrieval atoms. Because Factor-OPD combines ground-truth supervision with teacher guidance, the converted student is not constrained to reproduce teacher errors, allowing small downstream gains over the Exact reference. Among the evaluated non-reference models, SourceScale-VL also achieves the highest aggregate score on both benchmarks. These results show that interaction conversion can preserve the complete visual sequence while maintaining, and in some cases slightly improving, downstream reasoning accuracy.

Controlled attention conversion. Table 3 compares conversion strategies under identical visual sequences. Propagation-only approaches the Exact teacher with 57.61% accuracy and 0.030 SCE, demonstrating that visual context propagation can be efficiently captured by compact linear states.

Table 1: Performance comparison on XLRS-Bench-lite and LRS-VQA. Methods are grouped according to their primary strategy for handling ultra-long visual sequences. Bold and underlined values indicate the best and second-best results among non-reference methods, respectively, based on unrounded scores.

	XLRS Perception ↑								XLRS Reasoning ↑					XLRS ↑	LRS-VQA ↑			
Method	OC	RC	OLUC	RLUC	OCC	OCL	OMS	OSR	AD	ECR	RP	RCCD	CCR	Avg.	FAIR	Bridge	STAR	Avg.
General-purpose vision-language models																		
Qwen2.5-VL-7B	33.3	40.0	31.0	77.0	40.6	40.5	66.7	36.2	68.0	72.0	27.0	38.3	45.0	47.4	22.2	27.4	26.7	25.4
Qwen3-VL-8B	21.7	50.0	26.0	81.5	46.6	43.1	66.7	30.4	74.0	79.0	37.0	43.3	51.0	50.0	29.7	32.8	35.0	32.5
Matched-training baseline																		
Qwen3-VL-8B + RS-SFT	28.3	**51.0**	34.0	**82.0**	49.9	47.3	**70.0**	33.6	76.0	78.0	48.0	45.0	**53.0**	53.5	30.2	33.8	**35.9**	33.3
UHR visual content reduction methods																		
GeoLLaVA-8K	26.7	38.0	49.0	69.0	41.6	31.6	65.0	35.0	67.0	78.0	**66.0**	**50.0**	52.0	51.5	21.2	33.0	25.1	26.4
UHR-BAT	21.7	33.0	**50.0**	55.5	43.5	33.8	65.0	**44.8**	62.0	71.0	54.0	46.7	51.0	48.6	24.0	37.6	28.2	29.9
C2F	26.7	39.0	27.0	68.0	35.3	32.3	66.7	32.6	68.0	78.0	47.0	43.3	44.0	46.8	21.9	**38.2**	26.7	28.9
Adaptive visual selection methods																		
DeepEyes-7B	31.7	41.0	33.0	75.0	41.6	38.1	68.3	31.4	70.0	78.0	46.0	**50.0**	46.0	50.0	22.1	31.8	27.0	26.9
GeoEyes	**38.3**	40.0	24.0	73.5	**59.5**	**66.1**	68.3	32.2	**77.0**	80.0	56.0	40.0	50.0	54.2	23.6	34.9	26.9	28.4
Exact-attention reference and converted model																		
Teacher (Exact)	34.1	43.0	39.5	73.2	52.5	50.8	67.5	34.7	73.5	81.0	62.5	46.2	50.0	54.5	31.0	35.0	34.2	33.4
SourceScale-VL (Ours)	33.3	44.0	39.0	74.0	52.1	50.4	66.7	35.2	73.0	**82.0**	62.0	46.7	52.0	**54.6**	**30.7**	35.4	34.6	**33.6**

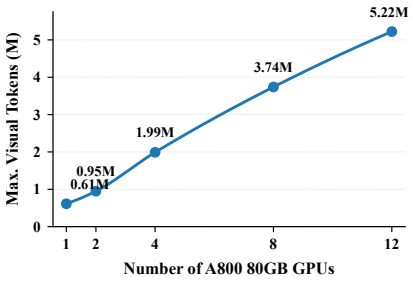

Figure 3: Maximum supported visual token length with increasing GPU resources.

Table 2: Full-sequence scalability with complete visual sequence retention. Memory and TTFT are profiled at a fixed 98K visual-token input.

