EVIVIT: EVIDENCE-ADAPTIVE VISION TRANSFORMERS FOR FINE-GRAINED PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained visual perception enables vision–language models to distinguish subtle attributes and ground their answers in visual evidence. In high-resolution scenes, processing the whole image at greater resolution spends visual tokens on irrelevant content, while isolated crops can lose the context needed to interpret the selected evidence. We introduce EviViT, a lightweight attachment that learns where a pretrained vision transformer should acquire detail. Human visual-search traces supervise a question-conditioned evidence density, which guides regional re-reading from the original pixels and the allocation of visual tokens. A sparse, coordinate-aware bridge then connects the regional features to the global scene, allowing the host to interpret precise evidence in context. Learned with the host backbone frozen, the attachment serves both the base model and compatible post-trained descendants without refitting. Experiments across nine hosts show consistent gains in average fine-grained accuracy. Matched-budget comparisons further show that EviViT outperforms global-only processing at every tested token ceiling while using fewer visual tokens.

1 INTRODUCTION

Visual understanding is fundamental to vision–language models (VLMs), providing perceptual grounding for question answering, multimodal reasoning, and decision making. In high-resolution scenes, an answer may hinge on a distant identifier, subtle attribute, short text, or spatial relation in a small image region. Resizing the image to a fixed visual-token budget can erase such evidence, whereas uniformly increasing resolution spends more computation on regions irrelevant to the question. Fine-grained benchmarks such as V*Bench and HR-Bench expose this tension (Wu & Xie, 2024; Wang et al., 2025). The question is thus not simply how many visual tokens a model receives, but where it spends them.

Existing approaches take two main directions. The first acquires detailed views through image tiling (Wang et al., 2025) or active search: methods such as V* and Mini-o3 repeatedly locate, crop, and inspect regions (Wu & Xie, 2024; Lai et al., 2026). This adaptive crop-then-forward loop asks the language model to choose a region, view another image, and reason again, adding inference overhead. It also ties evidence acquisition to the answering policy. The second direction reduces computation within an already encoded image through query-aware pruning, token merging, or complementary local–global retention (Li et al., 2026; Tong et al., 2026). These methods can remove redundant computation, yet selecting encoded tokens cannot recover detail lost during initial resizing. Together, these limitations motivate a different question: can the vision encoder itself learn which original-image regions need additional resolution before answer generation?

Human visual search provides supervision for this capability. Cursor movements, inspected regions, zooms, abandoned branches, and final evidence reveal how attention converges on an answer. Rather than using human search as an action sequence to imitate or reducing it to a final evidence box, we distill the full exploration process into a question-conditioned evidence density that directly supervises where the vision transformer should allocate resolution. This turns human search from behavioral supervision into visual-allocation supervision.

Doing so poses three challenges. First, the encoder must learn question-relevant locations from varied human trajectories, rather than visual salience alone. Second, it must select complementary views within a limited token budget: repeated zooms waste capacity, while a tiny isolated crop can lose

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.

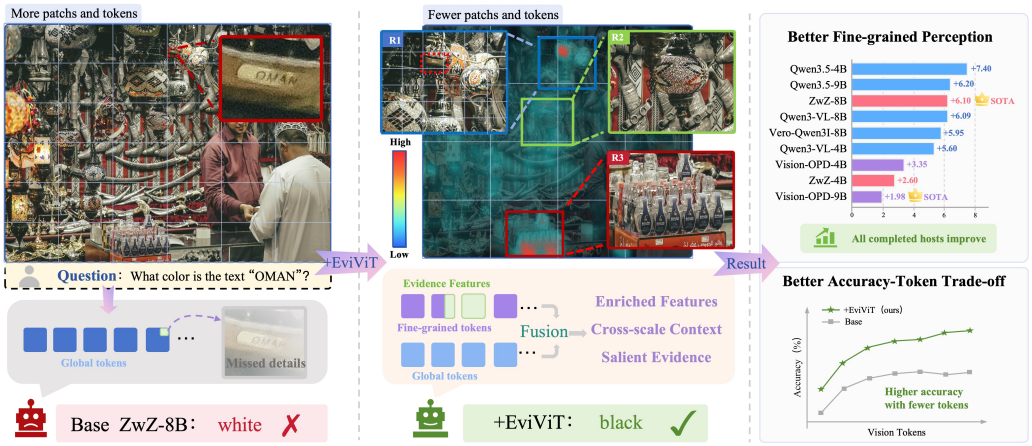

Figure 1: **Question-conditioned detail allocation.** A global-only host misses the small queried text despite processing more patches. EviViT predicts an evidence density learned from human search annotations, re-reads selected regions at native detail, and fuses them with the global scene. Across compatible hosts, this improves fine-grained accuracy using fewer tokens under matched ceilings.

necessary context. Third, it must fuse newly acquired detail with the original scene. Isolated crops can detach a clue from its spatial meaning; replacing the global view instead discards scene-level information.

We address them with EviViT, an intention-driven, evidence-adaptive ViT paradigm that learns where to spend visual resolution for the current question. As Figure 1 illustrates, a lightweight attachment augments a pretrained ViT while the host vision backbone and language model remain frozen. The Prompt-conditioned Token–Evidence Aligner (PTEA) first aligns question representations with intermediate ViT features to predict a dense evidence distribution supervised by human search traces. An evidence-aware planner then selects one or two non-redundant decisive regions and a complementary context region, dynamically allocates the local-token budget, and re-reads these regions from the original pixels at native detail. Unlike enlarging features that have already lost detail, this step acquires new visual evidence. Finally, a coordinate-aware Sparse Bridge establishes correspondence between regional and global tokens and exchanges information around spatially matched locations. This lets the language model interpret readable local evidence within the scene, without iterative crop commands.

Separating evidence acquisition from answering also makes the attachment reusable. Trained on 1K+ VisualProbe human-search records with the host ViT, merger, and LLM frozen, EviViT can serve compatible foundation models and their post-trained descendants without refitting the evidence allocator. Inference requires only the original image and question, not a human trace, evidence box, or explicit crop instruction.

The results support this design across model scales, post-training recipes, and visual-token budgets. EviViT improves average fine-grained accuracy for all nine foundation and post-trained hosts; on Qwen3-VL-4B/8B, the seven-benchmark average rises by 5.60/6.09 points. Under matched 1K–20K token ceilings, it consistently outperforms global-only processing while using only 90–97% as many visual tokens. Gains persist on stronger post-trained models and after language-side LoRA adaptation (Hu et al., 2021), indicating that learned evidence allocation is reusable beyond one answering policy.

Our contributions are summarized as follows:

- **Human-search-supervised evidence allocation.** We use the full search trajectory, not just its terminal box, to supervise question-conditioned resolution allocation inside a pretrained ViT.
- **An evidence-adaptive ViT paradigm.** EviViT combines evidence prediction, non-redundant native-detail acquisition, adaptive token allocation, and coordinate-aware sparse fusion within a lightweight attachment.

- **Reusable and efficient fine-grained perception.** Across 4B–9B foundation and post-trained hosts, EviViT improves fine-grained perception and accuracy–token trade-offs while remaining compatible with continued language-side adaptation.

2 Related Work

Vision Encoders. Vision Transformers represent images as patch sequences and learn their interactions through self-attention (Dosovitskiy, 2020). CLIP and SigLIP learn transferable visual representations from image–text pairs using contrastive and pairwise sigmoid objectives, respectively (Radford et al., 2021; Zhai et al., 2023). Dynamic-resolution encoders in the Qwen-VL family accommodate different image sizes with variable-length token sequences (Wang et al., 2024; Bai et al., 2025). This makes more detail available, but increasing the resolution of the entire image spends tokens on both decisive evidence and irrelevant background. EviViT instead learns where to re-read source pixels for a given question and fuses the resulting regional tokens with the global grid inside the frozen encoder.

Visual Perception. High-resolution perception methods recover details through tiling, selective cropping, or interactive search. V* uses language-guided search to locate small targets (Wu & Xie, 2024); FOCUS derives crop relevance from internal MLLM representations (Zhong et al., 2026). Mini-o3 scales multi-turn visual exploration, while AdaptVision learns when to acquire additional crop tokens through reinforcement learning (Lai et al., 2026; Lin et al., 2026b). A different line transfers detailed observations into the model during training: ZwZ distills region-grounded supervision into full-image inference, and Vision-OPD uses a crop-conditioned teacher for on-policy self-distillation (Wei et al., 2026; Yuan et al., 2026). Tool-using policies must decide when and where to inspect before incorporating the returned views; distillation removes this interaction by updating the answering model. EviViT instead learns a visual attachment while keeping that model frozen. It acquires regional detail before answer generation, without crop commands from the language model, and retains the global view alongside the local observations. This separates evidence acquisition from the answering policy, allowing the same attachment to complement compatible foundation and post-trained hosts.

Adaptive Attention. Adaptive computation changes which visual information is sampled or retained. DynamicViT prunes tokens using learned importance scores, whereas Token Merging combines similar tokens (Rao et al., 2021; Bolya et al., 2022). Deformable attention instead learns data-dependent sampling positions (Xia et al., 2022). For VLMs, OccamToken adds query-aware pruning, and Focus-Scan-Refine combines instruction-relevant local evidence with complementary global context (Li et al., 2026; Tong et al., 2026). Token compression reduces redundancy, but it can only select or combine features already available at the input resolution. EviViT also decides which detail to acquire: its predicted density allocates fresh tokens from the original pixels while preserving a global context stream. Human-attention and gaze datasets provide another source of spatial guidance by recording where people look while answering questions (Das et al., 2017; Sood et al., 2021; Chen et al., 2021). EviViT uses cursor movements, zoom windows, and final-evidence annotations to supervise a spatial evidence predictor. The human record therefore guides the allocation of resolution, rather than prescribing a sequence of actions for the answering model to imitate.

3 Method

EviViT separates two decisions that global resizing treats together: where an image needs more detail, and how that detail should be interpreted in context. It first predicts a question-conditioned evidence density and uses it to allocate regional resolution. It then connects the re-read observations with the global visual grid, preserving access to the surrounding scene. Figure 2 shows how both components fit inside the frozen host vision transformer, before answer generation.

3.1 Human visual-search data

Research-team members used our interface to collect human visual-search traces for 1K+ training questions from VisualProbe (Lai et al., 2026). They inspected the full-resolution image, zoomed

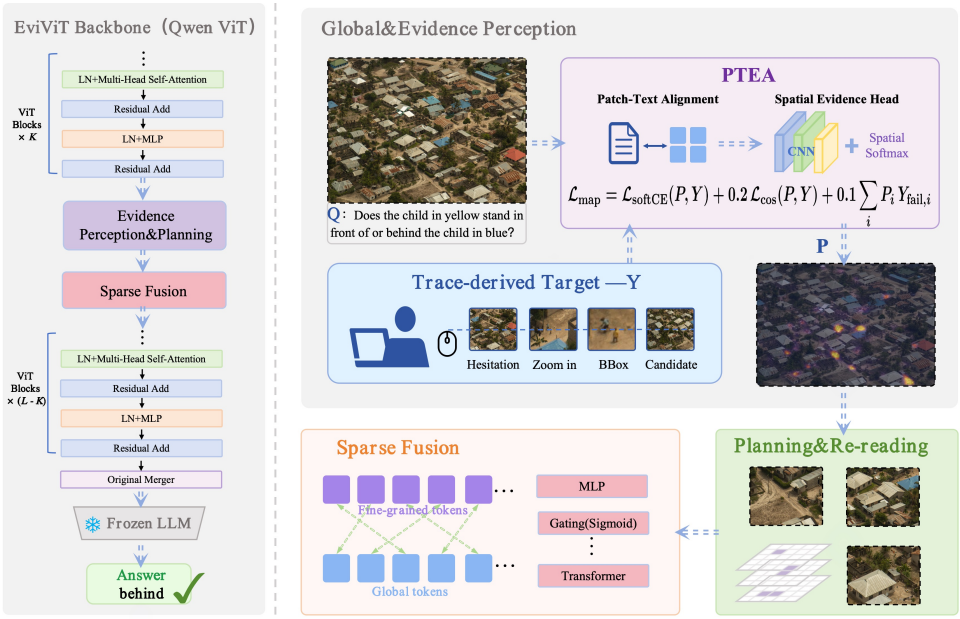

Figure 2: **EviViT pipeline.** Human interactions are distilled into a question-conditioned evidence target. PTEA predicts this distribution inside the host ViT and yields one or two non-redundant decisive regions plus one context region. These regions are re-read from the original pixels, fused with coordinate-matched global tokens, and processed by the remaining frozen ViT blocks and LLM.

into regions, returned to broader views, and marked the final evidence. The interface logs cursor positions, timestamps, zoom windows, resets, and final boxes in source-image coordinates, aligning observations across scales. Event intervals provide dwell cues for the spatial targets. These records capture both the selected evidence and the broader search that led to it. They supervise evidence and context prediction during training; inference takes only an image and question.

3.2 Learning a Spatial Evidence Density

From search events to evidence density. Let (I, q) be a high-resolution image and question. We render point events with Gaussian kernels and boxes as soft regional mass in source-image coordinates. The maps Y_{all}, Y_{succ}, Y_{last}, and Y_{final} describe all visited regions, the branch after the last reset, the last completed zoom, and the final box, respectively. Their mixture is

$$Y = \text{PeakNorm}(0.25Y_{\text{all}} + 0.35Y_{\text{succ}} + 0.25Y_{\text{last}} + 0.15Y_{\text{final}})\,. \tag{1}$$

Peak normalization rescales the maximum to one. Combining the precise final box with the surrounding inspected views supervises both evidence and context. Abandoned regions form a separate weak negative map Y_{fail}.

At inference, the Prompt-conditioned Token–Evidence Aligner (PTEA) predicts this target without a search trace. Let $V = \{v_i\}_{i=1}^{N_g}$ be the global grid at an intermediate visual block and $T = \{t_j\}_{j=1}^{L}$ the question tokens. PTEA projects both into a shared space, making the prediction depend on the question rather than visual salience alone. A text Transformer contextualizes the projected question tokens into $\hat{t}_j$, which are aligned with each patch:

$$s_{ij} = \tau \cos(W_v v_i, \hat{t}_j), \quad a_{ij} = \text{softmax}_j(s_{ij}), \quad c_i = \sum_j a_{ij}\hat{t}_j. \tag{2}$$

The aligned features, their interactions, and spatial coordinates feed a local convolutional head, whose spatial softmax yields P. We resize Y to this grid and normalize its mass to one before fitting PTEA with

$$\mathcal{L}_{\text{map}} = \mathcal{L}_{\text{softCE}}(P,Y) + 0.2\mathcal{L}_{\cos}(P,Y) + 0.1\sum_i P_i Y_{\text{fail},i}. \tag{3}$$

Soft cross-entropy and cosine loss align evidence mass and map shape; the final term discourages concentration on abandoned branches. Only PTEA is updated in this stage.

A complementary branch predicts a context distribution P_{ctx} from the same aligned features. Supervision from the broader inspected regions helps it identify surrounding information beyond the focal evidence. We fit this branch after PTEA while keeping the shared alignment and evidence branch fixed.

3.3 Non-redundant native-detail allocation

Selecting complementary regions. Given the predicted evidence density P, the planner selects complementary regions, allocates their token budgets, and then refines their spatial extent. Nearby peaks in P may correspond to the same evidence, so residual suppression retains up to two distinct decisive regions expected to contain answer-relevant evidence. Masking them in the context distribution P_{ctx} yields one complementary context region, giving two or three views of both focal evidence and its surroundings.

Balancing detail and context. ContextNeed is a lightweight budget predictor. From statistics of P and P_{ctx}, including concentration and residual context mass, it reserves 10–25% of local visual tokens for the context region. The remainder is divided among decisive regions according to their evidence mass, balancing contextual coverage with resolution for answer-relevant details.

Refining regions and setting resolution. A bounded box-refinement head, H-Safe, adjusts each selected region's center and size using its visual feature, query representation, and box geometry. The refined regions and allocation priorities become token grids for native re-reading alongside the retained global view; Appendix A.2 details the capacity and grid-alignment rules.

3.4 Native re-reading and sparse fusion

With boxes and token grids fixed, EviViT samples each region from the original image at its allocated resolution and encodes it to the insertion block using the same frozen ViT. This recovers detail lost in global resizing while retaining the global path for scene topology.

To connect a readable detail to its place in the scene, the bridge uses source coordinates to establish correspondence between the two grids. Each local token ℓ_i attends to the 3×3 neighborhood $\mathcal{N}(p(i))$ of its coordinate-matched global parent:

$$\tilde{\ell}_i = \ell_i + W_o \sum_{j\in\mathcal{N}(p(i))} \text{softmax}_j\left(\frac{(W_q\ell_i)^\top(W_k g_j)}{\sqrt{d}} + \phi(\Delta x, \Delta y, \Delta s)\right) W_v g_j. \tag{4}$$

The reverse update aggregates local messages by parent index and writes only to global tokens with local children. This exchange lets a local feature access its surroundings and lets the global grid incorporate a more detailed observation of the same location. Zero-initialized output projections and bounded residuals make the initial bridge an exact identity. Global and local tokens continue through the remaining frozen ViT blocks and original patch merger, then enter the LLM with rotary positions derived from source coordinates and scale.

3.5 Training and portability

Training follows the route from selecting evidence to using it. Search maps first fit PTEA, followed by its complementary context branch. Trace-derived context targets and box geometry then fit ContextNeed and H-Safe. With routing fixed, reference-answer cross-entropy and an identity penalty train the Sparse Bridge to use the acquired detail. The host ViT, merger, and LLM remain frozen throughout, so answer loss does not rewrite the spatial cue learned by PTEA.

We insert separate attachments into Qwen3-VL-4B and 8B after visual Blocks 16 and 18, respectively. The frozen-host fit lets each transfer to post-trained descendants that share its visual blocks, channel width, merger, and coordinate system, without refitting. Architectural sizes, box losses, and optimization details are given in Appendix A.

Table 1: **Fine-grained perception accuracy (%).** Top-three paired gains per benchmark are shaded orange dark – light ; top-three mean gains are shaded green dark – light .