Execution	GPUs	Max N_v ↑	Mem. ↓	TTFT ↓
Single Exact	1	491K	26.14	18.65
Sequence Exact	4	1769K	17.50	6.82
Propagation-only	4	1843K	16.73	6.39
All-Linear	4	**2212K**	**13.92**	**5.31**
SourceScale-VL	4	1990K	15.08	5.76

However, All-Linear degrades to 53.48% accuracy with substantially higher SCE, showing that query-conditioned retrieval still requires selective Exact computation. Under the same 25% Exact budget, RSR-Hybrid improves over KL-Hybrid by 0.82 points while reducing SCE, validating retrieval-sensitive allocation. After Source-Factored On-Policy Distillation, SourceScale-VL further reaches 57.46% accuracy and 0.041 SCE, closely approaching the Exact reference with efficient converted attention.

Full-sequence scalability. Table 2 shows that SourceScale-VL increases the maximum supported visual sequence from 1.769M to 1.990M tokens over sequence-parallel Exact attention while reducing memory and TTFT at a fixed 98K input. Figure **??** further shows context capacity scaling to 12 GPUs through ring-style propagation of compact states. All-Linear remains the most efficient but loses substantially more fidelity, motivating selective Exact retrieval.

4.3 Ablations and Analysis

RSR allocation quality. Table 4 compares selector scores with an Exact-restoration oracle, obtained by ranking retrieval atoms according to the SCE reduction from individually restoring them to Exact. RSR improves Spearman ρ from 0.44 to 0.63 and Recall@28 from 46.43% to 67.86% over KL, showing stronger agreement with capacity-critical retrieval atoms.

Table 3: Controlled conversion on the held-out in-domain split.

Method	Acc. ↑	KL ↓	SCE ↓
Teacher (Exact)	57.82	0.000	0.000
Propagation-only	57.61	0.014	0.030
All-Linear	53.48	0.084	0.166
AT-Linear	54.96	0.060	0.118
KL-Hybrid	56.21	0.045	0.091
RSR-Hybrid	57.03	0.037	0.063
SourceScale-VL	**57.46**	**0.025**	**0.041**

Table 4: Agreement with the Exact-restoration oracle over 112 retrieval atoms.

Selector	ρ ↑	Rc@28 ↑
Random	0.00	25.00
KL	0.44	46.43
RSR	**0.63**	**67.86**

Table 5: Recovery ablation with fixed RSR allocation.

Recovery	Acc. ↑	KL ↓	SCE ↓
RSR-Hybrid	57.03	0.037	0.063
+ Teacher KD	57.12	0.033	0.060
+ On-Policy KD	57.25	0.029	0.054
+ Source OPD	**57.46**	**0.025**	**0.041**

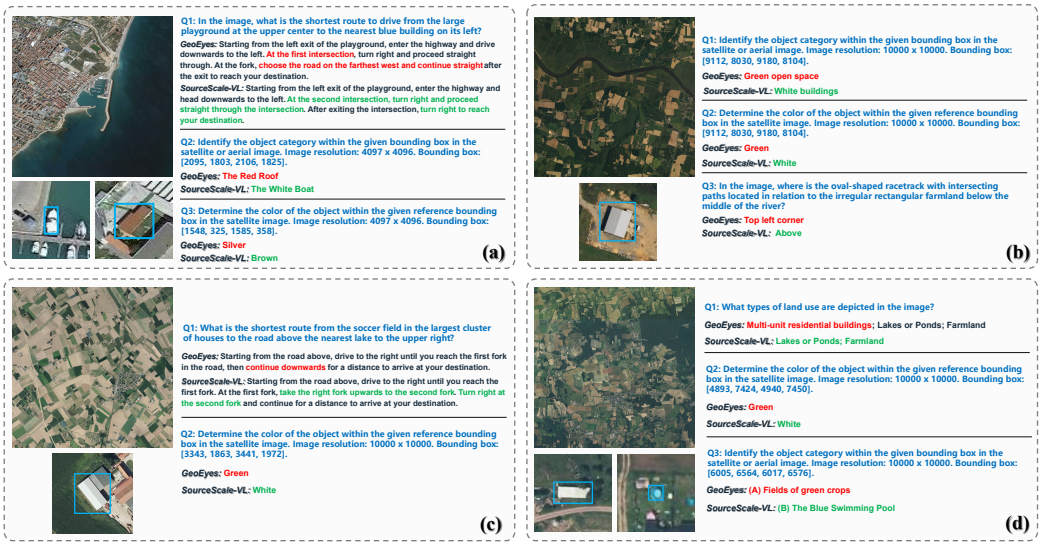

Figure 4: **Qualitative comparison on UHR RS vision-language tasks.** (a-d) Comparisons between GeoEyes and SourceScale-VL on object recognition, spatial grounding, and visual reasoning. SourceScale-VL produces more accurate responses by preserving fine-grained visual evidence.