Model	VP-E	VP-M	VP-H	V*	HR-4K	HR-8K	Zoom	Avg.	
Author-reported references; ‡ cells are matched local API re-evaluations									
GPT-4o†	47.50	15.40	11.20	71.20‡	67.12‡	62.50‡	47.10‡	46.00‡	
Gemini-3-Flash†	67.38	50.75	47.17	84.82	89.25	85.50	59.41‡	69.18‡	
GPT-5.2†	57.45‡	38.43‡	38.68‡	79.06	81.12	78.38	50.89	60.57‡	
GPT-5.4†	61.70‡	29.85‡	20.75‡	76.96	84.00	77.88	52.66	57.69‡	
Gemini-3.1-Pro†	58.16‡	41.42‡	41.51‡	87.96	89.63	86.88	61.18	66.68‡	
Gemini-3.5-Flash†	63.83‡	51.87‡	48.11‡	89.01	89.12	86.62	61.42	70.00‡	
Locally evaluated open checkpoints and paired EviViT attachments									
Qwen3-VL-4B	60.28	40.30	41.51	84.29	76.25	73.38	45.80	60.26	
+ EviViT	**63.83**	**50.37**	**48.11**	**87.96**	**79.62**	**77.75**	**53.37**	**65.86**	(+5.60)
Qwen3-VL-8B	67.38	42.16	44.34	85.86	76.00	72.75	43.91	61.77	
+ EviViT	**69.50**	**53.36**	**51.89**	**91.10**	**80.25**	**78.25**	**50.65**	**67.86**	(+6.09)
Qwen3.5-4B	63.83	44.40	45.28	81.68	79.00	76.38	49.11	62.81	
+ EviViT	**70.92**	**53.36**	**55.66**	**87.96**	**83.75**	**81.25**	**58.58**	**70.21**	(+7.40)
Qwen3.5-9B	65.96	44.40	42.45	84.82	76.62	73.38	51.01	62.66	
+ EviViT	**66.67**	**54.85**	**51.89**	**87.96**	**82.50**	**81.12**	**57.04**	**68.86**	(+6.20)
ZwZ-4B	**69.50**	47.76	34.91	90.05	78.12	76.62	55.62	64.66	
+ EviViT	66.67	**52.61**	**46.23**	**91.10**	**78.88**	**77.25**	**58.11**	**67.26**	(+2.60)
ZwZ-8B	71.63	46.64	46.23	89.53	80.50	77.12	55.98	66.80	
+ EviViT	**79.43**	**55.97**	**62.26**	**92.67**	**81.12**	**78.75**	**60.12**	**72.90**	(+6.10)
Vero-Qwen3I-8B	66.67	43.66	54.72	86.39	81.50	77.50	46.27	65.24	
+ EviViT	**72.34**	**55.22**	**61.32**	**87.43**	**83.12**	**83.88**	**55.03**	**71.19**	(+5.95)
Vision-OPD-4B	73.76	49.63	51.89	86.39	81.00	76.75	59.05	68.35	
+ EviViT	**80.85**	**52.24**	**53.77**	**89.53**	**82.62**	**80.75**	**62.13**	**71.70**	(+3.35)
Vision-OPD-9B	**74.47**	53.73	47.17	**93.19**	**80.88**	78.62	63.79	70.26	
+ EviViT	73.05	**58.58**	**54.72**	90.05	80.62	**82.62**	**66.04**	**72.24**	(+1.98)

†Author-reported cells: GPT-4o VisualProbe from Mini-o3 (Lai et al., 2026); Gemini-3-Flash values from PixelEyes (Gong et al., 2026); GPT-5.2/5.4 and Gemini-3.1-Pro/3.5-Flash values from Vision-OPD (Yuan et al., 2026). ‡Our single-call API evaluations use the local inputs, answer-only prompt, deterministic decoding, and Qwen3-VL-8B judge. Mixed-source averages are descriptive. Zoom uses the full image.

4 Experiments

Evaluation protocol. We evaluate fine-grained perception on VisualProbe Easy/Medium/Hard (515 questions) (Lai et al., 2026), V*Bench (Wu & Xie, 2024), HR-Bench 4K/8K (Wang et al., 2025), and the full-image view of ZoomBench (Wei et al., 2026). MMBench (Liu et al., 2024), MMStar (Chen et al., 2024), and the ten tasks of Visual-CoT (Shao et al., 2024) measure broader capability.

Each local host/host+EviViT comparison fixes the image–question inputs, language checkpoint, prompt, and deterministic decoding. A frozen Qwen3-VL-8B judge compares the candidate answer with the reference, given the question but no image or method identity. Published results and API supplements are marked separately in the tables; Appendix A gives the evaluation details.

4.1 Fine-grained perception and host transfer

Table 1 shows that attaching EviViT improves the seven-benchmark average for all nine foundation and post-trained hosts. The clearest benefits appear on VisualProbe Medium and Hard, which improve by 8.2 and 8.6 points on average, while the other five benchmarks also improve on average. These leading gains span several host families, rather than a single model. *The largest gains appear on the harder tiers, consistent with EviViT addressing evidence that is difficult to preserve in the global view.*

From evidence to a correct answer. The cases in Appendix D show how these gains arise when local detail is interpreted within its scene. EviViT reads a distant car identifier and resolves a ball's position relative to the long bench. On Vero-8B (Sarch et al., 2026), the base model considers the tower, facade, and logo before answering incorrectly; with EviViT, a shorter, correct rationale focuses on the signboard's straight edges and right angles. Better evidence thus helps keep reasoning on the queried object instead of drifting toward unrelated scene elements.

Transfer to strong open models. These benefits persist when the host has already undergone post-training. Among the locally evaluated open checkpoints from recent systems (He et al., 2026; Yuan et al., 2026; Bi et al., 2026; Gong et al., 2026) in Figure 3, ZwZ-8B+EviViT achieves the highest VisualProbe VP-All (63.69), while Vision-OPD-9B+EviViT leads ZoomBench (66.04). Both also improve over their own hosts in Table 1. This suggests that better question-relevant evidence complements answer-policy post-training: stronger reasoners still benefit from more focused perception. The foundation-family attachment provides it without refitting for each compatible descendant.

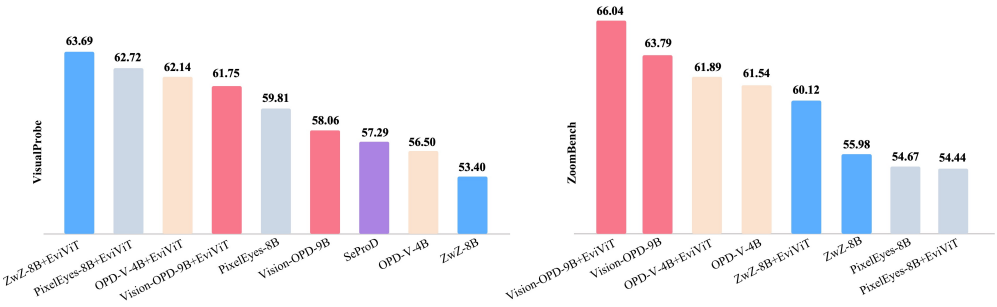

Figure 3: **Strong-model accuracy rankings.** VisualProbe VP-All (515 questions) and ZoomBench (845 questions), sorted within each panel. Open-checkpoint bars use our local evaluation. The SeProD VP-All bar is derived from its author-reported Easy/Medium/Hard scores under the original avg@32 protocol, so it is contextual rather than a matched comparison.

4.2 Accuracy–efficiency trade-off

Better accuracy with fewer visual tokens. To determine whether the gains require more visual input, we compare seven token ceilings from 1K to 20K on VisualProbe and V*Bench. Table 2 and Figure 4 show higher average accuracy at every ceiling with only 90–97% of the paired global-only tokens. At the 8B 4K setting, EviViT reaches 68.87 accuracy with 3.62K tokens, compared with 62.83 using 3.83K for the global-only model. It even exceeds the global-only 20K setting, which reaches 68.27 with 13.31K realized tokens. *Where visual tokens are spent matters as much as how many are available.* Global scaling samples the entire image more densely, whereas EviViT directs part of the budget toward predicted evidence while retaining the broader scene. This targeted allocation continues to help after the global-only gains begin to saturate.

Table 2: **Accuracy–budget scaling.** Mean VP/V* accuracy (515/191); G/E: realized tokens.

	Qwen3-VL-4B				Qwen3-VL-8B			
Budget	Global	EviViT	Δ	Tok. G/E	Global	EviViT	Δ	Tok. G/E
1K	50.30	**56.35**	+6.05	1.00/0.90K	50.14	**57.65**	+7.51	1.00/0.91K
2K	54.93	**60.11**	+5.18	1.99/1.88K	57.47	**64.64**	+7.17	1.99/1.89K
4K	60.69	**65.45**	+4.76	3.83/3.62K	62.83	**68.87**	+6.04	3.83/3.62K
8K	63.89	**66.39**	+2.49	6.65/6.41K	65.84	**72.24**	+6.40	6.65/6.43K
12K	65.06	**67.46**	+2.40	9.32/8.95K	67.69	**72.53**	+4.84	9.32/9.00K
16K	65.16	**69.30**	+4.15	11.55/11.15K	67.69	**74.18**	+6.49	11.55/11.19K
20K	65.54	**69.11**	+3.56	13.31/12.75K	68.27	**74.38**	+6.10	13.31/12.83K

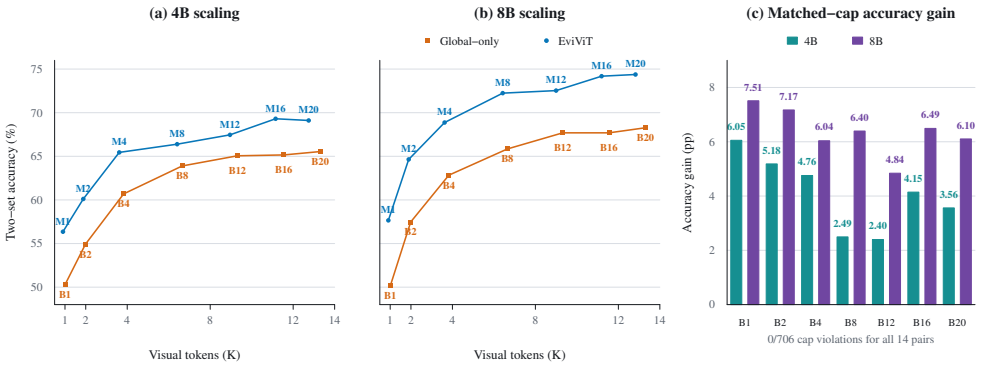

Figure 4: **Budget curves and paired gains.** Accuracy versus tokens on Qwen3-VL-4B/8B.

Table 3: **Whole-host-isolated efficiency.** Each budget pairs Global and EviViT. Tokens and accuracy use the complete sweep; latency and peak memory use 64 isolated requests per setting.

			TTFT (s)		E2E (s)			
Budget	Path	Vis. tok. (K)	Median	P90	Median	P90	Peak (GiB)	Δ Acc. (pp)
Qwen3-VL-4B								
1K	Global	1.00	0.206	0.218	0.557	0.619	8.62	–
	EviViT	0.90	0.384	0.403	0.738	0.804	8.66	+6.05
4K	Global	3.83	0.559	0.643	0.921	1.038	9.48	–
	EviViT	3.62	0.621	0.712	0.999	1.091	9.47	+4.76
16K	Global	11.55	0.541	0.637	0.902	1.021	9.97	–
	EviViT	11.15	0.643	0.739	1.082	1.192	9.96	+4.15
Qwen3-VL-8B								
1K	Global	1.00	0.254	0.264	0.696	0.706	16.72	–
	EviViT	0.91	0.465	0.477	0.892	0.904	16.76	+7.51
4K	Global	3.83	0.742	0.887	1.087	1.231	17.66	–
	EviViT	3.62	0.759	0.853	1.105	1.200	17.71	+6.04
16K	Global	11.55	0.740	0.887	1.088	1.233	18.25	–
	EviViT	11.19	0.745	0.841	1.091	1.187	18.17	+6.49

Runtime cost. Regional re-reading and fusion introduce overhead, but on 8B at 4K/16K, Table 3 shows 6.04/6.49-point gains for only 18/3 ms more median end-to-end time and nearly unchanged peak memory. The overhead is more visible at 1K and on 4B. Unlike multi-turn visual search (Lai et al., 2026), EviViT pays this overhead within one visual encoding process, without repeated language-model crop calls.

4.3 Preserving General Capabilities

EviViT is designed to sharpen local evidence while retaining the global information needed for broader visual understanding. Alongside the fine-grained gains in Table 1, Table 4 shows broadly stable MMBench, MMStar, and Visual-CoT results on four foundation hosts; Qwen3.5-4B improves on all three. This stability supports the global–regional design: selected evidence can be sharpened without discarding the broader scene representation. Appendix B extends the comparison to all nine hosts and individual tasks.

4.4 Continued Trainability

Beyond attaching EviViT after host training, we test whether the host can keep learning with it already in the visual pathway.

Table 4: **General-capability compatibility** on four foundation hosts.

	MMBench			MMStar			Visual-CoT		
Host	Base	+EviViT	Δ	Base	+EviViT	Δ	Base	+EviViT	Δ
Qwen3-VL-4B	87.36	88.43	+1.06	61.73	62.80	+1.07	77.53	75.52	-2.01
Qwen3-VL-8B	88.59	88.94	+0.35	65.07	65.27	+0.20	77.92	78.35	+0.43
Qwen3.5-4B	86.37	87.76	+1.39	63.20	64.40	+1.20	77.18	78.17	+1.00
Qwen3.5-9B	87.13	86.76	-0.37	64.33	65.20	+0.87	78.98	80.34	+1.36

Table 10 in Appendix A.5 compares Base+SFT and EviViT+SFT on the same 10K examples and language-LoRA schedule, with all visual weights frozen. Both improve on Visual-CoT, and EviViT's five-metric mean also rises. EviViT+SFT retains a 6.80-point VP-All and 7.34-point ZoomBench lead over Base+SFT. Thus, the learned evidence allocation remains useful as the language-side answering model continues to adapt.

4.5 Ablation Study

We first test query alignment by shuffling only the question seen by PTEA. Despite the unchanged host question and visual budget, VP-All falls by 4.47 and 3.31 points on the two large hosts in Table 7 of Appendix A.3. The added regional detail is therefore most useful when selected for the actual query, not simply because it supplies more pixels.

We next isolate location selection from learned fusion by passing the views through the host's native multi-image interface and moving only their centers. In Table 8 of Appendix A.3, pooled accuracy falls from 66.57 to 57.08 or 56.94 under random translation or cross-image shuffling. With view count, size, and budget fixed, this drop shows that the gain comes from acquiring detail at the right locations rather than merely adding regional views.

With query-matched locations established as useful, we examine how human exploration teaches the model to find them. Two complete attachments are trained from scratch on the same frozen Qwen3-VL-8B host and recipe, using either full-trace or final-box spatial supervision. Table 5 shows full-trace ahead on all three metrics, especially VP-Hard (56.60 versus 52.83). *Human search traces provide supervision beyond the final evidence box.* The box marks where search ended, whereas visited regions, the final branch, and the last zoom describe how attention narrows toward the evidence and its surrounding context. The gain therefore supports the value of supervision contained in the search process itself, beyond terminal localization labels. Appendix C.4 details the shared protocol.

Table 5: **Full-trace versus final-box supervision** at epoch three (%).

Supervision	VP-All	VP-Hard	V^*
Final-box	56.12	52.83	89.01
Full-trace	**57.28**	**56.60**	**90.05**
Δ	+1.17	+3.77	+1.05

5 Conclusion and Limitations

We introduce EviViT, a lightweight visual attachment that learns where to acquire detail from human search annotations and connects regional observations with the global scene. Across nine hosts it improves average fine-grained accuracy, using visual tokens more effectively at every tested budget. Its transfer to compatible post-trained hosts and continued utility during language adaptation make evidence allocation a reusable capability.

Limitation & Future Work. EviViT currently learns from 1K+ annotated searches. Scaling these data to more objects and scenes could extend its reach to small-object and embodied perception, where focused detail must remain grounded in the wider scene.

AI USE STATEMENT

Generative AI tools were used to refine the wording of an author-written manuscript and to assist with limited coding tasks related to training and inference. All AI-assisted text and code were reviewed and verified by the authors. The authors take full responsibility for the final manuscript and implementation.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. Our experiments use public visual benchmarks and visual-search annotations produced by members of the research team. The supplementary materials contain anonymized code and aggregate results, without raw interaction records or identifying information.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the architecture, training procedures, hyperparameters, and evaluation protocols. The anonymized supplementary code provides data-preparation, training, evaluation, and judging scripts, together with reported result summaries and training curves.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, Tat-Seng Chua, et al. Opd-v: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dengxian Gong, Yuanzheng Wu, Haobo Yuan, Zhengdong Hu, Tao Zhang, Yikang Zhou, Shihao Chen, Quanzhu Niu, Kai Wang, Jason Li, et al. Pixeleyes: Decoupling perception and reasoning for pinpoint visual evidence seeking. *arXiv preprint arXiv:2607.00115*, 2026.

Zhendong He, Qiyuan Dai, Guanbin Li, Liang Lin, and Sibei Yang. Self-prophetic decoding to unlock visual search in lvlms. *arXiv preprint arXiv:2605.28741*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

Geng Li, Guohao Chen, Ting Chen, Shilin Shan, Kuangji Zuo, Bofan Lyu, Tuo An, Gen Li, and Jianfei Yang. Occamtoken: Efficient vlm inference with training-free and budget-adaptive token pruning. *arXiv preprint arXiv:2605.29657*, 2026.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026a.

Zichuan Lin, Yicheng Liu, Yang Yang, Lvfang Tao, and Deheng Ye. Adaptvision: efficient vision-language models via adaptive visual acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11923–11932, 2026b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. *arXiv preprint arXiv:1902.09630*, 2019.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 27–43, 2021.