Source-factored recovery. Table 5 studies the contribution of recovery objectives under fixed RSR allocation. Teacher-forced token distillation provides limited improvement, while on-policy token distillation increases accuracy from 57.12% to 57.25% by reducing inference-state mismatch. Source-Factored OPD further improves accuracy to 57.46% and reduces SCE from 0.054 to 0.041, showing that source-wise contribution alignment provides complementary supervision beyond token-level matching.

4.4 Qualitative Results

Figure 4 shows that SourceScale-VL preserves fine-grained evidence across recognition, attribute understanding, and spatial reasoning after attention conversion, maintaining accurate query–region correspondence.

5 Conclusion

We presented SourceScale-VL, a post-training conversion framework that scales UHR RS-VLMs without reducing the post-encoder visual sequence. By combining linear visual propagation, RSR-guided hybrid retrieval, and Source-Factored On-Policy Distillation, SourceScale-VL improves full-sequence scalability while preserving source-specific reasoning fidelity.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish the writing, including improving grammar, clarity, and readability. They were not used for research ideation or execution, literature retrieval, substantive content generation, synthetic data generation, or mathematical proofs. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stańczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:263610088.

Joshua Ainslie, James P. Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit K. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *ArXiv*, abs/2305.13245, 2023. URL https://api.semanticscholar.org/CorpusID:258833177.

Shuai Bai, Yuxuan Cai, Rui-Zhe Chen, Ke qin Chen, Xiong-Hui Chen, Zesen Cheng, Liang-Hao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-Sheng Li, Mei Li, Kaixin Li, Zi-Cheng Lin, Junyang Lin, Xue-Jing Liu, Jiawei Liu, Cheng-Long Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chen-Xu Lv, Rui Men, Lingchen Meng, Lingchen Meng, Xin yi Ren, Sibo Song, Yu chen Sun, Jun Tang, Jianhong Tu, Jian-Qiang Wan, Peng Wang, Pengfei Wang, Qiu-Yue Wang, Yuxuan Wang, Tian-Bao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhi-Bo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bo-Wen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025a. URL https://api.semanticscholar.org/CorpusID:283262018.

Shuai Bai, Ke qin Chen, Xue-Jing Liu, Jia-Lin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jian-Qiang Wan, Peng-Fei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tian-Bao Xie, Zesen Cheng, Hang Zhang, Zhi-Bo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025b. URL https://api.semanticscholar.org/CorpusID:276449796.

Yukang Chen, Wei Huang, Baifeng Shi, Qing-Hao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. *ArXiv*, abs/2507.07966, 2025. URL https://api.semanticscholar.org/CorpusID:280017269.

Yuhong Chou, Zehao Liu, Ruijie Zhu, Xinyi Wan, Tianjian Li, Congying Chu, Qian Liu, Ji-Bin Wu, and Zejun Ma. Zeco: Zero communication overhead sequence parallelism for linear attention. *ArXiv*, abs/2507.01004, 2025. URL https://api.semanticscholar.org/CorpusID:280146584.

Yun-Kai Dang, Minxi Dai, Yue-Kun Yang, Zhan Li, Wen-Bin Li, Feng Miao, and Yang Gao. Uhrbat: Budget-aware token compression vision-language model for ultra-high-resolution remote sensing. *ArXiv*, abs/2604.13565, 2026. URL https://api.semanticscholar.org/CorpusID:287479849.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. URL https://api.semanticscholar.org/CorpusID:10163399.

Daniel Goldstein, Eric Alcaide, Janna Lu, and Eugene Cheah. Radlads: Rapid attention distillation to linear attention decoders at scale. *ArXiv*, abs/2505.03005, 2025. URL https://api.semanticscholar.org/CorpusID:278338959.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL https://api.semanticscholar.org/CorpusID:7200347.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, She-An Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:220250819.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman H. Khan, and Fahad Shahbaz Khan. Geochat:grounded large vision-language model for remote sensing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, 2023. URL https://api.semanticscholar.org/CorpusID:265456719.