Enwei Tong, Yuanchao Bai, Yao Zhu, Junjun Jiang, and Xianming Liu. Focus-scan-refine: From human visual perception to efficient visual token pruning. *arXiv preprint arXiv:2602.05809*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4784–4793. IEEE, 2022.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. *Advances in Neural Information Processing Systems*, 38:96498–96535, 2026.

A Reproducibility Details

A.1 Model and training instantiation

Frozen host. Qwen3-VL-4B has 24 visual blocks of width 1,024 and receives EviViT after Block 16. Qwen3-VL-8B has 27 blocks of width 1,152 and receives it after Block 18. The two scales use the same algorithm but separate PTEA, H-Safe, and Bridge weights. The vision backbone, merger, and LLM remain frozen during all four stages. For Qwen3.5-4B and Qwen3.5-9B, we fit separate attachments on the frozen hosts after visual Blocks 16 and 18, respectively; each attachment is reused on the corresponding Vision-OPD descendant without refitting.

A.2 Attachment architecture

PTEA. Frozen language embeddings are reduced to 512 dimensions by a fixed random projection and then to $d = 192$ by a learned projection. A one-layer, four-head text Transformer (48 dimensions per head, $192 \to 384 \to 192$ feed-forward network, pre-normalization, and dropout) contextualizes the question. Each visual position concatenates its projected feature, aligned text context, elementwise product and absolute difference, maximum similarity, attention entropy, and a learned coordinate embedding. Two 3×3 convolutions and a 1×1 head produce the evidence logits. The 4B evidence branch contains approximately 1.41M parameters. A complementary branch shares the visual–text alignment and has its own spatial residual block and prediction head for P_{ctx}. Its normalized target combines a map of dwell and zoom events across the full trace with a map of inspections before the last reset in a 0.70/0.30 mixture; when the latter is empty, it uses the full-trace exploration map alone. The context branch is initialized from PTEA and fitted with the shared alignment and evidence branch frozen.

Region decoder and H-Safe. Residual suppression proposes two decisive anchors but discards the second when it is redundant with the first. The context map is masked at every retained decisive anchor and supplies one context region. The deployed model therefore uses two or three regions under a three-slot cap. For anchor $b = (c_x, c_y, w, h)$, H-Safe predicts $\delta = (\delta_x, \delta_y, \delta_w, \delta_h)$ with an LN–256–128–4 residual MLP:

$$\begin{aligned} c'_x &= c_x + 1.5w \tanh(\delta_x), & c'_y &= c_y + 1.5h \tanh(\delta_y), \\ w' &= \mathrm{clip}(w \exp(1.2 \tanh(\delta_w)), .02, .95)\,, & h' &= \mathrm{clip}(h \exp(1.2 \tanh(\delta_h)), .02, .95)\,. \end{aligned} \tag{5}$$

The zero-initialized output makes the initial transform an identity. Smooth-L_1, GIoU (Rezatofighi et al., 2019), target-coverage, excess-area, and identity losses fit the bounded correction. ContextNeed is a 129-parameter MLP with six inputs and a 16-unit hidden layer. Its inputs are the secondary decisive-region mass, the normalized entropy of each map, their Jensen–Shannon divergence, context mass outside the decisive regions, and peak evidence probability. It predicts a target context share between 0.10 and 0.25, leaving the remainder for decisive views. These shares set local allocation priorities; the final token counts also depend on region size, the available capacity, and patch-grid alignment.

Sparse Bridge. The 4B bridge has one block, internal width 256, four attention heads, and approximately 1.91M parameters. Relative offset and log scale pass through a two-layer projection to form the attention bias. Output projections are zero-initialized, and each residual update is clipped to 20% of the receiving token's RMS. Reverse messages are aggregated by coordinate parent and written only to global tokens that have local children.

Four-stage fit. First, frozen ViT features are cached for the 1K+ training examples at candidate insertion blocks. Five-fold cross-validation using only the trace targets selects the insertion block and PTEA training duration before external QA evaluation. Second, PTEA is optimized for three complete passes over the cached features using Equation 3, after which the complementary context branch is fitted with PTEA frozen. Third, ContextNeed and H-Safe are fitted from the trace-derived context-share target and last-zoom/final-box geometry. The H-Safe output layer starts at zero, so the first model is exactly the evidence-anchor policy. Fourth, with routing fixed, the Sparse Bridge is trained with reference-answer token cross-entropy and a 0.01 identity penalty. We evaluate the checkpoint saved after the first complete pass. The native-capacity floor has no learned parameters.

Training loss alone does not identify PTEA's best spatial cue. In the 4B Block-16 five-fold run, both losses keep falling through epoch twelve, while the held-out evidence score in Figure 5 peaks at epoch three. The score reflects map alignment and coverage of final and last-zoom regions. We therefore fit a separate PTEA for three epochs on all 1,144 traces, choosing the duration from trace targets without using external QA results.

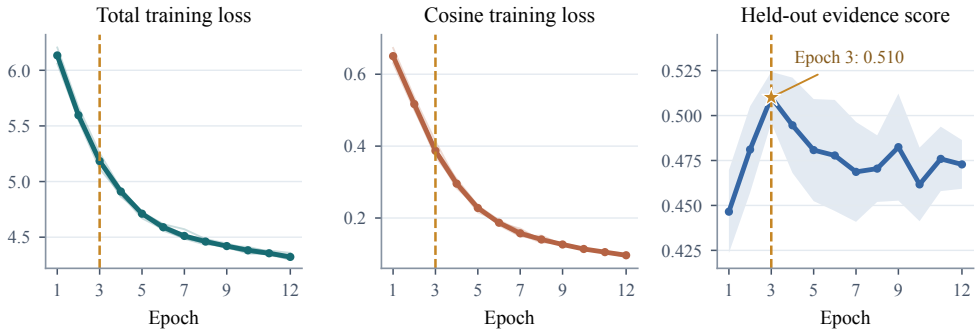

Figure 5: **PTEA training and checkpoint selection.** Five-fold trace-only training of the Qwen3-VL-4B Block-16 selector. Thin curves show individual folds and thick curves their mean. The held-out evidence score is shown with a one-standard-deviation band across folds; dashed lines mark the selected third epoch.

For the audited 4B core, frozen feature extraction, the three PTEA passes, and the selected Bridge pass require 1.046 observed A100 GPU-hours in the run logs; the maximum observed process memory is 36,993 MiB. This audit does not include H-Safe calibration or historical architecture-search compute. These logs estimate the cost of reproducing the attachment. They are not isolated throughput measurements: the stages ran under different shared-machine loads.

Native-capacity floor. Let N be the post-merger token count under the host's original 16M-pixel processor. The planner uses the image-dependent soft-capacity scale

$$F(N) = \begin{cases} 2048, & N \le 4096, \\ 2048 + (N - 4096)/4, & 4096 < N < 12288, \\ 4096, & N \ge 12288. \end{cases} \tag{6}$$

It combines this scale with native demand through the capacity envelope $\sqrt{N^2 + F(N)^2}$. The global view is planned from image size and native capacity; regional views receive the remaining capacity according to their area and allocation priorities, including the context share from ContextNeed. Each allocation is then aligned to the host patch grid, so $F(N)$ is a soft capacity parameter rather than a guaranteed minimum realized token count. The same rule applies across datasets. In the matched-budget sweep, an explicit ceiling instead caps each sample at the realized token count of its paired global-only baseline. This ceiling overrides the soft-capacity envelope when necessary, including every sample at 1K and larger images at 2K. All reported visual-token counts are measured after grid alignment.

A.3 Component comparisons

Table 6 separates the roles combined in the full model.

Table 6: **Where the gains enter.** Reread uses fixed-budget source-pixel views; Boxes adds continuous crop geometry; EviViT is the full model with input-dependent capacity allocation. Accuracy is in percent. Avg. equally weights the eight benchmarks, and Δ is relative to Global B16 (global-only, 16×1024^2-pixel cap) at the same scale. Bold marks each column's best score within a scale.

Variant	VP-E	VP-M	VP-H	V*	HR-4K	HR-8K	PB-1img	Zoom	Avg.	Δ
Qwen3-VL-4B										
Global B16	60.28	40.30	41.51	84.29	76.25	73.38	30.77	45.80	56.57	±0.00
+ Reread	**64.54**	47.01	39.62	86.91	79.50	77.12	**32.05**	53.25	60.00	+3.43
+ Boxes	63.83	**51.49**	45.28	86.39	79.25	76.12	31.60	51.60	60.70	+4.13
EviViT (full)	63.83	50.37	**48.11**	**87.96**	**79.62**	**77.75**	31.83	**53.37**	**61.61**	+5.04
Qwen3-VL-8B										
Global B16	67.38	42.16	44.34	85.86	76.00	72.75	32.32	43.91	58.09	±0.00
+ Reread	64.54	44.40	48.11	89.01	78.38	76.00	32.32	**51.95**	60.59	+2.50
+ Boxes	**70.92**	51.49	**55.66**	87.96	79.75	**81.38**	**32.39**	**51.95**	**63.94**	+5.85
EviViT (full)	69.50	**53.36**	51.89	**91.10**	**80.25**	78.25	32.20	50.65	63.40	+5.31

Fixed-budget re-reading raises the eight-task average at both scales, showing that accessing the selected source pixels matters even before later refinements. Continuous boxes contribute especially at 8B. Relative to global B16, the full model gains 5.04 points at 4B and 5.31 at 8B. At 8B, continuous boxes alone have a slightly higher eight-task mean, while the full configuration is stronger on VP-Medium and V*; the components therefore change the balance across tasks rather than adding a uniform gain at every stage.

Table 7: **Question-conditioned routing and sparse fusion on two large hosts.** Within each host, H-Safe, ContextNeed, the native-capacity floor, token accounting, prompt, decoding, and judge are fixed. Shuffled replaces only PTEA's matched question. Identity disables learned Bridge write-back but retains both Global and Fine streams. Macro is the unweighted mean over the five datasets.

Host	PTEA question	Bridge	VisualProbe	V*	HR-4K	HR-8K	PB-1img	Macro
Qwen3-VL-8B	Aligned	Learned	57.48	91.10	80.25	78.25	32.20	67.86
	Shuffled	Learned	53.01	87.96	80.12	79.62	31.86	66.52
	Aligned	Identity	56.89	90.05	80.25	78.25	31.90	67.47
	Shuffled	Identity	53.40	86.91	80.88	79.12	31.98	66.46
Qwen3.5-9B	Aligned	Learned	57.48	87.96	82.50	81.12	36.65	69.14
	Shuffled	Learned	54.17	87.96	81.12	79.38	36.27	67.78
	Aligned	Identity	56.70	86.39	82.25	81.50	36.80	68.73
	Shuffled	Identity	53.79	87.96	81.25	79.38	35.48	67.57

Table 7 disentangles two complementary roles in EviViT: selecting what evidence to acquire and integrating the acquired evidence. Shuffling only the question supplied to PTEA consistently reduces VisualProbe accuracy across both hosts, while the answering model still receives the original question; the effect on the remaining benchmarks is generally smaller and less uniform. This indicates that question alignment is particularly important for locating the fine-grained evidence targeted by VisualProbe. With the question correctly aligned, replacing the learned bridge with an identity path also lowers the five-dataset mean on both hosts. Together, these results separate the roles of the two components: query-guided selection determines which evidence is acquired, while learned sparse fusion improves how that evidence is incorporated into the host representation.

To isolate the placement of the resulting views from the learned bridge, we also route them through the host's native multi-image interface and perturb only their centers. Table 8 shows that both random translation and cross-image shuffling reduce pooled accuracy by about 9.5 points with view count, size, and token budget fixed. This control supports the role of query-matched locations beyond the effect of adding regional pixels.

Table 8: **Location control with native multi-image input.** Only crop positions change; view count, size, and token budget are fixed. Accuracy (%); Δ versus deployed locations.

Regional locations	VP-All	V*	Pooled	Δ
Deployed locations	57.09	92.15	66.57	–
Random translation	47.77	82.20	57.08	-9.49
Cross-image shuffled locations	47.77	81.68	56.94	-9.63

Table 9: **Paired gains come from rescues rather than answer churn.** Rescue/loss compares identical questions. Retain is the fraction of originally correct answers that remain correct; p is the two-sided exact McNemar p-value. Qwen3-VL and Qwen3.5 results come from completed answer-only paired runs using the same executor.

Host	VP tier	Base	+EviViT	Δ	Rescue/loss	Retain	p
Qwen3-VL-4B	Easy	60.28	63.83	+3.55	16/11	87.06%	0.44
	Medium	40.30	50.37	+10.07	41/14	87.04%	3.6e-04
	Hard	41.51	48.11	+6.60	17/10	77.27%	0.25
Qwen3-VL-8B	Easy	67.38	69.50	+2.13	14/11	88.42%	0.69
	Medium	42.16	53.36	+11.19	45/15	86.73%	1.3e-04
	Hard	44.34	51.89	+7.55	19/11	76.60%	0.2
Qwen3.5-4B	Easy	63.83	70.92	+7.09	19/9	90.00%	0.087
	Medium	44.40	53.36	+8.96	36/12	89.92%	7.2e-04
	Hard	45.28	55.66	+10.38	20/9	81.25%	0.061
Qwen3.5-9B	Easy	65.96	66.67	+0.71	17/16	82.80%	1
	Medium	44.40	54.85	+10.45	37/9	92.44%	4.1e-05
	Hard	42.45	51.89	+9.43	19/9	80.00%	0.087

Evidence localization and visual features. An external V* audit isolates two properties the attachment needs: its evidence map should rank relevant pixels, and its added path should specialize without erasing the host representation. The 186-image overlap with the official benchmark is disjoint from training. On these images, PTEA reaches 0.9294 pixel AUROC (0.2272 average precision), and its top-eight proposals cover at least 90% of every official target in 70.43% of cases. This proposal diagnostic is distinct from the deployed two- or three-region policy in Appendix C.3. Meanwhile, centered-kernel alignment is 0.948 at insertion, 0.895 at the final ViT block, and 0.926 after the merger. The two views tell a consistent story: relevant evidence becomes easier to extract while the frozen host's representational structure remains strongly aligned.

Table 9 tracks the same questions before and after attachment, distinguishing newly correct answers from answers lost. Rescues outnumber losses in every host–tier pair, so the gains reflect repaired errors rather than uniform answer churn. The Medium tier is clearest, with gains of 8.96–11.19 points, retention of 86.73–92.44%, and exact McNemar $p \leq 7.2 \times 10^{-4}$ across all four hosts. Repeating this pattern across Qwen3-VL and Qwen3.5, and across model scales, makes the improvement a family-level result rather than an artifact of one aggregate score.

A.4 Prompts, decoding, and scoring

Our paired evaluations isolate the visual attachment: base and attached hosts are evaluated with identical images, questions, prompts, and deterministic decoding. No evidence boxes are supplied, and the prompt requests a direct answer rather than intermediate reasoning. This keeps the comparison focused on the visual information available to the host.

We score answers with a frozen Qwen3-VL-8B semantic judge, which receives the question, reference answer, and candidate answer, but neither the image nor the method name. Semantic matching accommodates equivalent wording missed by exact string matching. We report a run only after every example has a prediction and a valid judgment; string-match scores serve only as diagnostics. For the longer responses from OPD-V's released policy, both paired runs use a 64K-token judge context with

deterministic Flash attention. This retains the response for scoring while keeping the judge and its instructions unchanged.

A.5 MATCHED RECOVERY SFT

Table 10: **Matched recovery SFT on Qwen3-VL-8B.** Matched 10K data, order, LoRA, and epoch-3 endpoint; EviViT stays frozen. Avg. equally weights VCoT, VP-E/M/H, and Zoom. Green gains compare each SFT row with its frozen path.

Model	Trainable	VCoT	Weak-5	Protect-5	VP-E	VP-M	VP-H	Zoom	Avg.	
Base	Frozen	78.50	89.53	70.86	67.38	42.16	44.34	43.91	55.26	
EviViT	Frozen	78.31	87.94	71.82	69.50	53.36	51.89	50.65	60.74	
Base+SFT	LLM LoRA	79.20	89.55	72.18	70.21	44.03	46.23	42.96	56.53	(+1.27)
EviViT+SFT	LLM LoRA	79.18	88.63	73.10	70.92	53.73	53.77	50.30	61.58	(+0.84)

Table 10 uses a shared 10,000-example QA manifest with 6,694 document/OCR examples, 2,900 examples from the five other Visual-CoT tasks, and 406 TextCaps/V7W replay examples. The document/OCR counts are DocVQA 754, InfographicsVQA 2,004, TextVQA 2,003, DUDE 1,681, and SROIE 252. Both arms consume only the image, question, and answer; evidence boxes are not additional policy inputs.

Language attention projections q, k, v, o use LoRA rank 8, $\alpha = 16$, and dropout 0.05. AdamW uses a learning rate of 10^{-6}, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, zero weight decay, 3% warm-up, and cosine decay. The micro-batch is one question with 20-step gradient accumulation: each epoch contains 10,000 sample steps and 500 optimizer updates. Both arms train for three epochs in the same seeded order, saving every half epoch; the reported pair uses the common final endpoint, not a separate best checkpoint per task. The vision tower and all EviViT modules remain frozen, while language LoRA learns from the visual tokens produced by the corresponding path.

This shared schedule also lets us compare how the two paths learn, not just where they finish. In Figure 6, answer loss falls and answer-token accuracy rises for both arms over three epochs, with EviViT+SFT generally maintaining the lower loss. The frozen attachment thus remains compatible with language-side optimization; Table 10 shows how the resulting models perform on evaluation sets.