Yanhong Li, Song-Lin Yang, Shawn Tan, Mayank Mishra, Rameswar Panda, Jia-Wei Zhou, and Yoon Kim. Distilling to hybrid attention models via kl-guided layer selection. *ArXiv*, abs/2512.20569, 2025. URL https://api.semanticscholar.org/CorpusID:284133045.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *ArXiv*, abs/2310.01889, 2023. URL https://api.semanticscholar.org/CorpusID:263608461.

Yi Liu, Jing Zhang, Di Wang, Xiao-Yu Tian, Hao-Nan Guo, and Bo Du. Seeing clearly without training: Mitigating hallucinations in multimodal llms for remote sensing. *ArXiv*, abs/2603.02754, 2026. URL https://api.semanticscholar.org/CorpusID:286228574.

Junwei Luo, Yingying Zhang, Xue Yang, Kang Wu, Qi Zhu, Lei Liang, Jingdong Chen, and Yansheng Li. When large vision-language model meets large remote sensing imagery: Coarse-to-fine text-guided token pruning. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9206–9217, 2025. URL https://api.semanticscholar.org/CorpusID:276928275.

Weigao Sun, Zhen Qin, Dong Li, Xuyang Shen, Yu Qiao, and Yi-Ran Zhong. Linear attention sequence parallelism. *Trans. Mach. Learn. Res.*, 2025, 2024. URL https://api.semanticscholar.org/CorpusID:268875936.

Weigao Sun, Disen Lan, Yi-Ran Zhong, Xiaoye Qu, and Yu Cheng. Lasp-2: Rethinking sequence parallelism for linear attention and its hybrid. *ArXiv*, abs/2502.07563, 2025. URL https://api.semanticscholar.org/CorpusID:276259019.

Feng-Xiang Wang, Ming-Shuo Chen, Yue-Ying Li, Di Wang, Haotian Wang, Zonghao Guo, Zefan Wang, Boqi Shan, Long Lan, Yu-Lin Wang, Hong-Zhen Wang, Wenjing Yang, Bo Du, and Jing Zhang. Geollava-8k: Scaling remote-sensing multimodal large language models to 8k resolution. *ArXiv*, abs/2505.21375, 2025a. URL https://api.semanticscholar.org/CorpusID:278911260.

Feng-Xiang Wang, Hong-Zhen Wang, Zonghao Guo, Di Wang, Yu-Lin Wang, Ming-Shuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025b. URL https://api.semanticscholar.org/CorpusID:277452282.

Feng-Xiang Wang, Ming-Shuo Chen, Yue-Ying Li, Ya-Jie Yang, Yi-Fan Zhang, Long Lan, Xue Yang, Hongda Sun, Yu-Lin Wang, Di Wang, Jun Song, Jing Zhang, and Bo Du. Geoeyes: On-demand visual focusing for evidence-grounded understanding of ultra-high-resolution remote sensing imagery. *ArXiv*, abs/2602.14201, 2026. URL https://api.semanticscholar.org/CorpusID:285615993.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 13:297–324, 2025. URL https://api.semanticscholar.org/CorpusID:280374118.

Fuzhao Xue, Yukang Chen, Da-Cheng Li, Qing-Hao Hu, Ligeng Zhu, Xiuyu Li, Yun-Hao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Jim Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *ArXiv*, abs/2408.10188, 2024. URL https://api.semanticscholar.org/CorpusID:271903601.

Yangfan Zhan, Zhi-Tong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ArXiv*, abs/2401.09712, 2024. URL https://api.semanticscholar.org/CorpusID:267035181.

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Frederick Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. Lolcats: On low-rank linearizing of large language models. *ArXiv*, abs/2410.10254, 2024. URL https://api.semanticscholar.org/CorpusID:273345158.

Ziwei Zheng, Michael Yang, Jack Hong, Chen-Xiao Zhao, Guo-Hai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *ArXiv*, abs/2505.14362, 2025. URL https://api.semanticscholar.org/CorpusID:278769859.

A Appendix

You may include other additional sections here.