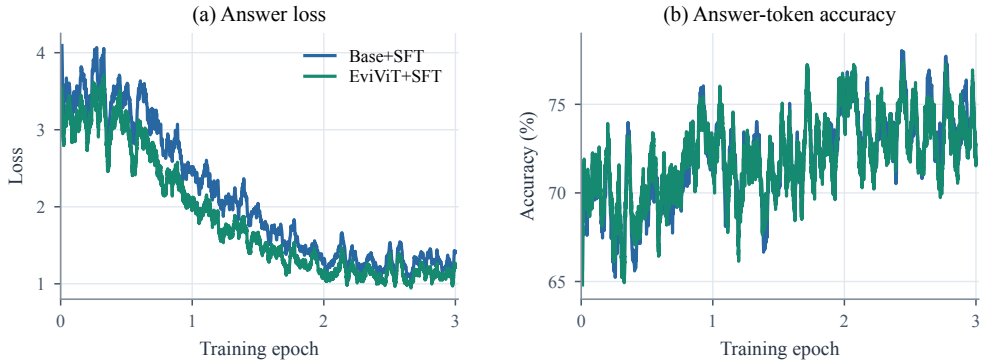

Figure 6: **Matched SFT training dynamics.** Qwen3-VL-8B Base+SFT and EviViT+SFT follow the same 10K-example sequence for three epochs. Answer loss (left) and answer-token accuracy (right) are shown as 400-example moving means; only language LoRA is updated.

Visual-CoT is micro-averaged over 7,427 questions. Weak-5 is the macro-average over DocVQA, InfographicsVQA, TextVQA, DUDE, and SROIE; Protect-5 is the macro-average over Flickr30K, GQA, OpenImages, VSR, and CUB. VP-All weights the 141/268/106 Easy/Medium/Hard questions;

ZoomBench contains 845 questions. The SFT audit uses separately evaluated frozen controls, with the same answer prompt and judge before and after training and judge input/output limits of 8,192/16 tokens. The compatibility audit retains the earlier 1,024/8-token runs on the same questions.

Table 10 shows that downstream language adaptation can use the frozen attachment. Both paths improve on Visual-CoT, while EviViT+SFT keeps its VisualProbe and ZoomBench advantages, improves the Protect-5 macro, and recovers part of the document/OCR gap. The attachment is therefore composable with ordinary language-side LoRA: it supplies specialized visual evidence that the host can learn to use without retraining the visual module.

B Additional Evaluation Results

Fine-grained gains are most useful when the host remains capable beyond the tasks targeted by the attachment. We therefore examine individual perceptual skills on PerceptionBench and broader multimodal capabilities on MMBench, MMStar, and Visual-CoT, comparing each host before and after attaching EviViT.

PerceptionBench subset. PerceptionBench (Lin et al., 2026a) complements the fine-grained benchmarks by testing a wider set of perceptual skills. We evaluate all 2,652 single-image questions from its 3,000-question suite, matching the current attachment's one-image input. We report this complete single-image slice as PB-1img; the remaining 348 questions require a multi-image setting outside the present comparison.

Table 11: **Broad-task compatibility across foundation and post-trained hosts.** Accuracy uses matched inputs, deterministic decoding, and the Qwen3-VL-8B semantic judge.

	MMBench			MMStar			Visual-CoT		
Host	Base	+EviViT	Δ	Base	+EviViT	Δ	Base	+EviViT	Δ
Qwen3-VL-4B	87.36	88.43	+1.06	61.73	62.80	+1.07	77.53	75.52	-2.01
Qwen3-VL-8B	88.59	88.94	+0.35	65.07	65.27	+0.20	77.92	78.35	+0.43
Qwen3.5-4B	86.37	87.76	+1.39	63.20	64.40	+1.20	77.18	78.17	+1.00
Qwen3.5-9B	87.13	86.76	-0.37	64.33	65.20	+0.87	78.98	80.34	+1.36
ZwZ-4B	87.43	85.59	-1.85	62.20	59.20	-3.00	77.77	74.96	-2.81
ZwZ-8B	88.36	86.83	-1.52	65.33	62.20	-3.13	79.08	78.71	-0.36
Vero-Qwen3I-8B	89.88	89.56	-0.32	73.73	72.40	-1.33	77.68	77.97	+0.30
Vision-OPD-4B	84.85	84.45	-0.39	61.53	62.33	+0.80	76.72	76.44	-0.28
Vision-OPD-9B	88.24	86.95	-1.29	66.40	64.27	-2.13	78.35	79.53	+1.18

MMBench, MMStar, and Visual-CoT. Table 11 covers all nine hosts, while Table 12 resolves ten Visual-CoT tasks for three foundation hosts. Across the four foundation hosts, EviViT improves 10 of 12 host–benchmark pairs, showing that its fine-grained gains usually coexist with broad-task preservation. The taskwise view sharpens the pattern: Flickr30K, OpenImages, VSR, and CUB improve for all three reported hosts. Post-trained models vary more because their learned policies consume the added visual sequence differently. Together with the matched SFT result in Appendix A.5, this suggests a coherent division of labor: EviViT supplies additional perceptual evidence, and lightweight language adaptation offers a direct route to aligning how the host uses it.

C Evidence Selection, Integration, and Supervision

C.1 Reusing evidence views across input interfaces

Selecting useful evidence and presenting it to the host are different parts of the design. Useful views should remain informative across input interfaces, and internal fusion should preserve what the host could learn from viewing the same crops separately. We test these properties on Qwen3-VL-8B using all 515 VisualProbe and 191 V^* questions. Every condition retains the deployed regions, their order, resize dimensions, and per-view visual-token counts, together with the same host weights, answer

Table 12: **Visual-CoT ten-task breakdown.** Each host uses one fixed checkpoint across all tasks, the answer-only protocol, and the deterministic 8B semantic judge.

		Qwen3-VL-8B			Qwen3.5-4B			Qwen3.5-9B		
Task	N	Base	+EviViT	Δ	Base	+EviViT	Δ	Base	+EviViT	Δ
Flickr30K	1546	76.97	77.49	+0.52	75.42	79.30	+3.88	76.84	79.24	+2.39
DocVQA	888	95.83	95.27	-0.56	96.06	96.06	±0.00	96.96	96.73	-0.23
GQA	978	68.71	70.35	+1.64	67.38	67.28	-0.10	70.86	72.70	+1.84
InfographicsVQA	360	85.56	82.22	-3.33	85.83	84.44	-1.39	86.39	86.39	±0.00
OpenImages	945	46.56	50.79	+4.23	51.75	53.33	+1.59	50.05	53.54	+3.49
TextVQA	526	93.16	92.59	-0.57	93.54	92.59	-0.95	94.49	94.68	+0.19
VSR	404	75.74	77.23	+1.49	73.02	74.26	+1.24	76.49	78.71	+2.23
DUDE	602	77.74	76.25	-1.50	76.91	77.24	+0.33	76.41	76.91	+0.50
SROIE	686	96.21	93.73	-2.48	96.36	96.21	-0.15	95.92	96.06	+0.15
CUB	492	81.71	83.33	+1.63	70.12	70.93	+0.81	84.55	84.76	+0.20

prompt, deterministic decoding, and Qwen3-VL-8B text-answer judge. The deployed records contain two regions in 22 cases and three in the remaining 684.

Table 13: **Interface portability with frozen evidence views.** Qwen3-VL-8B uses the same regions and per-view token counts. Pooled accuracy weights all 706 questions equally.

Interface	Regional pixel source	VP-All	V*	Pooled
Internal, learned bridge	Original image	57.48	91.10	66.57
Internal, identity bridge	Original image	56.89	90.05	65.86
Native multi-image	Original image	57.09	92.15	66.57
Internal, identity bridge	Resized global view	57.28	91.62	66.57

The full model fuses the views through its learned internal bridge. An identity control retains the global and regional streams and their source coordinates but disables learned bridge updates. The native multi-image control instead presents the global image and original-pixel crops as separate image inputs. Finally, a resized-global control follows the identity path but obtains its crops from the already resized global image, at the same regional input sizes.

Table 13 shows closely matched accuracy for internal fusion and native multi-image input: their pooled scores are identical, with small differences in opposite directions on the two benchmarks. The evidence therefore remains useful across interfaces: integrating it into the host's visual sequence preserves its aggregate answering value.

For deployment, the internal route connects local observations to the global grid through their source coordinates and passes one visual sequence to the host. The application still supplies an image and a question: it need not assemble a separate list of crop images, rewrite its prompt for multiple views, or rely on the host's native multi-image support. The attachment handles region acquisition and integration before answer generation, so the language model does not need to learn when to request a crop or how to manage the returned views. At the same time, coordinate-based fusion keeps local detail connected to its place in the scene. The comparable accuracy shows that this self-contained interface retains the benefit of exposing the evidence, making it practical to add focused perception to compatible hosts without redesigning their answering policy.

Keeping the views fixed separates selection from integration: the learned bridge improves over identity on both benchmarks, while native multi-image input shows that the views are already informative without it. Coordinate-aware exchange therefore adds a distinct second benefit. It binds local detail back to the scene representation, letting the host use the selected evidence inside its existing visual sequence.

C.2 WHY THE SELECTED LOCATIONS MATTER

The interface comparison raises a further question: are the selected views useful because they locate relevant evidence, or would additional crops of similar size work just as well? Under a limited visual budget, enlarging an uninformative part of the image spends tokens without helping to find the answer. We therefore keep the amount and form of local input fixed and change where it comes from.

All three conditions use Qwen3-VL-8B's native multi-image interface from Table 13, presenting the global view and crops as separate image inputs. This tests the selected evidence independently of learned bridge fusion, rather than perturbing locations inside EviViT's internal path. For each of the same 706 questions, we preserve the global view, region count and order, crop pixel dimensions, output grids, roles, and per-view token allocations. Host weights, answer prompt, decoding, and judge also remain unchanged.

Random translation moves each crop to a sampled valid position in the same image. Cross-image shuffling borrows ordered centers from a different image in the same benchmark with the same region count, retaining the current sample's crop sizes. This preserves positions selected for another image–question pair but breaks their alignment with the current one. Both controls use a seed fixed before inference and do not consult question content, answers, predictions, or target annotations.

Table 8 shows that the selected locations outperform both alternatives on VisualProbe and V*, with pooled margins of 9.49 and 9.63 points. The controls retain equally sized and detailed views, with comparable mean crop-union area, yet lose much of the answering benefit. The important factor is therefore not simply seeing more crops, but spending those views on evidence relevant to the current image and question. Together with Appendix C.1, this explains the attachment's reusable value: its selected evidence remains informative across interfaces because of what the views reveal, not just how they are presented.

C.3 COVERAGE BY THE REGIONS ACTUALLY DEPLOYED

EviViT allocates local detail to evidence, rather than trying to cover the whole scene with crops. We examine whether its selected regions reach the objects needed to answer a question, including cases that require more than one target. We evaluate the final post-H-Safe crops on 186 V* examples with 240 official target boxes. These 132 single-target and 54 two-target questions are disjoint from the attachment-training manifest. We use the saved integer-pixel crop rectangles, excluding the global view. For target T_{ik} and the union U_i of local crops in sample i, coverage is $c_{ik} = |T_{ik} \cap U_i|/|T_{ik}|$. We report the mean target coverage in each sample, its minimum target coverage, and the fraction of samples for which every target reaches 90% coverage.

Table 14: **Coverage of official targets by deployed local crops.** All 186 V* examples and 240 targets are retained. Mean averages target coverage within each sample, then across samples; Min averages each sample's least-covered target. All90 requires every target to reach 90% coverage. Area is crop-union/image area. All values are percentages.

Region layout	Mean	Min	All90	Area
Deployed EviViT regions	73.69	68.28	64.52	12.39
Joint random translation	18.47	15.95	12.38	12.39
Joint centering	25.83	22.55	15.59	12.39
Independent random translation	14.47	11.73	9.00	12.20

Table 14 shows that the selected regions concentrate useful detail into a small part of the image: they occupy 12.39% of its area while achieving a mean target-area coverage of 73.69%. To test whether this comes from the choice of locations, we translate each complete crop layout to 10,000 random valid positions, preserving its sizes, relative positions, overlap, and exact union area. We also center the same layout, and separately randomize individual crops while preserving their sizes. None of these controls uses target annotations. Joint-random placement reaches only 18.47% mean target coverage, and centering also leaves a substantial gap. The selected views are therefore targeted observations, not simply a large or centrally placed window onto the image.

Finding the most prominent evidence is only part of the task. A relation question may require another object outside that focal region, which is why EviViT retains a complementary context view. Every target receives at least 90% coverage in 120 of the 186 questions; removing context from the saved crop set reduces this to 98. Of the 22 additional cases covered with context, six contain two targets. In five of these, the decisive views already cover one object but miss the other completely; context supplies the missing side of the relation. This geometric comparison gives a concrete role to the context view: it complements the focal evidence instead of repeatedly inspecting the same object.

These observations clarify how local selection works with the global view. The crops provide concentrated detail for the objects most relevant to the question, while the full-image stream retains the surrounding scene and positions outside the selected regions. Coordinate-based fusion connects the two, so the host can interpret a readable local attribute in its broader spatial setting. The model is therefore not restricted to answering from isolated evidence boxes: focused observations enrich, rather than replace, its view of the scene.

C.4 The value of full-trace supervision

A final evidence box records where a search ends, but not the observations that led there. The full record also retains inspected regions, changes of focus, and surrounding context. These cues provide a richer spatial target for learning where to allocate detail. We test their value by independently retraining the complete attachment under two supervision conditions, with the Qwen3-VL-8B host frozen throughout.

The Final-box condition learns its evidence map from terminal boxes and forms context targets by geometrically expanding those boxes. Full-trace uses the human-event mixture for evidence and the recorded exploration for context. Both conditions draw from the same 1,144-example training set, fitting spatial targets on its valid search records and the bridge on all training QA examples. They share the attachment architecture, training schedule, and seed. Each follows the complete staged fit: a newly initialized PTEA, a context predictor initialized from its own PTEA, independently fitted H-Safe and ContextNeed, and a new identity-initialized bridge. The box-refinement targets and the training-QA criterion for fitting ContextNeed are shared. We evaluate the common three-epoch bridge endpoint under the same native adaptive allocation policy, prompt, decoding, and Qwen3-VL-8B semantic judge.

Table 5 shows why the extra spatial information is useful in this matched run. Full-trace improves all three displayed metrics, with the largest gain on VP-Hard. The final box identifies the target at the end of a search, whereas the visited regions, final branch, and last zoom also locate nearby objects and plausible alternatives. Their combined density can guide the selector toward detail that is both readable and situated in the scene. This offers a plausible account of the Hard result; the comparison tests the full supervision recipe rather than isolating any one trace cue. Native ZoomBench is tied between the two trained attachments, so the observed advantage is strongest on the displayed fine-grained tasks rather than uniform across datasets.

D Qualitative Evidence-Localization Cases

Figures 7–9 make the mechanism concrete across small-text recognition, relative-position reasoning, and evidence-grounded explanation. In each case, EviViT recovers the decisive local cue while preserving the scene needed to bind that cue to the question. The examples further suggest that better localized evidence can make the resulting reasoning more focused and concise by reducing attention to irrelevant scene elements. The Vero-8B example additionally includes both complete generated rationales.



Figure 7: **Small-text repair on VisualProbe Hard.** The base host misreads the roof identifier as "6C71A." EviViT recovers "6C71" while retaining the street context needed to locate the queried car.



Figure 8: **Relative-position repair on V*Bench.** Qwen3-VL-4B predicts that the soccer ball lies to the right of the long bench. EviViT allocates native detail to the ball, the relevant bench, and a complementary scene anchor, changing the answer to the correct left-side relation.

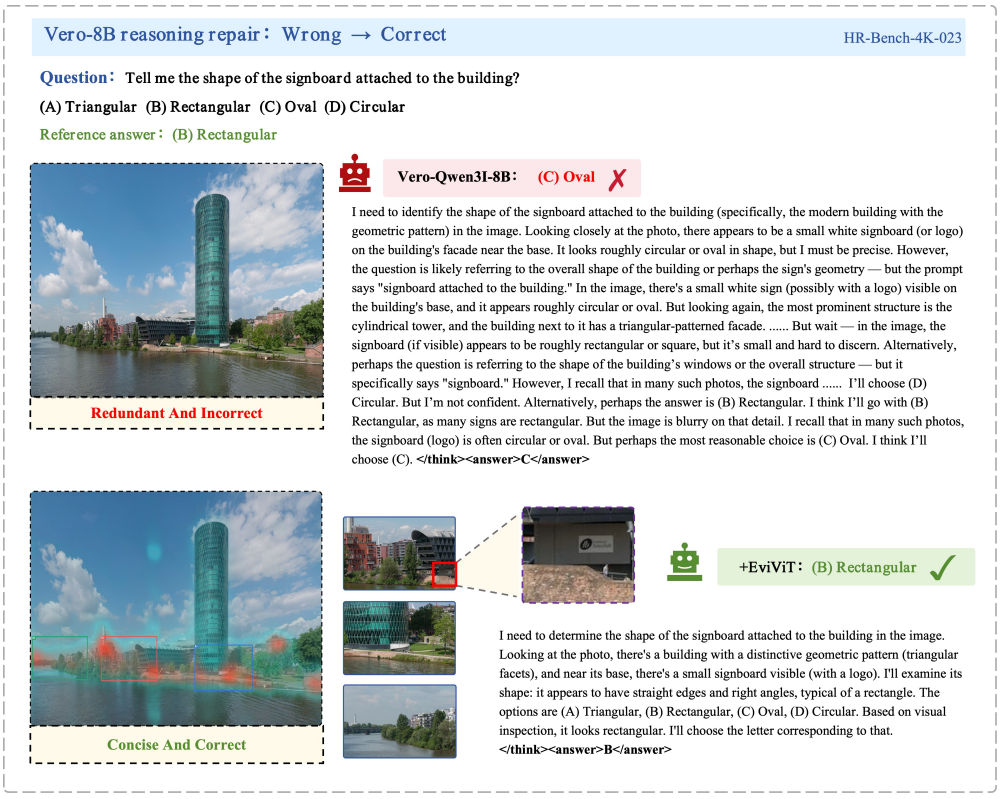

Figure 9: **Evidence-grounded reasoning repair on Vero-8B.** The base response considers the tower, facade, and logo before incorrectly answering "oval." EviViT localizes the signboard and grounds a shorter, correct rationale in its straight edges and right angles. Both outputs are reproduced in full.